2000년도 국고 보조금 지원 사업

국어 정보화 연구 인력 양성 사업
— 국어정보화 아카데미 개최 및 국어정보화 교재 개발 —

제2과제 : "한국어 말뭉치의 활용"(교재)의 개발

최종보고서

2000.12.

연세대학교 언어정보개발연구원

2000년도 국고 보조금 지원 사업

국어 정보화 연구 인력 양성 사업
국어정보화 아카데미 개최 및 국어정보화 교재 개발 —

제2과제 : "한국어 말뭉치의 활용"(교재)의 개발

연구책임자 : 한영균(울산대)
연 구 원 : 서상규(연세대)
연구 보조원 : 차윤선(연세대 대학원)
우민주(연세대 대학원)
이철성(연세대 대학원)
정성린(울산대 국제한국학교육센터)

2000.12.

연세대학교 언어정보개발연구원

2260197

제 출 문

문화관광부 장관 귀하

본 보고서를 국고지원 <국어 정보화 연구 인력 양성 사업— 국어정보화 아카데미 개최 및 국어정보화 교재 개발 —(제2과제 : "한국어 말뭉치의 활용"(교재)의 개발)> 사업 보고서로 제출합니다.

2000년 12월

연구책임자 : 한 영균(울산대)

요 약 문

1. 사업명

국어 정보화 연구 인력 양성 사업
— 국어정보화 아카데미 개최 및 국어정보화 교재 개발 —
제2과제 : 한국어 말뭉치의 활용"(교재)의 개발

2. 연구의 목적

○ 세종계획을 비롯하여, 그간 국내에서 구축된 각종 국어 말뭉치를 실제로 다루면서, 각종 공개 프로그램(텍스트 편집기, 용례 추출기, 자동 품사 주석기, 음절·단어·어절 등의 단위를 계량화하기 위한 각종 유틸리티 등)을 이용하여, 언어 자료를 다루는 방법과 실제 사용법을 중심으로 구성한, "한국어 말뭉치의 활용"(국어정보학 교재)을 개발함.

3. 연구의 필요성

○ 2단계에 들어 더욱 활성화될 '국어정보화 아카데미'의 기본 교재가 부족함.
○ 21세기 세종계획에 의해 개발된 각종 말뭉치와, 말뭉치 분석 처리 도구(소프트웨어)를 일반인이나 연구자들이 쉽게 활용할 수 있도록 하는 입문적 실습서를 개발·보급함으로써, 세종계획에 의한 각종 결과물의 활용도를 극대화할 필요가 있음.
○ 이제까지 국책 또는 기관 및 개인 연구자에 의해 개발되어 활용되고 있는 한국어 말뭉치의 각 분야에서의 활용을 활성화함으로써, 국어 정보화의 저변 확대에 기여함.

4. 연구내용 및 결과물

○ 국어정보학의 각 영역에서 전자말뭉치를 활용하는 방법에는 어떠한 것이 있으며, 실제 말뭉치를 활용한 연구 결과가 기존 연구와 어떻게 다른가 하는 점을 구체적인 사례를 중심으로 설명하고, 국어정보학에서의 말뭉치 활용 방안 및 말뭉치를 구축, 분석하는 데에 주의하여야 할 점에는 어떤 것들이 있는가를 보여줌으로써 구축된 말뭉치의 활용도를 높이고, 향후의 말뭉치 구축 방법론의 개선에 참조할 수 있도록 하였다.
○ 이제까지 국책 또는 기관 및 개인 연구자에 의해 개발되어 활용되고 있는 한국어 말뭉치를 실제로 활용하기 위한 방법론과 도구를 소개함으로써, 말뭉치 활용 능력을 배양토록 하는 교재를 개발하였다.

○ 결과물 : "한국어 말뭉치의 활용"(교재 1책 원고)

5. 연구결과의 활용

○ 2단계에 들어 더욱 활성화될 '국어정보화 아카데미'의 기본 교재로 활용함.

○ 홈페이지에 공개함으로써, 다양한 분야와 목적의 일반인과 연구자에게 활용할 수 있도록 보급함으로써, 세종계획의 성과를 활발히 이용할 수 있게 함.

○ 단행본으로 출판·보급하여, 세종계획에 대한 인지도를 더욱 높임.

차 례

제1부
국어정보학에서의 말뭉치 활용의 제측면 1

1. 언어학적 측면에서의 말뭉치활용 2

1. 순수언어학적 측면에서의 말뭉치의 활용 3
1.1. 서론 3
1.2. 음운론적 측면에서의 말뭉치 활용 5
1.3. 형태론적 측면에서의 말뭉치 활용 6
2. 언어 정책 수립과 말뭉치의 활용 9
3. 언어 교육에 있어서의 말뭉치의 활용 11

2. 자연언어처리 측면에서의 말뭉치활용 13

1. 서론 13
2. 언어처리 시스템 개발과 말뭉치의 활용 15
2.1. 서론 15
2.2. 형태분석과 말뭉치 15
2.3. 구문해석과 말뭉치 19
2.2 기계번역과 말뭉치 21

3. 사전편찬학 측면에서의 말뭉치 활용 23

1. 서론 24
2 거시 구조 구성에 있어서의 말뭉치 활용 25
2.1. 표제항의 적정규모 25
2.2. 표제항이 되는 언어단위의 선정 27

제2부
한국어 말뭉치의 현황과 활용도구 30

1. 한국어 말뭉치의 현황 31

1.1 연세 한국어 말뭉치(Ysc :Yonsei Korean Corpora) 32
1.2. 국립국어연구원 말뭉치 41
1.3. 국어 정보 베이스 말뭉치 44
1.4. 고려대학교 한국어 말모둠 1 50
1.5. 고려대학교 장르별 텍스트 코퍼스(KGENRE Corpus) 51
1.6. <21세기 세종계획> 국어 기초자료(1998) 52
1.7. <21세기 세종 계획> 국어 기초자료(1999) 55
1.8. ETRI 품사태그 부착 말뭉치 (시험판 CD) 76
1.9. 말뭉치 서비스 기관 : 21세기 세종 계획 자료실 80

2. 말뭉치 활용 도구 92

2.1 말뭉치 구축을 위한 텍스트의 개념 이해와 텍스트 편집기 93
2.1.1 텍스트 파일이란? 93
2.2 말뭉치를 가공하기 위한 자동품사 주석기와 그 활용 96
2.2.1 품사 주석이란? 96
2.2.2 자동 품사 주석기와 형태소 분석기 97
2.2.3 지능형 형태소 분석기 98
2.2.4 반자동 품사 주석기 '꼬리달개' 108
2.3 말뭉치를 이용한 용례의 추출과 활용 114
2.3.1 말뭉치와 용례 추출이란? 114
2.3.2 용례 검색기 '글잡이 1.0'(윈도용) 114
2.3.3 '글잡이 1.0' 부속 도스용 용례 추출기 129
2.3.4 패턴매치를 이용한 용례 추출기 hgrep 142
2.4 언어정보를 계량화하기 위한 처리 도구와 활용 146
2.4.1 국어정보화 프로그램(21세기 세종 계획) 146
2.4.2 21세기 세종계획 기초언어자료 처리 도구 147
2.4.3 말뭉치 주석 및 계량 도구 'Ytagger' 164
2.4.4 반자동 의미 주석용 프로그램 '참뜻' 173
2.4.5 전처리 프로그램 "전처리개 Ver1.5" 181

3. 실습용 말뭉치의 활용 184

한국어 말뭉치의 활용

연구책임자: 한영균(울산대)

공동연구원: 서상규(연세대)

2000.12.

<차 례>

제1부
국어정보학에서의 말뭉치 활용의 제측면 1

1. 언어학적 측면에서의 말뭉치활용 2

1. 순수언어학적 측면에서의 말뭉치의 활용 3
1.1. 서론 3
1.2. 음운론적 측면에서의 말뭉치 활용 5
1.3. 형태론적 측면에서의 말뭉치 활용 6
2. 언어 정책 수립과 말뭉치의 활용 9
3. 언어 교육에 있어서의 말뭉치의 활용 11

2. 자연언어처리 측면에서의 말뭉치활용 13

1. 서론 13
2. 언어처리 시스템 개발과 말뭉치의 활용 15
2.1. 서론 15
2.2. 형태분석과 말뭉치 15
2.3. 구문해석과 말뭉치 19
2.2 기계번역과 말뭉치 21

3. 사전편찬학 측면에서의 말뭉치 활용 23

1. 서론 24
2 거시 구조 구성에 있어서의 말뭉치 활용 25
2.1. 표제항의 적정규모 25
2.2. 표제항이 되는 언어단위의 선정 27

제2부
한국어 말뭉치의 현황과 활용도구 30

1. 한국어 말뭉치의 현황 31

1.1 연세 한국어 말뭉치(Ysc :Yonsei Korean Corpora) 32
1.2. 국립국어연구원 말뭉치 41
1.3. 국어 정보 베이스 말뭉치 44
1.4. 고려대학교 한국어 말모둠 1 50
1.5. 고려대학교 장르별 텍스트 코퍼스(KGENRE Corpus) 51
1.6. <21세기 세종계획> 국어 기초자료(1998) 52
1.7. <21세기 세종 계획> 국어 기초자료(1999) 55
1.8. ETRI 품사태그 부착 말뭉치 (시험판 CD) 76
1.9. 말뭉치 서비스 기관 : 21세기 세종 계획 자료실 80

2. 말뭉치 활용 도구 92

2.1 말뭉치 구축을 위한 텍스트의 개념 이해와 텍스트 편집기 93
2.1.1 텍스트 파일이란? 93
2.2 말뭉치를 가공하기 위한 자동품사 주석기와 그 활용 96
2.2.1 품사 주석이란? 96
2.2.2 자동 품사 주석기와 형태소 분석기 97
2.2.3 지능형 형태소 분석기 98
2.2.4 반자동 품사 주석기 '꼬리달개' 108
2.3 말뭉치를 이용한 용례의 추출과 활용 114
2.3.1 말뭉치와 용례 추출이란? 114
2.3.2 용례 검색기 '글잡이 1.0'(윈도용) 114
2.3.3 '글잡이 1.0' 부속 도스용 용례 추출기 129
2.3.4 패턴매치를 이용한 용례 추출기 hgrep 142
2.4 언어정보를 계량화하기 위한 처리 도구와 활용 146
2.4.1 국어정보화 프로그램(21세기 세종 계획) 146
2.4.2 21세기 세종계획 기초언어자료 처리 도구 147
2.4.3 말뭉치 주석 및 계량 도구 'Ytagger' 164
2.4.4 반자동 의미 주석용 프로그램 '참뜻' 173
2.4.5 전처리 프로그램 "전처리개 Ver1.5" 181

3. 실습용 말뭉치의 활용 184

제1부

국어정보학에서의 말뭉치 활용의 제측면

1. 언어학적 측면에서의 말뭉치활용

이 장에서는 언어학적 측면에서의 말뭉치를 활용한 연구의 제측면을 개관한다. 순수언어학적 측면에서는 대체로 말뭉치 언어학이라고 지칭되어 온 연구방법을, 언어정책적 측면에서는 한자혼용 문제 및 국어 순화와 관련된 문제를, 언어교육과 관련해서는 주로 제2언어로서의 한국어 교육에서의 말뭉치 활용 방법에 대해서 간단히 살펴본다.

1. 순수언어학적 측면에서의 말뭉치의 활용

1.1. 서론

말뭉치를 활용한 언어 연구는 뭉치언어학Corpus Linguistics이라는 명칭으로 언어학의 하위 영역의 하나로 자리 잡았다.

영미 학계에서 corpus linguistics라는 용어가 언제부터 사용되었는지는 확실하지 않다. 다만 1950년대에 미국의 실증주의-행동주의적 구조언어학자들이 실제 언어 자료를 제일차적 설명 대상으로 삼은 것을 효시로 볼 수 있을 것이다. 한편 일단의 영국 언어학자들이 1959년에 더럼Durham 대학에서, 1960년 이후에는 런던 대학에서, "영어용법조사말뭉치Survey of English Usage Corpus"라는 이름의 프로젝트로 영어의 실제 쓰임을 광범위하게 장기적으로 조사하기 시작한 바 이로써 실질적인 말뭉치를 다루는 언어 연구가 재출발했다고 할 수 있다. 그러나 말뭉치에 대한 연구가 본격화된 것은 1963-4년에 미국 브라운Brown 대학의 프랜시스Nelson Francis와 쿠체라Henry Kučera가 15 범주에 걸친 500편의 글에서 2000 어절씩 추출하여 총 100만 어절 말뭉치를 구축하고 컴퓨터를 이용하여 이를 분석해서 어휘에 관한 여러 가지 통계를 제시한 이후라고 할 수 있다. 이것이 이른바 브라운 코퍼스Brown Corpus라는 것으로서 현대적 의미의 뭉치언어학의 시초가 된다고 할 수 있다.[1)]

국내에서 말뭉치언어학이 관심을 끌게 된 데에는 1980년대말부터 본격화되기 시작한 새로운 사전의 편찬을 위한 움직임이 큰 영향을 미쳤다.[2)] 다양한

1) 그러나 Corpus Linguistics라는 용어가 본격적으로 사용되기 시작한 것은 것은 1980년대라고 할 수 있다. 1984년에 아르츠Jan Aarts와 메이스Willem Meijs가 영어 말뭉치를 컴퓨터를 이용하여 분석한 결과들을 다룬 논문을 모아 *Corpus Linguistics*라는 책자를 간행한 이후 비슷한 이름의 단행본들이 계속 출간되고 있으며, 1996년에는 *International Journal of Corpus Linguistics*라는 잡지가 John Benjamins에서 창간되었다.

어휘의 용법을 조직적·체계적으로 제시하고, 실제 용법을 바탕으로 한 뜻풀이 및 용례의 제시를 중시하는 현대의 사전 편찬에 말뭉치의 이용이 필수적이라는 것은 두말할 나위도 없는 것이어서,[3] 국어의 여러 문법 단위에 대한 철저한 기술적descriptive 연구도 부족하고 사전 편찬의 경험도 그리 많이 축적하지 못한 국어학계에 뭉치언어학의 경험적empirical·기술적descriptive 연구 방법론은 양자의 필요성을 함께 충족할 수 있는 유일한 대안으로 제시되었던 것이다.

그러나 말뭉치언어학의 방법론을 수용하기 위해서는 여러 가지 문제가 먼저 해결되어야 했다. 말뭉치언어학은 기본적으로 대량의 전산화된 자료와 그 자료를 처리하고 필요한 정보를 추출할 수 있는 언어처리도구, 그리고 추출된 정보의 해석에 동원되는 통계 도구 등을 필요로 한다. 그러나 말뭉치언어학이라는 개념이 처음 소개된 1980년대 후반, 국내에는 활용할만한 전산화된 자료나 도구가 전무한 상태였고, 전산학 분야에서도 한국어의 처리를 위한 연구가 태동하는 단계였다. 이러한 상황에서 사전 편찬을 계획한 이들은 직접 말뭉치를 구축하고, 그를 가공하기 위한 도구와 언어정보의 추출을 위한 도구를 개발해야 부담을 안고 있었던 것이다.

그럼에도 불구하고 지난 10년 사이 국내에서의 전산화된 말뭉치를 이용한 연구는 그 방법론의 개발, 자료의 축적, 언어처리 도구의 개발 등에서 측면에

2) 1980년대 말부터 활성화되기 시작한 새로운 국어사전의 필요성에 대한 논의는 몇 개의 새로운 사전 편찬 프로젝트의 출발로 이어졌다. 현재 널리 알려져 있는 것만으로도 연세대학교 사전편찬실에서 시도하고 있는 한국어 대사전 편찬 프로젝트, 국립국어연구원이 추진하고 있는 『(가칭)종합국어대사전』 편찬 프로젝트, 고려대학교 민족문화연구소에서 추진하고 있는 사전편찬 프로젝트를 들 수 있다. 이들 각각은 지향하는 사전의 규모나 성격은 조금씩 다르지만, 기본적으로 실제 문헌에 출현하는 용례를 바탕으로 한 문법정보 및 뜻풀이, 용례 제시 등을 기본으로 한다는 점에서 같은 방법론을 채택하고 있다고 할 수 있다.

3) 1987년 출간된 『코빌드 영어 사전(COBUILD English Dictionary)』은 전적으로 말뭉치를 이용하여 편찬된 것으로 사전 편찬의 역사에 중요한 획을 그은 것으로 평가받고 있다.

서 비약적인 발전을 보였다. 본 강좌에서는 이러한 발전의 성과를 바탕으로 언어정보의 추출을 목적으로 한 기초자료 말뭉치의 구축, 보다 정확하고 효율적인 정보의 추출을 위해 필요한 말뭉치의 주석 등 말뭉치의 구축 방법론과 관련된 여러 문제를 정리하고, 나아가 말뭉치를 활용한 언어 연구의 결과물로서뿐만 아니라 그 자체가 언어 연구를 위한 가장 훌륭한 텍스트가 될 수 있는 사전의 구조, 그리고 전자화된 사전의 정보학적 활용을 위한 사전 구조화의 문제를 정리한다. 다른 한편으로 언어 연구에 있어서의 말뭉치의 활용이 현대국어연구에 그치는 것이 아니라, 역사언어학·방언학·국문학 연구에도 적용될 수 있는 것이다.

1.2. 음운론적 측면에서의 말뭉치 활용

국어 음운론 연구는 구조주의 및 생성음운론의 방법론에 따른 음소 목록의 확인 및 규칙 설정 등을 중심으로 진행되어 왔다. 그러나 정작 가장 기초적인 음운론적 정보는 아직 제대로 정리되지 않은 것들이 많다.

우선 국어의 음소 혹은 음성의 빈도에 관한 정보가 전무한 것을 지적할 수 있다. 문어를 중심으로 한 자소 빈도는 여러 연구를 통해서 어느 정도 밝혀졌다고 할 수 있지만, 음성언어의 기초를 이루는 음소 및 음성에 관한 정보는 아직 제대로 밝혀진 것이 없는 것이다. 따라서 음소 및 음성의 연결 빈도, 음성언어에서의 음절 빈도 등에 대한 것도 아직 정리되지 않았다고 할 수 있다.

두 번째로는 국어의 음성·음운 규칙과 관련된 통계적 정보가 제대로 얻어지지 않은 것을 지적할 수 있다. 생성음운론의 도입 이후 현대국어의 음운규칙은 대체로 그 전모가 밝혀졌다고 할 수 있지만, 구체적으로 어떤 음성·음운규칙이 어느 정도의 비율을 차지하는지는 밝혀진 바가 없는 것이다. 이는 국어의 음성·음소의 빈도와 함께, 국어 음성언어의 현실을 파악하는 데에 아주 중요한 정보라고 할 수 있으며, 나아가 외국인을 대상으로 한 한국어교육, 기계 번역 및 통역, 음성인식시스템의 개발, 문서 => 음성 변환시스템의

개발 등 여러 가지 응용 분야에 활용될 수 있는 것으로서 그 중요성은 아무리 강조해도 지나치지 않은 것이라고 할 수 있다.

이러한 연구는 구어 말뭉치의 구축이 이루어진 후에야 본격적인 연구가 가능한 것이지만, 문어말뭉치를 바탕으로 한 연구도 전혀 불가능한 것은 아니다. 그것은 문자=>음성 변환 시스템에서 사용하는 기술을 이용해서 문어 말뭉치를 구어 형식으로 변환하고, 그 결과를 이용하여 음소 차원의 여러 가지 기초정보를 획득할 수 있는 것이다. 이는 적절한 규모의 구어말뭉치가 구축되면 상호 비교자료로서도 활용할 수 있으므로 현 단계로서는 읽기 변환 규칙을 정교화하고, 그 결과를 바탕으로 통계를 얻어내는 것도 한 방법이 될 수 있다.

1.3. 형태론적 측면에서의 말뭉치 활용

실제로 말뭉치를 활용한 언어학적 연구 중에서 가장 큰 성과를 얻을 수 있는 분야 중의 하나가 형태론적 연구라고 할 수 있다.

음성 • 음운이라는 것이 언어의 제2분절을 이루는 요소로서 그 수가 상대적으로 적고 따라서 조합 역시 어느 정도 한정되어 있는 것이지만, 제1차분절의 단위인 형태소 및 단어는 그 수가 음성 • 음운과 비교할 수 없을 정도로 수가 많을 뿐 아니라, 그 분포와 쓰임이 다양한 것이기 때문에 그를 대상으로 한 통계적 연구는 직관적으로는 도저히 감당해 낼 수 없는 일인 것이다.

문장 속에서 어떤 단어나 형태소가 보이는 분포는 두 가지 관점에서 살펴볼 수 있다. 하나는 그것들이 어떤 문장의 특정 위치에 나타날 수 있는가 없는가를 살펴 보는 것이며 다른 하나는 문장 속에서 그것들이 나타나는 경향을 살펴 보는 것이다. 문법적 특성에 비추어 보면 나타날 수 있는 자리이지만 실제 사용에서는 그 자리에 쓰인 경우가 발견되지 않는 경우나, 여러 자리에 나타날 수 있는 형태소가 실제 쓰임에서는 특정 위치에만 나타나는 경우 등에 대하여 그 추세를 파악하는 것을 의미한다.

한 예로 조사가 어떤 위치에 나타날 수 있는가 하는 문제는 그 동안의 조사 연구를 통하여 다루어진 반면, 조사가 사용되는 경향은 각 조사의 어휘적인 특수성에 의해 표출되는 것이며 단일한 원리로 규칙화되기보다는 실제 쓰임의 추세 즉 빈도를 통해 파악될 수밖에 없는 것이기 때문에 기존의 연구를 통해 구체적으로 다루어지지 못하였다. 그러나, 사전편찬이나 자연언어 처리 등의 응용을 전제로 한 문법 기술은 규칙을 밝혀내고 그 규칙을 실현시키기 위한 어휘의 선택제약을 기술하는 것만으로는 충분하지 못하며, 실제 쓰임의 경향에 대한 빈도 정보를 요구하는 것이기 때문에, 조사의 분포를 가·부의 관점에서가 아니라 경향의 관점에서 살펴 보는 작업이 필수적이라고 할 수 있을 것이다(이러한 문제를 중심으로 다룬 것이 남윤진(2000)이라고 할 수 있다).

남윤진(2000)에서 밝혀진 바에 따르면 국어의 조사류들은 그 분포상 다음과 같은 특성을 보인다.

1. 고른 분포를 보이는 조사

ㄱ. 격조사: 이/가, 을/를, 에, 에서, 으로

ㄴ. 접속조사: 의, 이나(접속)

ㄷ. 보조조사: 은/는, 이나(강조), 뿐, 들, 부터, 까지, 이야, 도, 이라도, 이라고(양보), 이라고(인용), 요

2. 명사 뒤에만 오는 조사

ㄱ. 격조사: 께서, 께(방향), 에게, 한테, 더러, 보고, 이랑, 하고, 아

ㄴ. 접속조사: 이랑, 하고

ㄷ. 보조조사: 마다, 마따나, 이야말로, 일랑, 께(어림), 꼴, 깨나, 가량, 마냥, 가지고, 말고, 끼리, 짜리, 치고

3. 명사 및 명사형 뒤에만 오는 조사

만(비교), 마저, 보다, 이란, 서, 같이, 으로써, 은/는커녕

4. 용언의 종결형 뒤에만 오는 조사

그려, 마는, 고

5. 용언의 연결형 및 종결형 뒤에 오지 않는 조사
쯤, 처럼, 서, 씩, 이나마, 껏,
6. 부사류 뒤에 오지 않는 조사
와/과, 밖에, 만큼, 보다, 조차, 다가

이러한 연구 결과는 조사는 체언 뒤에 붙는다고 하는 일반적 상식과는 전혀 다른 모습을 보여 준다. 중요한 점은 이러한 사실을 직관을 통해서 파악하는 것이 거의 불가능하다는 점이다. 결국 말뭉치를 활용한 경험적 연구만이 국어의 제1차분절을 이루는 요소들의 분포와 쓰임을 제대로 파악할 수 있다고 할 수 있는 것이다.

2. 언어 정책 수립과 말뭉치의 활용

우리나라의 언어 정책 중에서 가장 큰 혼란을 겪어온 것으로 이른바 한자의 혼용 여부 및 한자 교육 문제를 들 수 있다. 이는 근본적으로 한글의 창제 원리와 운용 원리에 대한 이해 및 적용과 깊은 관련을 가지고 있는 문제이지만, 다른 한편으로는 국어라는 총체를 이루고 있는 구성요소를 어떻게 파악하느냐 하는 문제와도 결부되어 있는 것이라고 할 수 있을 것이다. 한자어가 우리말이 아니라고 생각하는 경우는 별개의 문제라고 하겠지만, 한자어 역시 우리말을 이루고 있는 요소 중의 하나라고 한다면, 한자어가 차지하고 있는 비중을 정확히 파악하는 것과 개개 한자의 분포 및 빈도에 관한 정확한 빈도를 얻어내는 일은 올바른 한자정책의 수립을 위해서 무엇보다도 시급하고 반드시 해내야 할 일이라고 할 수 있다.

이를 위해서는 일반적인 언어연구를 위해서 말뭉치를 구성하는 것과는 다른 방법이 필요하다.

우선 적정한 규모의 표준 말뭉치를 구성할 필요가 있다는 것은 일반적인 말뭉치를 기반으로 한 연구와 동일하다고 할 수 있다. 그러나 국어교육 혹은 국어 정책의 수립을 위한 것임을 감안한다면 말뭉치를 구성하는 텍스트의 장르별 비율은 일반 말뭉치와는 달라져야 할 것이다. 초 • 중 • 고등학교의 주요 교과목의 교과서가 전체적으로 포함되어야 함은 물론, 일반인들이 즐겨 읽는 신문 • 잡지의 비율이 다른 경우보다는 높아질 필요가 있을지도 모른다.

분석 대상 말뭉치의 규모 역시 그 신뢰성을 높이기 위해서는 상당이 커야 할 것인데, 문제는 규모 자체보다는 말뭉치를 목적에 맞게 가공하는 작업이 필요하다는 점이다. 즉, 말뭉치에 들어 있는 모든 한자어를 한자로 변환하고, 그를 대상으로 해서 한자어 및 한자의 빈도, 텍스트 내에서의 비율 등등을 분석할 필요가 있는 것이다. 이러한 작업은 일반 말뭉치를 구축하는 작업과는 달리 한자 및 한자어에 대한 정확한 지식을 가지고 있는 이만이 할 수 있는 일이다.

이와 함께 국어의 순화를 위한 노력에도 말뭉치의 적극적인 활용이 필요하다. 한 통계에 따르면 일간 신문의 경제면을 구성하는 어휘의 교체 비율은 5년에 30%가 넘는다고 한다. 그만큼 새로운 용어들이 등장하고 사라지는 비율이 높다는 뜻이다. 이러한 새로운 용어는 많은 경우, 번역어가 아닌 음역어인 경우가 많다. 뿐만아니라 일본 한자어를 그대로 들여다 쓰는 경우도 적지않다. 이러한 문제를 해결하기 위해서는 새로운 용어가 나타나는 대로 적절한 국어 번역용어를 심의를 통해 제시해주는 것 이상의 방법이 없는 바, 이를 위해서는 새로운 용어를 확인하는 작업이 우선적으로 이루어져야 한다.

그런데 새로운 용어가 등장하는 것을 관찰하는 방법으로 말뭉치를 이용하는 것보다 좋은 방법이 없다. 결국 이는 정책적인 지원을 필요로 하는 것으로 문어 텍스트를 대상으로 지속적으로 신어를 추출해 내고 그를 대상으로 심의 작업이 진행되어야 하는 것이다.

3. 언어 교육에 있어서의 말뭉치의 활용

한 언어가 제2언어로서 교육 및 학습의 대상이 될 때, 어휘 교육 및 학습과 관련하여 제기되는 문제는 다음과 같이 정리할 수 있다(Carter & McCarthy 1988).

① 하나의 외국어를 익히는데, 어느 정도의 단어가 실용적인 어휘가 되는가?

② 어떤 단어가 우선적으로 배워야 할 단어인가?

③ 제2언어 학습의 초기 단계에 학습자에게 다른 단어보다 유용한 단어들이 있는가?

④ 단어들을 학습 난이도에 따라 등급화하는 것이 가능한가?

⑤ 다른 단어보다 배우기 어려운 단어라는 것이 존재하는가?

⑥ 어휘 학습의 가장 좋은 방법은 무엇인가?[4)]

⑦ 여러 가지 의미를 지니고 있는 단어는 어떻게 학습하는가? 그러한 단어들은 초기 단계의 학습 대상에서 제외하는 것이 좋은가? 그렇지 않다면 학습 대상이 되는 의미를 구분해 내야 할 것인가?

⑧ 문어에서보다 구어에서 접하기 쉬운 단어들이 있는가? 그렇다면 우리는 그것들을 알고 있는가?

이들 중에서도 특히 언어교육에 말뭉치를 활용할 때에 관심을 두어야 할 부분은 ① - ④의 네 가지인데, 이들은 대체로 제2언어 교육 특히 어휘 교육에 있어서의 어휘 통제와 관련된 것이라고 할 수 있다.

문제는 외국어로서의 한국어 교육과 관련해서 볼 때 이러한 관점에서의 연구가 전무한 상태라는 점이다. 한국어 학습에 있어서 일차적으로 익혀야 할

4) 이는 별도의 목록 안에 들어 있는 개별 항목을 학습하는 것이 좋은가 아니면 쌍을 이루는 단어와 함께(예를 들어 번역 대응어) 학습하는 것이 좋은가 혹은 문맥 속에서 학습하는 것이 좋은가 하는 문제이다.

단어 즉 기본어휘에는 어떤 것이 있는지, 그러한 기본어휘의 실질적인 용법은 어떤 것인지, 그들 어휘를 여하히 한국어 학습 교재에 수용할 것인지 등에 대해서 우리는 아직 별로 아는 것이 없다. 이러한 상황 하에서의 한국어의 국제화를 둘러싼 논의는 공허한 것일 수밖에 없다. 한국어 내지 한국학의 국제화와 함께 한국어 학습에 있어서의 어휘 학습의 중요성을 감안할 때, 이제 우리는 어휘 통제와 관련된 연구가 필요한 시점에 도달했다고 할 수 있는 것이다.

특히 기본어휘의 효율적 학습을 위해 익혀야 할 어휘의 용법과 의미에 관련된 제반 정보를 충실히 담고 있는 학습 도구로서의 기본어휘 빈도용례 사전 개발이나 한국어 학습에서의 어휘 통제에 관한 연구의 필요성에 적극적으로 대응하기 위한 방안으로는 외국인을 위한 ≪한국어 기본어휘 빈도용례 사전≫의 개발 이상의 것이 없다. 즉 제2언어로서의 한국어 교육에서의 어휘 교육의 효율화를 위해서는 기초적 어휘 통계 자료를 만들어 내고 그를 바탕으로 한 어휘 교육 자료의 체계화 하는 일이 필수적이라고 할 수 있는 바, 그 첫번째 작업이 되는 기본어휘의 선정을 위해서 제2언어로서의 한국어 학습에 필요한 기본어휘의 선정 방법론을 정립하고, 기본어휘 선정 방법론에 따라 필요한 기초자료를 확보하는 방법에는 어떤 것이 있는가를 검토한 후, 정리된 방법론에 따라 실제 자료를 구축하고, 구축된 자료를 분석하여 기본어휘를 선정하는 절차가 필요한 것이다.

2. 자연언어처리 측면에서의 말뭉치활용

이 장에서는 자연언어처리에서의 말뭉치를 활용한 연구의 제측면을 개관한다. 우선 언어처리시스템의 기본이 되는 형태소 분석 시스템 및 구문 분석 시스템 개발에 있어서의 말뭉치 활용을 살피고, 응용분야 중에서 가장 중요한 것의 하나인 기계번역에서의 말뭉치 활용에 대해서 검토해 보기로 한다.

1. 서론

자연언어처리에 있어서는 여러 분야에서 통계적인 자료를 필요로 하게 된다. 자연어 처리를 위한 문법 개발이나 자연어처리 시스템용 어휘부 개발, 단어 사이의 의미 관계 조사 등에 있어서 이전에는 사람이 직접 텍스트를 분석하고 필요한 정보를 분석하는 수밖에 없었다. 그러나 최근에는 정확한 통계자료를 바탕으로 한 분석 방법이 발전하고 있으며, 이와 아울러 분석 대상이 되는 문서의 양이 대량화함에 따라, 인간이 언어정보를 처리하는 것은 능률이나 정확성에서 한계에 부딪히고 있으며 컴퓨터를 활용한 정보 추출이 필연적인 것이 되고 있다.

말뭉치에서 통계정보를 구하기 위해서는 우선 말뭉치를 이루고 있는 언어단위, 즉 형태소, 어절, 구, 문장 등에 적절한 언어학적 정보가 부가되어 있어야 한다. 특히 자연언어처리에서는 어절이나 단어 단위에 적절한 형태•통사정보를 부가한 자료를 활용함으로써 단어나 품사간의 결합확율을 쉽게 구할 수 있는 것이다.

여기서는 이러한 자연언어처리에서의 말뭉치의 활용이 어떤 측면에서 이루어지는가와 말뭉치를 활용하기 위해서 필요한 작업들을 정리해보기로 한다.

2. 언어처리시스템 개발과 말뭉치의 활용

2.1. 서론

대규모의 말뭉치는 자연언어처리의 여러 분야에서 유용하게 사용된다. 예를 들어 문어를 구성하는 단위 중 가장 단순한 것인 국어의 낱글자 중에서 가장 많이 사용되는 자소가 어떤 것인가에 대한 정보나 가장 많이 사용되는 음절에 관한 정보도 코드 체계의 개발, 문서 압축 기술의 개발 등에 응용될 수 있어서 자연언어처리의 관점에서는 아주 긴요한 정보인 것이다. 그러나 본격적인 자연언어처리 시스템의 개발을 위해서는 이러한 간단한 정보로는 부족하다. 정확한 확률정보가 필요하기 때문이다. 즉 자연언어처리시스템의 개발을 위한 대부분의 통계정보는 말뭉치를 이루는 언어 단위별로 그 쓰임이 명확히 정의되어야 의미가 있기 때문이다.

이 때문에 유럽어권에서는 오래 전부터 말뭉치를 구축하고, 그를 기반으로 한 언어에 관한 통계정보 추출과 그러한 통계정보 추출에 사용할 시스템의 개발에 노력하고 있다.

2.2. 형태분석과 말뭉치

자연어의 입력 문자열로부터 문의 해석에 필요한 단위문자열로 잘라서 입력문에 대한 각종 지식정보, 즉 품사의 카테고리, 활용형 등 각종 정보를 추출해내는 과정을 형태소 해석이라 말한다.

영어의 경우는 단어의 구분이 분명하고, 어형변화의 부분도 -ed, -ing 또는 복수형 -s, -es 등 비교적 단조롭게 정형화 되어 있다. 그러나 한국어에서는 단어 단위의 분리가 일정하지 않고, 명사에 결합된 조사와 용언의 처리는 대단히 복잡하다. 조사의 종류도 다양하고 또 용언에서는 한 어간에서 수십개 이상의 파생어가 발생하므로 정확한 분석이 요구된다.

형태소 해석을 위해선 크게 4가지의 단위 모듈(module)이 필요하다.

① 언어적 최소의 단위인 형태소 단위로 사전을 구성

② 어절을 이루는 규칙성을 형태소 분석 문법으로 작성
③ 대상 어절로부터 사전참조와 문법검사를 통해 형태소 단위로 분석
④ 형태소 단위의 사전 정보로부터 어절에 대한 언어적 정보 합성

이렇게 어절을 형태소 단위로 분석하는 과정은 1990년대 초반까지는 주로 자소를 단위로 하는 방법이 주류를 이루고 있었다. 그러나 자소단위의 처리는 불필요한 사전의 탐색이 잦고 계산값이 많다는 부담을 안고 있다. 이를 극복하기 위해 근래에 들어서는 이른바 음절 정보를 이용한 형태소 해석이 주목받고 있다. 이는 국어의 형태소가 지니고 있는 음절 구조의 특성을 통계적으로 분석한 결과를 바탕으로 형태분석에 임하는 방법이다. 이론적으로 한글의 음절은 11,172개가 존재할 수 있지만(초성 19 * 중성 21 * 종성 28), 형태분석에서 문제가 되는 음절은 상대적으로 그 수가 적다는 점을 적극적으로 반영함으로써 사전 탐색에 소요되는 시간과 계산값을 가능한 한 줄이는 것이다.[5] 예를 들어 '철수는'이라는 어절이 입력된 경우 체언 중에는 "어느, 여느" 밖에는 "느"로 끝나는 것이 없고, "철, 수, 는" 중에서 조사의 초성으로 사용될 수 있는 음절은 '는' 밖에 없으므로[6] 여러 차례의 사전 검색을 거치지 않고도 "철수+는"으로 분석할 수 있다는 것이다.[7]

또한 최근에 들어서는 대형 코퍼스의 통계적 분석을 통해서 현대 한국어 문어 텍스트에서 출현 빈도가 높은 어절을 중심으로 별도의 어절 사전을 구축하여 두고, 해당 어절이 분석 대상이 되는 경우 매번 형태분석 과정을 거치지 않고 그대로 사전의 정보를 이용해 분석 결과를 내는 방법도 고려되고 있

5) 음절정보를 이용하는 경우 한 어절을 분석하는데 문법형태소 사전을 탐색한 회수는 평균 2.13회이고, 음절정보를 이용하지 않는 경우는 평균 5.27회인 것으로 보고되었다.(강승식 1993 : 131)
6) 국어 조사의 초성으로 사용되는 음절은 "가, 같,. 게, 고, 과, 까, 께, 나, 는, 니, 다, 대, 더, 도, 든, 라, 랑, 로, 를, 마, 만, 말, 며, 밖, 보, 부, 서, 아, 야, 에, 엔, 엘, 여, 와, 요, 으, 은, 을, 의, 이, 인, 일, 조, 처, 치, 커, 토, 하"의 48개이다.(강승식 1993 : 부록 5-1.)
7) 이때 "철수는"이 체언+조사 구성임을 파악하는 과정은 별개의 것이다.

다. 이러한 방법은 상대적으로 소수의 어절이 차지하는 텍스트 내에서의 비중이 크다는 데 근거하고 있다. [표 2][8]에서의 통계를 보면 약 144만 어절의 코퍼스에서 빈도수 101 이상의 어절 16,580 종(어절 종수 비율로는 전체의 약 7.78%에 불과하다)의 누적 빈도수가 전체 어절 수의 71.58%를 차지하는바, 동일 어절의 반복 비율이 상당히 큼을 알 수 있다. 이러한 특성을 이용하여 자주 사용되는 어절만 별도의 사전을 구성해서 어절의 분석 결과를 사전에서 그대로 가져다 쓸 수 있게 한다는 것이다.

어절의 빈도	어절 종수	어절 빈도	비율	누적빈도
500회 이상	262	341,068	23.72%	23.72%
301-500회	196	75,591	5.25%	28.97%
101-301회	1,220	201,370	14.00%	42.97%
51-100회	1,895	132,226	9.19%	52.16%
11-50회	13,007	279,315	19.42%	71.58%
5-10회	18,600	125,682	8.74%	80.32%
4회	8,628	34,512	2.40%	82.72%
3회	14,368	43,014	2.99%	85.71%
2회	32,028	64,056	4.45%	90.16%
1회	140,730	140,730	9.18%	99.34%
계	230,934	1,437,663	100.00%	100.00%

[표 1] 어절의 발생 빈도에 대한 통계

언론연구원의 신문기사 데이터베이스의 명사 색인어의 분석 결과를 보아도 이러한 분석은 상당한 근거가 있는 것으로 생각된다. 언론연구원의 KINDS 색인의 색인어 총 단어수는 796,039 항이고, 누적단어수 49,059,096 항으로 색인어당 평균 빈도는 61.6회인데, 실제 상위 빈도 10,000 어종의 누적빈도는 37,158,903 항에 달한다. 이는 전체의 1.3%에 불과한 명사류의 누적빈도가 총 누적 단어수의 75.7%를 점하고 있음을 보여주는 것으로 상대적으로 아주 적은 수의 단어들이 아주 빈번하게 쓰임을 의미한다.[9]

8) [표 2]는 김재한(1994 : 15)의 것을 빌어온 것이다.
9) 언론연구원의 색인어 구성에 대해서는 한영균(1994) 참조.

② 형태소의 원형 복원이란 어절 분석 단계에서 형태음소적 교체에 의한 이형태를 지니는 형태소가 나타나는 경우, 해당 형태소를 사전 등재형으로 복원시켜주는 과정을 의미한다. 이 경우 동일한 형태소의 원형 복원에서도 기술되는 변형 조건이 달라지는 경우가 있어서 처리가 간단치 않다. 그러나 이는 본질적으로 규칙에 의해 해결될 수 있는 문제다.

③의 옳은 후보를 선정하는 과정은, 형태소의 분석 과정에서 여러 가능태가 출현하는 데에서 생기는 문제를 해결하는 것을 의미한다.

한국어 어절이 지닐 수 있는 중의성은 ㉠ 형태 상의 중의성(한 어절이 최소한 두 가지 이상의 형태소 유형으로 분석될 수 있는 경우. 예를 들어 '감기는'이라는 어절은, 최소한 "감기(명사)+는(조사)", "감기(동사)+는(어미)", "감(동사)+기(어미)+는(조사)"이라는 세 가지 의미로의 분석이 가능하다). ㉡ 품사 중의성 (한 단어가 둘 이상의 품사를 가지는 경우.) ㉢ 의미 중의성 (이른바 동철이의어의 경우)의 세 가지가 존재하는데, 경우에 따라서는 이들 세 문제가 동시에 나타나기도 한다(ex. 줄(絃), 줄(주+ㄹ), 줄(줄+ϕ)).

또한 중의성에는 문장 단위의 문맥에 의해 해소될 수 있는 것과 그렇지 않은 것의 두 유형이 있다. 즉 다음의 (2)㈎, ㈏에서 (2)㈎의 중의성은 문장단위의 정보로도 해소될 수 없는 중의성을 지닌 것이며, (2)㈏는 문장 단위의 정보만으로도 해소될 수 있는 것이다.

(1) ㈎ 나는 새를 보았다.
㈏ 푸른 하늘을 나는 새를 보았다.

한국어 형태소 해석기 개발의 초기 단계에는 (2)㈏과 같은 중의성 문제도 거의 해결할 수 없었지만, 근래에 들어서는 대형 코퍼스의 구축을 통해서 통계 정보를 확보한 결과 어절 분석시의 중의성을 상당히 해소할 수 있게 되었다.[10]

10) 김병희 외(1993)에서는 약 95% 형태소 분석 성공률을 보였다고 보고하고 있다. 그러나 이는 중의성을 완전히 해소한 결과의 비율이 아니라, 형태소 분석기가 제시한 가

2.3. 구문해석과 말뭉치

구문해석은 입력된 문장에 대해서 형태소를 분석한 결과로부터 문법규칙 및 어휘정보를 조사하여 문장의 구조를 해석하는 과정이다.

현재 한국어 구문해석에 대한 연구에서 가장 많은 연구 기반으로 삼고 있는 문법은 의존 문법이다. 이는 한국어가 자유 어순을 가지고 있고 교착어로써 기능어가 문장에서의 어절의 기능을 지시하기 때문이다.

과거의 연구에서는 구문해석의 문제는 문법을 설정하고 설정된 문법이 올바르며, 이를 이용해 구현된 구문해석기에 의해 문장이 얼마나 정확히 해석되는지의 문제였다. 그러나 실제 자연어 처리 응용 시스템을 위해서는 가장 올바르다고 판단되는 하나의 결과가 산출되어야 한다. 그래서 최근의 연구는 최적의 해를 결정하기 위한 중의성 해결의 문제가 초점이 되고 있다.

한국어는 교착어로서 어휘 형태소와 기능 형태소가 결합되어 하나의 어절을 이룬다. 그러므로 한국어의 어절은 그 자체가 구문 구조를 가지고 있는 형태이다. 따라서 어절간의 관계를 파악하는 것은 구문 구조간의 관계를 밝히는 것이라고 볼 수 있다. 다음 <예시 1>은 의존문법의 방법론에 따라서 "미래에 마음을 두는 자에게는 그 이상 더딘 것이 없고, 오늘을 즐기는 자에게는 그 이상 빠른 것이 없다."는 문장을 분석한 것인데, 문장을 구성하고 있는 각 요소 사이의 관계가 나무 그림 형식으로 나타나 있다. 이 나무의 연결 부분이 결국 문장을 구성하는 요소들 사이의 관계를 드러내는 것이라고 할 수 있는데, 각 연결점(node)의 문법적 성분과, 그들 사이의 관계에 대한 통계적 정보와 국어 문장을 이루는 구성 성분의 계량적 분포에 대한 정보가 확보되면, 국어 구문 분석기의 개발에 적극적으로 기여할 수 있을 것이다.

((((((미래/nct+에/jca 마음/nc+을/jc 두/pv+는/exm) (자

능태 중에 옳은 분석 결과가 포함된 비율로 생각된다. 이 문제에 대해서는 5장에서 다시 다룬다.

/nbu+에게는/jca ((((그/md 이상/nc) 더디/pa+ㄴ/exm) 것/nb+이/jc) 없/pa+고/ecq) ,/s, 오늘/nct+을/jc 즐기/pv+는/exm) 자/nbu+에게는/jca 그/md 이상/nc) 빠르/pa+ㄴ/exm) 것/nb+이/jc) 없/pa+다/ef) ./s.)

<예시 1>

```
;                                +---미래/nct+에/jca
;                                |---마음/nc+을/jc
;                          +---두/pv+는/exm
;                          |     +---자/nbu+에게는/jca
;                          |     |                  +---그/md
;                          |     |              +---이상/nc
;                          |     |          +---더디/pa+ㄴ/exm
;                          |     |     +---것/nb+이/jc
;                          |     |---없/pa+고/ecq
;                          |     |---,/s,
;                          |     |---오늘/nct+을/jc
;                          |---즐기/pv+는/exm
;                          |---자/nbu+에게는/jca
;                          |---그/md
;                     +---이상/nc
;               +---빠르/pa+ㄴ/exm
;          +---것/nb+이/jc
;+---없/pa+다/ef
;./s.
```

3. 기계번역과 말뭉치

기계번역(machine translation)은 번역의 자동화를 추구하는 자연언어처리의 대표적인 응용분야의 하나다. 최근 들어, 지능형 컴퓨터와 지식 정보화 사회에 대한 비전이 제시되면서, 그 핵심적인 요소 가운데 하나가 바로 자연언어처리라는 인식이 확산되고 있고, 컴퓨터의 비약적인 발달과 더불어 자연언어처리 기술이 크게 향상되어서 실용적 수준의 기계번역 시스템에 대한 시장 전망을 밝게 해주는 요인이 되고 있다. 또 인공지능 기술이 발달하고 언어학 분야의 이론도 많은 발전을 이룩하였으며, 이러한 이론적 배경의 강화는, 그 동안의 연구에서 축적되고 향상된 자연처리기술과 함께 새로운 발전을 가능하게 할 요인으로 평가되고 있다.

기계번역은 컴퓨터가 하나의 언어를 다른 하나의 언어로 변환하는 것이며 이를 위해서는 번역 원문을 해석하고 문장의 의미를 이해하여 번역문장을 생성하여야 하는 것이다. 따라서 기계번역이란 자연언어처리분야의 기본 기술을 망라한 하나의 응용분야인 동시에 인공지능 연구분야의 기본단계라고 말할 수 있다.

자연어문장을 다룰 때는 기본적으로 크게 두 가지의 입장에서 생각할 수 있다. 첫째는 입력이 자연어문이고, 이것을 문법규칙, 의미소성을 사용하여 해석하고, 그 문장이 문법적으로 타당한가를 판단하던가, 그 문법의 구조를 표현하던가, 또 문장의 의미를 기술하던가, 그 의미에 따른 처리를 하는 일이다. 둘째는 의미의 기술이나, 해석 데이터로부터 그것에 적합한 자연어를 생성하던가, 추론하는 일이다. 이러한 문제들의 첫 단계가 자연어 문장의 해석이다. 자연언어처리에서 의미해석이란 단어 또는 문장이 나타내고자 하는 의미를 의미자질(semantic feature), 격(case), 의미표현(semantic representation), 논리(logic)등의 수단으로 표현하는 것이다.

기계번역에서 의미해석은 다음 두 가지 경우에서 특히 그 필요성을 알 수 있다.

첫째는 구문해석은 되나 의미해석이 안 되는 문장을 볼 수 있다. 예를 들

어 '그가 배를 샀다'라는 문장은 구문해석은 되나 '배'의 의미를 이 문장만 가지고서는 'pear'로 해석할지 'boat'로 해석할지 알 수 없다. 이 경우는 위 문장의 앞. 뒤 문맥을 살펴보고 이를 결정하여야 한다.

둘째는 구문해석의 결과가 모호한 문장이다. 예를 들어 'Harry borke the glass with the hammer'와 같은 문장은 뒤에 위치한 전치사구(with the hammer)가 동사(broke)를 수식하는지 명사(glass)를 수식하는지 구문해석의 단계에서는 알 수가 없다. 이렇게 모호한 구문 해석 결과를 의미해석 단계에서 해결해야 한다.

규칙기반 기계번역에서, 숙어적 표현과 같이 어휘 치환과 구조 변형이 복합적으로 수반되는 표현을 처리하기 위해서는 복잡하고도 많은 변환 규칙이 필요하다. 그런데 규칙은 서로 간에 간섭 현상이 심하기 때문에, 대량의 복잡한 변환 규칙을 개발하는 것은 가장 큰 문제점이다. 규칙 기반 기계 번역이 지니고 있는 이와 같은 문제점을 극복하기 위해, 최근 새로운 방법이 시도되고 있는데, 이에는 예문기반 기계번역(example-based machine translation), 숙어기반 기계번역(idiom-based machine translation)등이 있다.

예문기반 기계번역에서는 대역 말뭉치(bilingual corpus)로부터 추출된 번역 예문을 참조하여, 원문과 번역 예문사이의 유사 일치(similar match)를 통해 변환이 이뤄지고, 유사 일치시에 어휘들 사이의 의미적 유사성을 평가하기 위해 시소러스(thesaurus)를 사용한다.

숙어기반 기계번역에서는 단순 어휘치환이나 구조 변형으로는 번역할 수 없는 표현을 숙어(idiom)로 간주하고, 원문에서 이러한 숙어적 표현을 파악하여 그 대역어를 결정하는 숙어인식(idiom recognition)에 중심을 두고 변환이 진행된다. 그리고 연어(colloction)관계에 기초한 변환도 함께 이뤄지는데, 예문기반 기계번역과는 달리 숙어인식에서 완전 일치(exact match)를 사용하기 때문에 시소러스(thesaurus)가 꼭 필요한 것은 아니다.

3. 사전편찬학 측면에서의 말뭉치 활용

이 장에서는 사전편찬 특히 학습자 사전의 편찬에 있어서의 말뭉치 활용의 문제를 거시구조 구성의 측면과 미시구조 구성의 측면 두 가지로 나누어 살펴본다.

1. 서론

지난 10여 년 사이 국내에서의 사전 연구 및 사전 편찬과 관련된 움직임은 괄목할 만한 진전을 보이고 있다. 영역별로 특수한 사전(학습자 사전 동사 구문 사전 조사 사전 관용구 사전 용례 사전 갈래 사전 고유어 사전 시어 사전 작가 사전 고어 사전 등)들이 속속 새로 간행되었거나 간행을 준비 중인가 하면, 사전 편찬에 컴퓨터와 말뭉치가 도입되어 컴퓨터를 활용한 말뭉치 분석 결과를 반영한 사전이 만들어지고 있는 것이다.

그러나 그러한 변화가 사전학의 이론적 발전을 의미하지는 않는 것으로 보인다. 이제 얻고 있는 가시적 업적들은 이미 이루어졌어야 할 것들이 보충된 것이거나 종래의 언어학적 사전학적 틀 속에 말뭉치 활용이라는 새로운 방법을 도입한 것일 뿐, 오늘날의 사전 나아가 앞으로의 사전이 지향해야 할 방향에 대한 이론적 탐구나 그러한 이론을 바탕으로 한 실천과는 거리가 있는 것으로 생각되기 때문이다.

사전학의 발전을 위해서는 사전학 관련 분야에 대한 전문 교육의 활성화 등 저변 확대를 위한 노력과 함께, 오늘날의 사전과 사전 편찬 방법론이 환경의 변화와 학문의 발전에 걸맞게 변화해야 한다는 사실을 정확히 인식하고, 그러한 인식을 바탕으로 새로운 이론을 정립해야 할 시점에 도달했다. 이것이 비로 말뭉치를 활용한 사전 편찬인 것이다.

2. 거시구조 구성에 있어서의 말뭉치 활용

2.1. 표제항의 적정 규모

표제항의 적정 규모라는 문제는 기본적으로 사전이라는 상품을 생산하는 과정에서 제기되는 것이라고 할 수 있다. 이런 관점에서 본다면 상업적인 문제를 떠나서 순수히 학문적인 관점에서는 표제항의 적정 규모를 결정하는 것이 대사전과는 무관한 것으로 생각하기 쉽다. 그러나 그렇지 않다. 대사전이라고 하더라도 국어의 모든 어휘를 표제항으로 담을 수는 없는 것이고, 따라서 적절한 규모와 내용을 결정할 필요가 있는 것이다. 그러나 여기서는 기본적으로 학습용 사전의 표제항목의 크기와 관련된 문제를 다룬다. 아울러 표제항목의 크기를 다룸에 있어서는 어휘형태소를 중심으로 한다. 기능형태소는 가능한한 모두를 사전의 표제항으로 삼는 것이 옳다고 판단되기 때문이다

2.1.1. 어휘 총량과 학습용 사전의 표제항 수

학습자용 사전의 표제항의 크기가 어느 정도여야 하는가는 기본적으로 국어 어휘의 총량 및 한국어 습득에 필요한 어휘의 양을 결정하는 문제와 밀접한 관계를 가지고 있다. 그러나 국어 어휘의 총량을 정확히 측정한다는 것은 불가능하다. 사전에 등재된 표제항의 수, 실제 문헌에서 사용되는 어휘의 양 등을 참조할 수 있을 뿐이다. 따라서 학습용 사전의 표제항 수는 이러한 점들을 고려하여 결정되어야 할 것이다.

일반적으로 초/중/고등학생의 학습에 요구되는 어휘의 수와 일반인 혹은 외국인에게 요구되는 생활상의 기초어, 한국학의 학습을 위해 요구되는 어휘는 그 분포와 범위가 달라질 수 밖에 없다. 또한 한국의 사회문화, 인문지리, 자연지리, 역사, 정치/경제 등에 관한 어휘는 별도로 고려되어야 한다.

한국어 사전에 등재된 표제항의 수는 사전의 종류에 따라서 달라진다.

- 한 언어 사전의 경우
- 두 언어 사전의 경우

- 외국인을 위한 한국어 사전의 경우
- 한국인을 위한 두 언어 사전의 경우(한 -> 영 등)

한 언어사전(국어사전의 경우)도 대사전, 중사전, 소사전의 경우로 나누어 검토할 수 있다.

□ 실제 문헌에서 사용되는 어휘의 양

●일간지의 경우

- 하루분 신문의 어휘 총량
- 한달분 신문의 어휘 총량
- 일년분 신문의 어휘 총량

●잡지의 경우

- 한달치 시사종합지의 어휘량
- 한달치 주부잡지의 어휘량

●단행본 : 정보전달적 단행본과 문학작품의 경우로 대별할 수 있으나, 최소한

- 문학작품의 경우
- 일반인을 대상으로 한 서적의 경우
- 학술서적의 경우
- 종교서적의 경우

등으로 나누어 검토할 수 있다.

●교과서의 경우

- 국민학교 교과서의 어휘 총량
- 중학교 교과서의 어휘 총량 : 국어/사회/도덕/기타
- 고등학교 교과서의 어휘 총량 : 국어/사회문화/국사/정치경제/기술/기타

2.1.2. 한국어 어휘부의 구성과 사전 표제항의 구성

어휘 구성은 국어의 어휘부에 등재되는 어휘 항목의 품사별/기원별/형태구성별 분포를 의미한다.

□ 기원별 분포는 한자어, 외래어, 고유어
□ 품사별 분포는 명사/동사/형용사/관형사/수사/감탄사 등,
□ 형태구성별 분포는 단일어/복합어/파생어 등
□ 용법(범주)별 분포로는 ● 보통명사/고유명사, ● 일반어/학술용어

등으로 나뉘는데 어휘구성에 대한 정확한 정보는 한국어 어휘 교육 및 사전 편찬에 중요한 의미를 갖는다. 사전의 표제항 구성은 기본적으로 어휘구성 비율에 대한 고려를 바탕으로 할 것이기 때문이다.

2.2. 표제항이 되는 언어 단위의 선정

2.2.1. 기능 형태소

기능형태소란 조사 어미 등의 문법형태소와 파생접사류를 가리킨다. 이들 기능형태소는 어휘형태소와는 달리 폐쇄집합을 이룬다. 이런 관점에서 기능형태소는 그 분포 및 의미 기능에 대한 정밀한 기술과 함께 가능한한 빠짐없이 사전에 등재할 필요가 있다. 그러나 이러한 기능형태소의 사전적 처리에는 간단하지 않은 문제들이 있다.

첫째, 기능형태소 중에서 조사와 어미의 경우 파생접사류와는 달리 여러 형태소가 결합한 복합형태가 많다. 이들 복합형태를 사전의 표제항으로 채택할 것인가 여부가 문제로 제기될 수 있다. 특히 복합형태 중에는 단순히 어미나 조사로만 이루어진 것 외에도 어미에 조사가 결합한 형태, 문법적 기능이 전혀 다른 조사가 결합한 형태 등이 적지 않아서 이들을 처리하는 기준의 설정이 필요하다.

둘째, 국어의 맞춤법은 기본적으로 형태음소적 표기를 원칙으로 하고 있지만. 조사나 어미와 같은 문법 형태소의 경우는 이형태를 그대로 표기에 반영하는 것을 원칙으로 하고 있다. 따라서 문에 텍스트의 이해를 위한 사전에서는 이들 이형태에 관한 정보를 어떤 방법으로든 제공할 필요가 있다.

2.2.2. 한 어기 어휘소와 여러 어기 어휘소

어휘 형태소 중에는 여러 어간으로 이루어진 단어가 적지 않다. 특히 명사와 동사는 이른바 복합어가 단일어보다 훨씬 다양하게 사용된다고 할 수 있다. 이들 복합어 중에서 어떤 것을 사전의 표제어로 등재할 것인가는 간단치 않은 문제를 남긴다.

2.2.3. 파생어

문법의 기술이라는 관점에서도 파생어가 어휘부의 등재소가 될 수 있는가 하는 문제는 분석자의 관점에 따라서 상당한 편차가 있다. 그러한 순수한 문법 기술의 관점에서 뿐만 아니라 실질적인 사전 편찬의 관점에서는 훨씬 더 복잡한 문제를 내포하고 있다. 기본적으로 파생어는 사전의 중요한 구성 요소가 된다. 파생어의 생성규칙을 알지 못하는 사람에게 그것을 학습할 수 있도록 해 주는 것이어야 하기 때문이다.

2.2.4. 성구와 속담

성구와 속담은 관용어라는 용어와 함께 언어학적 정의가 쉽지 않은 단위들이다. 특히 성구 중에서 이른바 고사성어 내지 한자성어라고 지칭되는 것의 처리와 속담의 한계를 정하는 문제, 성구와 연어의 구분 배열 방식 등이 검토되어야 한다.

2.2.5. 약어와 외래어

외래어의 처리 문제는 현존하는 국어사전의 평가에 있어서 가장 큰 문제중의 하나로 지적되어 온 것이다. 특히 이 경우가 말뭉치를 분석한 결과를 활용할 때에 가장 바람직한 결과를 얻을 수 있는 것이라고 할 수 있다.

2.2.6. 고유명사와 보통명사

일반적으로 언어사전에는 고유명사를 포함하지 않는 것이 원칙이라고 할 수 있다. 국어사전이 백과사전적 요소를 많이 지니고 있다는 점이 국어사전의 가장 큰 단점으로 지적되어 온 것이다. 그러나 실제 백과사전적 요소가 반드시 국어사전에서 배제되어야 할 것인가는 단정하기 어려운 문제다. 사용자들은 국어사전에서 무엇을 요구하고 있는가가 검토되어야 하는 것이다. 이런 점에서 사용자의 요구 분석이 제대로 이루어지지 않은 것을 문제점으로 지적할 수 있을 것이다.

2.2.7. 어휘화한 단위의 처리

어휘화한 단위란 원래는 둘 이상의 형태소 특히 어휘형태소와 문법형태소의 결합으로 이루어진 단위(우리는 흔히 그것을 어절이라고 부른다)가 특정한 용법을 지니게 됨으로써 한 단위로 인식되거나, 경우에 따라서는 하나로 처리하는 것이 국어의 이해에 더욱 편리하게 된 것을 가리킨다.

특히 부사의 경우는 이러한 형태로 이루어진 것들이 많다. 명사의 곡용형이나 동사의 활용형이 부사어로의 용법이 굳어지면서 부사로 전용되는 경우가 많은 것이다.

제2부

한국어 말뭉치의 현황과 활용도구

1. 한국어 말뭉치의 현황

이 장에서는, 현재 국내에서 구축, 개발되어 활용되고 있는 주요 말뭉치 가운데, 그 내용이 공개되어 있거나 실제로 얻어 쓸 수 있는 말뭉치를 소개한다. 이를 통해 특히, 말뭉치의 구축의 목적을 비롯하여, 규모, 구성, 활용 방법, 자료의 여부 등의 제반 요소에 대한 정보를 알 수 있다.

1. 한국어 말뭉치 구축의 현황

1.1 연세 한국어 말뭉치(Ysc :Yonsei Korean Corpora)

【구축자】 연세대학교 언어정보개발원구원 (이전 한국어사전편찬실)

【구축 기간】 1987년~1999년 현재

【목적】 연세 한국어 말뭉치 1~9, 표준말뭉치 등은 "연세 한국어 사전"(1998.10.발간)의 편찬을 위한 목적으로 구축되었다. 1990년대 중반 이후부터는, 주로 언어학적 연구와 각종 특수 사전의 편찬, 초등 교육 및 외국어로서의 한국어 교육 등의 기초 어휘 사전 등의 편찬을 위한, 각종의 연세 한국어 특수 말뭉치가 구축되고 있다.

【원 자료 성립 시기】 1960년대~1990년대

【규모】 연세 한국어 말뭉치 1~9(9종, 약 4,300만 어절), 표준 말뭉치(2종, 약 2,900만 어절), 특수 말뭉치(8종, 약 2,500만 어절), 품사 표지 부착 말뭉치(2종, 180만 어절), 의미 표지 부착 말뭉치(1종, 100만 어절)

【입말/글말】 "연세 한국어 말뭉치 4, 세종 입말뭉치1999" 외에는 모두 글말(문어), 버금입말(준구어) 등을 중심으로 이루어져 있다.

【구성】 연세 한국어 말뭉치는 『연세 한국어 사전』(1998)의 편찬에 이용된 "연세 한국어 말뭉치 1-9"(약 4,200만 어절)를 비롯하여, 최근 개발되어 활용되고 있는 각종 특수 말뭉치, 품사 표지 부착 말뭉치, 의미 표지 부착 말뭉치 등으로 구성되어 있다.

【주석】 특수 말뭉치 중, "세종 말뭉치"는 TEI 헤더로, 그밖의 말뭉치에는 자체 개발한 고유한 태그로 서지 정보와 쪽 정보를 부착하고 있으며, 1998년 말부터 품사 표지 및 의미 표지 주석이 이루어지고 있다.

【응용 사례】 가장 직접적으로는 『연세 한국어사전』(1998) 편찬에 이용되었으며, 언어학의 제 분야 연구는 물론, 음성 인식, 문서 인식, 자연어처리 등의 여러 분야에서 다양하게 이용되고 있다.

【관련 자료】 정기 간행물인 『사전편찬학 연구』(2000년 8월 현재 9집까지 간행)와 『"언어 정보의 탐구』('99년 5월 창간), 말뭉치 기반 국어 연구 총서(2000년 8월 현재 7종 발간) 등에서 실제 말뭉치를 이용한 구체적 연구 결과를 볼 수 있으며,

http://clid.yonsei.ac.kr/에서 상세한 정보를 제공하고 있다.

종 류	말뭉치의 이름	내용의 구성	규모(어절)
표준 말뭉치	표준 말뭉치 A	말뭉치1-9를 토대로, 가능한 여러 장르와 분야의 표본들이 고르게 분포할 수 있도록 재구성하여, 용례 추출의 효율을 높인 말뭉치.	1,160만
	표준 말뭉치 B		1,788만
균형 말뭉치	연세 말뭉치1	연세 말뭉치 중 가장 이르게 구축된 말뭉치로, '현대 한국인의 독서 실태'에 관한 설문 조사에 의해 구축된 말뭉치. 표본의 대표성을 확보하고 표본 선정 기준을 마련하기 위해, '전문가 조사'에서는 교과서의 배제 또는 표집 비율 조정, 문제 번역물의 표집과 관계된 문제, 독서물의 유목 분류 문제에 대한 조사가 이루어졌으며, '일반인 조사'에서는 전문가 연구의 결과를 토대로 독서물 유목을 작성하여 독서물 유목별 독서량을 조사한 결과를 토대로 구축된 균형 말뭉치. 샘플 텍스트 말뭉치. 자세한 것은, 정찬섭 외(1990:7-70)를 참조 바람.	290만
	연세 말뭉치2	도서 대출 빈도에 따른 균형 잡힌 말뭉치를 구성할 목적으로 주제별 선정 기준을 사용함. 문헌정보학적 방법을 응용한 것으로, 가능한 모든 주제 분야의 낱말이 망라되도록 하기 위해 듀이 십진 분류법(Dewey Decimal Classification)을 이용하여 한국어 문헌을 크게 열 가지 대범주(총류, 철학, 종교, 사회, 과학, 언어학, 순수과학, 응용과학, 예술, 문학, 역사 등의 분야)로 나누어 수집함. 또한 도서의 대출 빈도는 각 낱말의 인지도를 간접적으로 나타낸다는 전제 아래, 주로 1987-1988년의 도서를 대상으로 대출 빈도가 높은 자료를 중심으로 하여, 234개의 표본을 선정하여 1990년~1991년에 걸쳐 구축됨. 주제별 구성은, 총류(7.8%), 철학(9.9%), 종교(10.7%), 사회과학(12.8%), 언어(5.7%), 순수과학(11%), 응용과학(11.7%), 예술(8.1%), 문학(11.2%), 역사(11.3%) 등으로 구성되어 있음.	110만
합계			3,058만

[표2] '연세 한국어 말뭉치 1-9'의 구성

종 류	말뭉치의 이름	내용의 구성	규모(어절)
시대별 말뭉치	연세 말뭉치3 (1980년대)	사람들이 많이 접하는 문헌이 각 낱말의 인지도를 간접적으로 반영한다고 보고, 1980년대 우수 출판물 목록에서 표본을 선정.	598만
	연세 말뭉치4 (1990년대 입말뭉치)	실제 사용된 입말을 녹음하여 전사하는 '순수 입말'(구어)과, 희곡, 방송 대본, 시나리오 따위를 중심으로 한 '버금 입말'(준구어)로 구성. 대화(26%), 강연(24%), 상담(14%), 희곡·대본(13%), DJ방송(13%), 토론(8%), 회의(2%) 등의 내용으로 구성되어 있다. 다른 글말뭉치와는 달리, 여기에는 발화참여자의 나이, 성별, 직업에 대한 정보, 발화참여자의 수와 발화의 성격, 전사자 정보, 녹음 시간 정보 등이 덧붙여져 있다.	77만
	연세 말뭉치5 (1970년대)	1970년대 문헌을 대상으로 하여, 교과서에서 신문에 이르는 다양한 문헌 자료로 구성. 1970년대 문헌을 대상으로 하여, 신문(10%), 소설·수필(50%), 일반서적(35%), 교과서(5%) 등의 문헌 자료를 수집함으로써 구축됨.	860만
	연세 말뭉치6 (1960년대)	해방 이후의 우리말의 모습을 제대로 반영한 말뭉치 구축을 위해, 1980년대 이전을 10년 단위로 거슬러 올라가면서 말뭉치를 구축하기 시작하였으며, 1000만 어절을 목표로 하고 있음.	723만
	연세 말뭉치7 (1990년대)	1990년대 중반까지의 자료로,1994년부터 1995년에 걸쳐 구축됨..	1,367만
	연세 말뭉치8 (교과서)	초등학교 전 과목 및 중·고등학교의 국어, 사회를 중심으로 구성. 제5차(한영균 말뭉치), 제6차 교과 과정의 두 종류.	87만
	연세 말뭉치9 (아동 도서)	순우리말의 사용 빈도가 높고 문장도 비교적 우리말의 구조에 맞게 사용되는, 아동 교육용 도서 중에서 적합한 자료를 선정하여, 1996년에 구축됨.	150만
합계			3,862만

[표3] '연세 한국어 말뭉치 1-9'의 구성

종 류	말뭉치의 이름	내용의 구성	규모(어절)
특수 말뭉치	인용구 말뭉치	1910년대에서 1970년대까지의 주요 문학 작품을 선정하여 구성.	350만
	현대시 말뭉치 (한영균 말뭉치)	'한국 대표 시인 100인 선집'을 입력, 운문으로만 구성한 말뭉치.	85만
	기초 학습 말뭉치1	기초(초등) 학습용 사전 편찬을 목적으로, 교과서(5%), 아동신문(10%), 아동잡지(10%), 백과사전(15%), 아동 창작 작품(5%), 만화(5%), 동화·소설·전기(35%) 등의 글말과, 아동 방송용 시나리오(10%), 기타 보충용(5%) 등의 구성으로, 1,000만 마디를 목표로 하며, 1차분 20%를 구축 완료.	177만
	한국어교육 말뭉치1	외국어로서의 한국어 교육을 위한 기초 어휘 조사를 위해, 소설, 수필, 수기/기행, 논설 등의 일반 교양 텍스트에, 한국어 교재, 초등 교과서를 중심으로 하고, 입말, 버금 입말 등을 더해 구성.	100만
	세종 말뭉치 1998	문화관광부 주관 '21세기 세종 계획'의 '국어 기초 자료 구축' 과제 중 본 연구원 수행의 결과물. 글말, 버금 입말, 입말의 세 부분으로 이루어져 있음. TEI 헤더 부착.	825만
	전문용어 말뭉치1	경제 분야 용어 연구를 위해 구축된 것으로, 용어 해설 자료와 교재, 논문, 일반 자료로 구성됨. 내용 면에서 경제 전반, 증권, 금융, 국제 경제, 경영 등의 세부 분야로 구성됨.	140만
	세종 글말뭉치 1999	문화관광부 주관 '21세기 세종 계획'의 '국어 기초 자료 구축' 과제 중, 본 연구원 수행의 결과물. 입말뭉치 구축이 독립 과제로 분리되면서, 글말과 버금 입말만으로 이루어짐. TEI 헤더 부착.	800만
	세종 입말뭉치 1999	문화관광부 주관 '21세기 세종 계획'의 '국어 기초 자료 구축' 과제 중, 본 연구원 수행의 결과물로, 화면 있는 방송과, 라디오 등 화면 없는 방송, 기타 강의나 강연, 상담 자료 등의 일상 대화를 녹화(녹음), 전사한 자료로 구성. TEI 헤더 부착.	75만
합계			2,552만

[표4] '연세 한국어 특수말뭉치'의 구성(1)

종 류	말뭉치의 이름	내용의 구성	규모(어절)
품사 표지 부착 말뭉치	초등 교과서 말뭉치	기초 학습 말뭉치1 중, 초등학교 전 과목의 교과서에, 자동 태깅, 수동 태깅의 두 단계에 걸쳐, 품사 표지를 부착한 말뭉치.	81만
	한국어교육 말뭉치2	한국어 교육 말뭉치1에, 자동 태깅, 수동 태깅의 두 단계에 걸쳐, 품사 표지를 부착한 말뭉치.	100만
의미 표지 부착 말뭉치	한국어교육 말뭉치3	한국어교육 말뭉치2에, "연세 한국어 사전"의 의미 기술을 기준으로 하여, 기초 어휘 1,000개에 의미 표지를 부착한 말뭉치. '언어정보개발연구원 제 2차 어휘빈도 조사'의 상위 1000 어휘를 대상으로 작업이 진행중이며 2000년 말에 구축 완료 예정임.	100만

[표5] '연세 한국어 특수말뭉치'의 구성(2)

【날말뭉치(일부, 연세 한국어 말뭉치9)】

Ysc9_0011.txt
<ㅍㅂ><ㅅㅈ><ㅈ>811.37</ㅈ>
<ㅈㅈ>호영송</ㅈㅈ>
<ㅈㅁ>호영송소설집(파하의안개외7편)</ㅈㅁ>
<ㅊㅊ>문학과지성사</ㅊㅊ>
<ㄴㅇ>19781100</ㄴㅇ></ㅅㅈ><ㅈㄹ>
<ㅈㄱ=파하의 안개><ㅈㅈ=7>
말은 고작 엉성한 그물에 불과합니다. 그 그물로 행의 또는 행위의 진실을 건지려고 했을 때 잡으려는 고기는 그물 사이로 빠져 나가 버리고 마는 것입니다.
손을 그릇삼아 샘물을 떠마셔 본 적이 있으시죠? 물은 이미 손가락 틈으로 다새어 나가고 아주 조금의 물이 입술과 혀를 축여 줄 뿐 아닙니까. 네? 말의 유희라고요? 그럴지도 모르겠읍니다. 그 점은 부인하지 않겠읍니다. 혓바닥과 입술을 나불대고 또는 종잇장 위에 펜으로 뭘 좀 끄적끄적해 보았자, 고트프리트 벤이 그의 시행을 재주껏 전개시켜 보았자 그의 시행들은 그의 독자들에 의해서 말장난이라고 걷어채리고 마는 일도 나는 잘 알고있으니까요.
그럼 고트프리트 벤의 정신의 진실은 어디 있겠으며, 그의 행위와 삶의 궁극적 가치는 어떻게 확인할 수 있겠읍니까. 절망적이지요. 말의 유희에 떨어지지 않고 혹은 말려 들지 않고 행위의 진실, 사실의 사실됨을 선생께 전당한다는 일은 불가능할 것만 같습니다.
여하튼 계속하라구요. 그리고 내 말은 어떻든간에 선생이 나의 행위의 진실을

파악하는데 도움이 된다구요. 아, 술요? 그 술을 조금 주시렵니까. 아, 혀를 쏘는 맛이 그럴 듯하군요.</ㅈ><ㅈ=8>많이는 못 합니다. 그만하면 됐읍니다. 네?

【품사 표지 부착 말뭉치(일부)】

00001: <ㅈㄹ=승구의 일기_ssk>
00002: <note>서상규 기초학습말뭉치</note>
00003: <Date>2000/06/23</Date>
00004: @
00005: 이 이 AN
00006: 일기는 일기 NN 는 PA
00007: 우리 우리 NP
00008: 반의 반 NN 의 PA
00009: 조승구라는 조승구 NNX ㅣ VC 라는 EF
00010: 6학년 육# NU 학년 NX
00011: 남자아이가 남자아이 NN 가 PA
00012: 쓴 쓰 VV ㄴ EF
00013: 것입니다 것 NX 이 VC ㅂ니다 EF
00014: . . SP
00015: @
00016: 승구는 승구 NNX 는 PA
00017: 1990년 1990 NU 년 NN
00018: , , SP
00019: 불의의 불의 NN 의 PA
00020: 교통 교통 NN
00021: 사고로 사고 NN 로 PA
00022: 머리를 머리 NN 를 PA
00023: 다쳐 다치 VV ㅓ EF
00024: 뇌 뇌 NN
00025: 수술을 수술 NN 을 PA
00026: 두 두 NU
00027: 번이나 번 NX 이나 PA
00028: 받았습니다 받 VV 았습니다 EF
00029: . . SP
00030: @
00031: 수술 수술 NN
00032: 후 후 NX
00033: 회복 회복 NN

00034:	기간을	기간	NN	을	PA
00035:	거쳐	거치	VV	ㅓ	EF
00036:	승구가	승구	NNX	가	PA
00037:	다시	다시	AV		
00038:	학교에	학교	NN	에	PA
00039:	나왔을	나오	VV	ㅏㅆ을	EF
00040:	때	때	NX		
00041:	,	,	SP		
00042:	승구의	승구	NNX	의	PA
00043:	머리	머리	NN		
00044:	한가운데에	한가운데	NN	에	PA
00045:	어린애	어린애	NN		
00046:	주먹	주먹	NN		
00047:	하나가	하나	NU	가	PA
00048:	들어갈	들어가	VV	ㄹ	EF
00049:	정도로	정도	NX	로	PA
00050:	움푹	움푹	AV		
00051:	패인	패이	VV	ㄴ	EF
00052:	자국이	자국	NN	이	PA
00053:	있었다면	있	VJ	었다면	EF
00054:	믿겠습니까	믿	VV	겠습니까	EF
00055:	?	?	SP		
00056:	@				

【의미 표지 부착 말뭉치(일부)】

ID	어절번호	어절	어절내 순서	형태소	품사	의미[11]	참고
1	1	<ㅈㄹ=CA000206-SSK>					
2	2	@	1				
3	3	제목	1	제목	NN		
4	4	:	1	:	SP		
5	5	[	1	[	SP		
6	6	98	1	98	NU		

11) 의미 표지에 쓰인 로마자와 아라비아 숫자는, "연세 한국어사전"(1998, 두산동아)의 의미 항목을 기호화한 것임.

7	7	한국일보	1	한국일보	NNX	
8	8	신춘문예	1	신춘문예	NN	
9	9	희곡	1	희곡	NN	
10	10	당선작	1	당선작	NN	
11	11	]	1	]	SP	
12	12	만행	1	만행	NN	
13	13	(1)	1	(1)	SP	
14	14	잦아든다	1	잦아들	VV	
15	14	잦아든다	2	ㄴ다	EF	
16	15	.	1	.	SP	
17	16	@	1			
18	17	밝아지면	1	밝	VJ	2-Ⅰ①㉡
19	17	밝아지면	2	아	EF	
20	17	밝아지면	3	지	VX	5-①
21	17	밝아지면	4	면	EF	
22	18	늦가을	1	늦가을	NN	
23	19	오후	1	오후	NN	
24	20	4시경	1	네#	NU	4-
25	20	4시경	2	시	NX	2-Ⅱ
26	20	4시경	3	경	SN	7-
27	21	.	1	.	SP	
28	22	@	1			
29	23	동물원	1	동물원	NN	
30	24	코끼리	1	코끼리	NN	
31	25	우리	1	우리	NN	1-Ⅰ
32	26	앞	1	앞	NN	-Ⅰ①
33	27	벤치	1	벤치	NN	
34	28	.	1	.	SP	
35	29	@	1			
36	30	승복	1	승복	NN	
37	31	차림의	1	차림	NX	
38	31	차림의	2	의	PA	
39	32	철주가	1	철주	NNX	
40	32	철주가	2	가	PA	
41	33	목탁을	1	목탁	NN	
42	33	목탁을	2	을	PA	
43	34	두드리며	1	두드리	VV	-
44	34	두드리며	2	며	EF	
45	35	주변을	1	주변	NN	
46	35	주변을	2	을	PA	
47	36	조심스레	1	조심스레	AV	
48	37	두리번거리고	1	두리번거리	VV	

49	37	두리번거리고	2	고	EF	
50	38	서	1	서	VV	-Ⅰ①
51	38	서	2	ㅓ	EF	
52	39	있다	1	있	VX	2-Ⅱx
53	39	있다	2	다	EF	
54	40	.	1	.	SP	
55	41	@	1			
56	42	이따금	1	이따금	AV	
57	43	불경을	1	불경	NN	
58	43	불경을	2	을	PA	
59	44	중얼거리고	1	중얼거리	VV	
60	44	중얼거리고	2	고	EF	
61	45	있고	1	있	VX	2-Ⅰ①
62	45	있고	2	고	EF	
63	46	,	1	,	SP	
64	47	벤치에는	1	벤치	NN	
65	47	벤치에는	2	에는	PA	
66	48	역시	1	역시	AV	2-③
67	49	승복차림의	1	승복	NN	
68	49	승복차림의	2	차림	NX	
69	49	승복차림의	3	의	PA	
70	50	기봉이	1	기봉	NNX	
71	50	기봉이	2	이	PA	
72	51	웅크린	1	웅크리	VV	
73	51	웅크린	2	ㄴ	EF	
74	52	채	1	채	NX	3-
75	53	떨고	1	떨	VV	1-Ⅱ③
76	53	떨고	2	고	EF	
77	54	있다	1	있	VX	2-Ⅰ①
78	54	있다	2	다	EF	
79	55	.	1	.	SP	
80	56	@	1			

1.2. 국립국어연구원 말뭉치

【구축자】 국립국어연구원

【구축 기간】 1992년~1999년

【목적】 국립국어연구원에서 1992년부터 시작한 "표준국어대사전" 편찬과 국어 연구를 위한 목적으로 구축되었다. 국가적 사업으로 구축한 대규모 말뭉치로는 최초이며, 전산 파일을 쉽게 얻을 수 있는 현대 문헌보다는 국어사, 고전문학, 개화기 자료, 사전, 구비문학 등 1980년대 이전의 자료들을 위주로 구축하였다.

【원 자료 성립 시기】 1400년대~1990년대

【규모】 총 67,650,000 어절

【입말/글말】 구어 전사 자료 외에는 모두 글말을 중심으로 이루어져 있다.

【구성】 국립국어연구원 말뭉치는 『표준국어대사전』(1999)의 편찬을 위하여 구축한 말뭉치와 연구원 내부에서 입력한 북한 문헌 말뭉치, 그리고 문화관광부에서 <국어정보처리기반 구축> 사업으로 구축한 말뭉치를 국립국어연구원 형식에 맞게 수정, 보완한 말뭉치로 구성되어 있다.

【주석】 1992년 『표준국어대사전』 편찬을 위하여 개발한 용례 검색 프로그램(HGREP)의 형식에 맞추어 파일 처음에 서지 정보를 주고 그 이후에는 각 페이지마다 파일 이름과 페이지 번호를 주는 방식으로 되어 있다.

【응용 사례】 『표준국어대사전』(1999) 편찬과 국립국어연구원의 각종 연구 사업에 활발하게 이용되고 있으며 국어 연구 및 자연어 처리 등의 여러 분야에서 다양하게 이용되고 있다.

【관련 자료】 말뭉치를 이용하기 위한 용례 검색 프로그램(HGREP)을 소개한 "문자열 검색, 사전 검색 프로그램의 연구와 개발(1997)"이 국립국어연구원 홈페이지 자료실에 등록되어 있다. 말뭉치 자료 및 관련 프로그램은, 국립국어연구원의 홈페이지인 http://www.korean.go.kr와 21세기 세종계획의 홈페이지인 http://www.sejong.or.kr에서 받아볼 수 있다.

장르 구분	내용의 구성	규모(어절)
국어사 문헌	15세기부터 19세기까지의 국어사 문헌들을 정리하여 입력. 이본이 있을 경우 별도로 구축하였음	4,450,873
고전 문학	고전 소설과 고전 시가 중 대표적인 작품을 선정. 이본이 있을 경우 별도로 구축하였음	4,205,872
소설류	신소설, 현대 소설, 번역 소설 중에서 문인들의 선호도가 높은 작품을 선정	15,857,202
시	대표적인 시집을 선정하여 구축	148,014
수필	대표적인 수필집을 선정하여 구축	2,533,343
신문	주요 일간지, 전문 분야 신문 및 개화기 신문 중 일정 기간을 선택하여 입력	5,429,208
잡지	주요 잡지, 전문 분야 잡지 및 개화기 잡지 중 일정 기간을 선택하여 입력	3,065,844
교과서	국립국어연구원에서 보유하고 있는 교과서를 입력	925,511
사전류	개화기에 나온 국어사전류와 남, 북한의 대표적 국어사전을 사전 검색 프로그램(HDB)의 형식에 맞게 입력	19,408,118
국립국어연구원 간행물	국어연구소와 국립국어연구원에서 발간한 보고서와 정기 간행물 중 말뭉치로 구축하기에 알맞은 간행물을 선정	1,523,000
북한 문헌	국립국어연구원에서 보유하고 있는 소설, 잡지, 신문, 교과서를 입력	1,887,460

장르 구분		내용의 구성	규모(어절)
기타	구비문학	구어 자료를 채록한 문헌들을 선정	8,215,113
	구어 전사 자료	토크쇼와 같은 텔레비전 방송 프로그램의 내용을 글로 풀어 쓴 자료	
	민속 자료	판소리 대본, 무가, 민요 등을 수집한 문헌을 선정하여 입력	
	언어 교육 자료	외국인을 위한 한국어 교육 자료와 한국인을 위한 외국어 교육 자료를 선정	
	아동, 만화물	어린이들을 위한 잡지, 동화책, 그림책, 만화책 등을 선정	
	대본 자료	희곡, 시나리오, 방송 대본 등을 선정	
	표준어 문헌	염상섭, 나도향 등 서울과 경기 지역 출신 작가들의 작품을 선정	

[표10] '국립국어연구원 말뭉치'의 구성

【날말뭉치(일부)】
(국립국어연구원 말뭉치)

/*
파일이름 : MCU00040.HWP
입력기 : 훈글 2.5
입력자 : 한동완
입력시작 : 96년 9월 6일
입력완료 : 96년 9월 14일
교정자 : 한동완
교정시작 : 96년 9월 15일
교정완료 : 96년 9월 19일
서지 : 月刊野談 第一卷 제4호, 1935년 1월. 癸酉社
입력범위 :
마지막파일 :
분량 : 135페이지.
어절 : 28625
기타 :
*/
- MCU00040.HWP -
<MCU00040,002>
¶新年漫談^
¶逸民^

오늘은 새해를 맞는 첫날이다. 이런 날은 다른 날과 달을 것이 업지마는 해가 박귀는 첫날이오 일년의 처음인 까닭으로 누구나 지난 열두달 동안에 격거온 - 여러가지 걱정하던 것은 버서버리고 낫버하든 것은 진보되여 가기를 자긔도 원하면 남에게도 비(禱)는 것이다.

이런 새해를 전일에는 양력 일월일일이라 하더니 지금은 을해(乙亥)년 원조(元旦)라 보통 쓴다. 그러나 양력이란 것은 지구(地球)가 저혼자 도는 것으로 하로라 하고 지구가 태양을 끼고 한번 도는 삼백륙십 오일로 일년을 삼어 그것을 열두달에 별넛스니 아모날이나 신년 원조를 삼을 수 잇스며 더구나 갑자(甲子)니 을축(乙丑)이니 하는 륙갑을 부치는 것은 본래부터 업는 것이다.

그런즉 오늘의 신년 원조를 을해년 처음날이라 함은 음력의 을해년을 인용(引用)함이다. 음력이란 것은 텬톄가 삼백륙십오도(度)와 사분지일인데 태양은 하로 한번을 돌고 태음은 이십구일 좀넘어서 한번을 돌아 다달이 해와 달이 한번식 모힘으로 한달을 삼고 태양이 텬톄보다 하로 한도(度)식 적게 가게 되여 삼백륙십오일 좀 넘어서 제자리에 잇게 됨으로 열두달로 일년을 삼고 삼년만콤 열 석달로 일년을 삼는 것이다.

<이하 생략>

1.3. 국어 정보 베이스 말뭉치

【구축자】 한국과학기술원(문화체육부와 과학기술처의 지원)

1. 주관 : 한국과학기술원 전산학과 최기선
2. 협력 : 시스템 공학연구소 자연어 처리부

【구축 기간】 1994~1997년

【목적】 '국어 정보 처리 기반 구축을 위한 연구'(1994-1997) 용역(연구책임자 최기선 교수)으로 이루어짐.

【규모】 약 7,158만 어절

년도	1994년도	1995년도	1996년도	1997년도	합계
규모(어절)	1,025만	1,489만	2,533만	2,110만	7,158만

[표 12] 한국과학기술원 말뭉치 연도별 어절 수

【응용 사례】 1998년의 21세기 세종 계획 "국어 기초 말뭉치 구축"의 일환으로 이를 표준화, TEI 헤더를 부착하여 세종 말뭉치에 통합됨.

【관련 자료】 『21세기 세종 계획 국어 기초 자료 구축』 연구보고서, 문화관광부,1998.

《국어정보베이스 I》 (시디롬 시험배포판)

○ Raw Corpus : 1000만 어절

한국어 분석 자료(가공하지 않은 원문)로, 각 화일은 TDMS입력 형태로 변환되어 있음(TDMS로 직접 입력하여 관리 가능)

○ Tagged Corpus : 100만 어절

한국어 분석 자료 - 품사 태그를 붙인 가공 말뭉치

○ TreeTagged Corpus : 1만 문장

한국어 분석 자료 - 한국어 구문 구조를 분석 가공한 말뭉치

《국어정보베이스II》(시디롬 시험배포판)

◎ 본 CD-ROM에는 1994. 10. - 1997. 12. 동안에 문화관광부[국어정보처리기반 구축을 위한 연구]와 과학기술부[통합국어정보베이스]의 지원을 받아 구축된 데이터베이스를 기초로 하여, 과학기술부 과학기술정책관리연구소 지원 STEP2000(Software Technology Enhancement Plan 2000) 과제인 "대용량 국어정보 심층처리 및 품질관리 기술개발"에 의하여 데이터베이스의 품질을 보완한 기초 자료 데이터베이스(raw corpus)와 가공 자료 데이터베이스(pos tagged corpus)를 수록하고 있다.

【규모와 구성】

○ 기초자료 데이터베이스 (1,000만 어절 수준)

○ 가공자료 데이터베이스 (20만 어절 수준)

○ 구문분석용 국어용례 데이터베이스[12)]

원시 말뭉치(날말뭉치)는, 각 분류된 체계 내에서 최대한 균형성을 이루도록 구축되었으며, 문학 분야(KDC 800) 30%, 비문학 분야 서적(35%) 및 신문 및 뉴스(35%)를 포함하고 있다. 또한 각 말뭉치 표본은 1만 어절을 넘지 않도록 구성함.

【주석】 말뭉치에 대한 정보를 제공하기 위해 헤더 정보가 부착됨. 원시 말뭉치와 품사 부착 말뭉치에 대한 마크업(헤더 정보 및 분류 정보 입력, tdms화를 위한 마크업) 정보에는, 제목, 저자, 출판년도, 발행처가 포함되며, 일반적인 출판물에 대한 분류 정보는 kdc(한국십진분류)와 ddc(듀이십진분류)를 기록하고, 신문에 대해서는 KAIST CS Lab에서 정한 분류 체계를 이용하여 kdc란에 news.XX 형태로 기록함.

【날말뭉치(일부)】
《국어정보베이스 I》(시디롬 시험배포판)】

12) http://csfive.kaist.ac.kr/kcp를 참조 바람.

<project> 국어정보처리 기반 구축을 위한 연구 95 </project>
<class> Corpus </class>
<language> Korean </language>
<textcode> KSC-5601-1988 </textcode>
<process> raw </process>
<version> 수동 V 0.5 </version>
<serialbumber> 96-11-ycpark-0002 </serialnumber>
<filename> mh2b0002.txt </filename>
<title> 한국적 슬기가 세계를 이긴다 </title>
<author> 임무생 </author>
<date> 1993.7.5. </date>
<typist> 한영균 </typist>
<inputstatus> 수동수정 1차 </inputstatus>
<note> </note>

<tdmsfiletext>
;저자:임무생
;출전:한국적 슬기가 세계를 이긴다
;발행일:1993.7
;출판사:신구미디어
;corpus 작성책임:한영균
;corpus 작성일자:1995.

&{
1. 어떤 기업이 21세기를 이끌어 나갈 것인가

어떤 기업이 21세기를 이끌어 나갈 것인가. 우리나라에서만도 하루에 수백 개의 기업이 새로 생겨나고 도산한다. 전세계적으로는 하루에도 수천 수만 개의 기업들이 새로 문을 여는가 하면 이와 반대로 그만큼의 기업이 흔적 없이 사라져가는 일이 되풀이 된다.

기업들은 또 하루에도 엄청난 숫자의 신제품을 쏟아내고 있다. 저마다 기업들이 온갖 아이디어를 짜내고 정성을 들인 것들이다. 그러나 이 중에서 소비자들이 이름을 기억할 만큼 히트하는 제품은 극히 소수에 불과하다. 소비자의 손에 미처 닿기도 전에 창고에서 폐기처분되는가 하면 상품 진열대에 변변히 얼굴도 내보지 못한 채 몇몇 소비자의 손에서 불합격 판정을 받고 사라지는 제품이 대부분이다. 기업들은 제품 하나 잘못 만들어 도산하기도 하지만 히트 상품 하나로 일약 세계적 기업으로 부상하기도 한다. <이하 생략>

【품사 표지 부착 말뭉치(일부)】
《국어정보베이스 I》(시디롬 시험배포판)】

중국문화의	중국/nq+문화/ncn+의/jcm
위기와	위기/ncn+와/jcj
재건	재건/ncpa
1981	1981/nno
년	년/nbu
11	11/nno
월	월/nbu
10	10/nno
일	일/nbu
,	,/sp
생전에	생전/ncn+에/jca
지위나	지위/ncn+나/jcj
명성이	명성/ncn+이/jcs
그다지	그다지/mag
드러나지	드러나/pvg+지/ecx
않았던	않/px+았/ep+던/etm
청나라의	청나라/nq+의/jcm
신하였던	신하/ncn+이/jp+었/ep+던/etm
거천	거천/nq
巨川	巨川/nq
양제	양제/nq
梁濟가	梁濟/nq+가/jcs
북경	북경/nq
북쪽의	북쪽/ncn+의/jcm
적수담	적수담/nq
積水潭에	積水潭/nq+에/jca
몸을	몸/ncn+을/jco
던져	던지/pvg+어/ecx
자살하였다	자살/ncpa+하/xsv+었/ep+다/ef
.	./sf

【구문분석 표지 부착 말뭉치(일부)】
《국어정보베이스 I》(시디롬 시험배포판)】

```
##########################################################
#
#  Tree Tagged Corpus  1996.8
#  원문
#    제목 : "현대인의 ㅅ과 종교"
#    저자 : 윤이흠
#    출판사 : 고려한림원
#    발행일 : 1993.12.27
#
##########################################################
; 한국의 불교는 기복불교이며, 한국의 기독교는 기복화되었다는 말을 흔히 한다.
(S
  (VP
      (NP
          (VP
              (VP
                  (NP (NP 한국/nq )+의/jcm 불교/ncn )+는/jxt
                  (VP (NP 기복/ncpa+불교/ncn )+이/jp ))+며/ecc+,/sp
              (VP
                  (VP
                      (NP (NP 한국/nq )+의/jcm
                          기독교/ncn )+는/jxt
                        기복화/ncpa+되/xsv  )+(AUXP  었/ep  )))+다/ef+는/etm  말
                        /ncpa )+을/jco
      (VP (ADVP 흔/paa+히/xsa ) 하/pvg ))+ㄴ다/ef+./sf )
; 한국의 종교들이 기복화되어간다는 것을 의미한다.
(S
  (VP
      (NP
          (VP
              (VP
                  (NP (NP 한국/nq )+의/jcm
                      종교/ncn+들/xsn )+이/jcs
                    기복화/ncpa+되/xsv )+(AUXP 어/ecx+가/px ))+ㄴ다/ef+는/etm 것
```

```
                    /nbn )+을/jco
        의미/ncpa+하/xsv )+ㄴ다/ef+./sf )
; 종교가 있는 곳에 기복행위가 같이 있다.
(S
   (ADJP
         (NP
             (ADJP (NP 종교/ncn )+가/jcs 있/paa )+는/etm 곳/ncn )+에/jca
         (ADJP (NP 기복/ncpa+행위/ncpa )+가/jcs
               (ADJP (ADVP 같이/mag ) 있/paa )))+다/ef+./sf )
; 기복과 종교는 서로 완전히 분리될 수는 없으되, 분명하게 서로 구분되는 현상이다.
(S
   (VP (NP 기복/ncpa+과/jcj 종교/ncn )+는/jxt
       (VP
           (VP
               (VP (ADVP 서로/mag )
                   (VP (ADVP 완전/ncps+히/xsa ) 분리/ncpa+되/xsv ))+(AUXP ㄹ/etm
                    수/nbn+는/jxt 없/paa ))+으되/ecs+,/sp
           (VP
               (NP
                   (VP (ADJP 분명하/paa )+게/ecs
                       (VP (ADVP 서로/mag ) 구분/ncpa+되/xsv ))+는/etm
                   현상/ncn )+이/jp )))+다/ef+./sf )
; 우리는 이 둘이 어떻게 구분되는가를 알 필요가 있다.
(S
   (ADJP (NP 우리/npp )+는/jxt
         (ADJP
               (NP
                   (VP
                       (NP
                           (VP
                               (NP (MODP 이/mmd ) 둘/nnc )+이/jcs
                               (VP (ADJP 어떻/pad )+게/ecs
                                   구분/ncpa+되/xsv ))+는가/ef )+를/jco
                       알/pvg )+ㄹ/etm 필요/ncpa )+가/jcs 있/paa ))+다/ef+./sf )
```

1.4. 고려대학교 한국어 말모둠 1

【구축자】 고려대학교 민족문화연구원

【구축 기간】 1995년

【목적】 매체 및 내용에 따라서 한국어를 고르게 대표하고자 하였다.

【원 자료 성립 시기】 1910년 이후의 자료를 대상으로 하되, 1970~1990년대 자료가 중심이 되었다. 언어 생활 및 입말 교육에서 활용 비중이 높은 1920~1960년대의 자료를 채택하였다.

【규모】 약 1,000만 어절

【입말/글말】 입말 및 버금 입말 12% / 글말 88%.

【구성】

구어/준구어 (전체의 12%)		
신문 인터뷰		40.2%
잡지 인터뷰		2.7%
연극	7.4%	
드라마	22.0%	
TV방송전사		18.6%
라디오방송전사	4.8%	
기타녹음전사	0.9%	
신문 (전체의 20%)		
잡지 (전체의 10%)		
책-정보 (전체의 34%)		
책-상상 (전체의 21%)		
기타 (전체의 2%)		

[표 16] 고려대학교 민족문화연구원 말모둠1

【주석】 TEI 호환 방식으로 태깅되어 있다.

【응용 사례】 언어학, 국어학, 전산언어학 연구 자료로 활용되고 있다.

【관련 자료】 김홍규, 강범모(1996), "고려대학교 한국어 말모둠 1 (KOREA-1 CORPUS) : 설계 및 구성"(고려대학교 한국어학회, 한국어학 3, 233-258), 강범모(1997), "고려대학교 한국어 말모듬(코퍼스)의 구축과 활용", 기초자료 데이터베이스의 구축 및 활용 ('97한국정보관리학회 추계학술특별세미나 발표자료집, 1997년 9월 12일, 국립중앙도서관).

1.5. 고려대학교 장르별 텍스트 코퍼스(KGENRE Corpus)

【구축자】 고려대학교 민족문화연구원
【구축 기간】 1997년
【목적】 다양한 특성을 가진 언어 장르별 비교를 위해 구축하였다.
【규모】 약 40만 어절
【입말/글말】 입말 및 버금 입말 12% / 글말 88%.
【구성】 입말 : 대화, 연극, 드라마, 인터뷰, 뉴스, 스포츠중계, 연설, 전화, 만화, TV 다큐멘터리, 한국어 회화집, 소설 대화 등
글말 : 신문 1,2,3, 서평, 취미, 대중잡지, 전기, 수필, 백과사전, 학문, 학술 논문, 학생 작문, 법조문, 절차적 텍스트, 개인 편지, 소설, 공문서, 공적 편지 등
【주석】 TEI 호환 방식으로 태깅되어 있다.
【응용】 통계적 방법에 의한 한국어 텍스트 유형 및 문체 분석 연구.
【관련 자료】 강범모. 1997. "고려대학교 한국어 말모듬(코퍼스)의 구축과 활용", 기초자료 데이터베이스의 구축 및 활용 ('97한국정보관리학회 추계학술특별세미나 발표자료집), 1997년 9월 12일, 국립중앙도서관.

1.6. 〈21세기 세종계획〉 국어 기초자료(1998)

【구축자】 대한민국 문화관광부

▲ 총괄 책임자: 김홍규

▲ 분야별 구축 책임자

△ 현대국어 자료: 김홍규(고려대), 협력책임자 이상섭(연세대)

△ 북한, 연변 자료: 소강춘(전주대)

△ 역사 문헌 자료: 한영균(울산대)

△ 구비문학 자료: 김병선(정신문화연구원)

【구축 기간】 1998년

【목적】 현대 국어, 북한/해외 한국어, 국어사 자료, 구비문학 자료에 걸친 기초 언어 말뭉치를 개발하는 한편, 이미 축적된 <국어정보처리기반 구축> 사업(1994-1997)의 말뭉치, 국립국어연구원 구축 말뭉치에 대한 표준화와 후처리 작업을 통해, 통합적 국가 말뭉치를 구축하였다.

【원 자료 성립 시기】 1900년대 초~1998년

【구성 및 규모】

(1) 현대국어 말뭉치(1,650만 어절)[13]

(2) 세종 말뭉치 통합 표준화 · 후처리(1억 2000만 어절)[14]

(3). 북한 및 해외 한국어 말뭉치(북한 : 35만 어절, 해외(연변) 한국어 : 100만 어절)

(4). 역사 자료 말뭉치(70만 어절)

(5). 구비문학 말뭉치(100만 어절)

【입말/글말】 입말 및 글말

【구성】 기존 말뭉치의 통합이 아닌, 이 프로젝트에 의해 새로이 구축된 현대 국어

13) 구어 자료 전사 50만 어절, 한자 비율 10% 이상의 국한혼용 자료 100만 어절을 포함

14) 국비 지원에 의해 이미 축적된 말뭉치. <국어정보처리기반 구축> 사업(1994-1997)의 말뭉치(7000만 어절), 국립국어연구원 구축 말뭉치(5000만 어절)를, 표준 지침에 따라 공통된 양식으로 모든 자료를 후처리하여 표준화하고, 향후에 대규모 말뭉치를 다양하게 활용하는 데 필요한 헤더 정보를 붙임. 이를 기반으로 전문연구기관 배포 및 실험용의 균형 말뭉치(1000만 어절 규모)를 제작하고, 향후의 말뭉치 구축 설계에 필요한 장르 · 시대 · 특성별 통계 및 분포도를 작성함.

말뭉치의 구성은 다음과 같다.

장르 구분		입말	글말	총 합계
기타 녹음	대화/대담, 좌담	17271		17271
	연설, 강의 ,강연	12678		12678
	기타 녹음 합계	29949		29949
기타 비출판물	교육자료		65144	65144
	대화/대담, 좌담	4419		4419
	방송극, 영화	352526		352526
	생활		28917	28917
	연극(희곡)	84001		84001
	체험 기술적 텍스트		10977	10977
	기타 비출판물 합계	440946	105038	545984
신문	대화/대담, 좌담	113655		113655
	대화/인터뷰	361677		361677
	생활		41577	41577
	신문		3253901	3253901
	연극(희곡)	20329		20329
	신문 합계	495661	3295478	3791139
잡지	교육자료		144567	144567
	대화/대담, 좌담	23360		23360
	대화/인터뷰	7278		7278
	사회		193440	193440
	생활		389925	389925
	예술론		152284	152284
	인문		312117	312117
	자연		207025	207025
	체험 기술적 텍스트		124330	124330
	총류		109429	109429
	잡지 합계	30638	1633117	1663755

[표 17] 세종 말뭉치(1998)의 구성(1)

장르 구분		입말	글말	합계
전자적 출판	대화/인터뷰	14941		14941
	사회		17265	17265
	생활		36025	36025
	인문		13858	13858
	자연		93347	93347
	체험 기술적 텍스트		20242	20242
	총류		28995	28995
	전자적출판 합계	14941	209732	224673
책	교육자료		895154	895154
	대화/대담, 좌담	12368		12368
	방송극, 영화	48036		48036
	사회		813628	813628
	상상적(예술적) 텍스트		3042248	3042248
	생활		107143	107143
	연극(희곡)	97039		97039
	연설, 강의 ,강연	1278		1278
	예술론		856030	856030
	인문		1186834	1186834
	자연		410845	410845
	체험 기술적 텍스트		1496124	1496124
	총류		925886	925886
	토론, 회의	13925		13925
	책 합계	172646	9733892	9906538
화면이 없는 방송	대화/대담, 좌담	49636		49636
	연설, 강의 ,강연	60455		60455
	화면이 없는 방송 합계	110091		110091
화면이 있는 방송	대화/대담, 좌담	340911		340911
	방송극, 영화	3242		3242
	총류	48807		48807
	화면이 있는 방송 합계	392960		392960
총 합계		1687832	14977257	16665089

[표 18] 세종 말뭉치(1998)의 구성(2)

【기타】 자료 파일은 <한글> 3.0의 문서 형식이며, <21세기 세종계획>의 표준 지침에 의한 TEI 헤더와 마크업이 첨부되어 있다.

【관련 자료】 이에 관한 상세한 정보는 <21세기 세종계획 국어 기초자료 구축 1998년도 보고서>에 기술되어 있다.

1.7. 〈21세기 세종 계획〉 국어 기초자료(1999)

【구축자】 대한민국 문화관광부

▲ 총괄 책임자: 김홍규

▲ 분야별 구축 책임자

△ 현대국어 문어(준구어 포함) 자료: 김홍규(고려대)

△ 현대국어 형태소 분석 자료: 김홍규(고려대)

△ 현대국어 구어 자료: 서상규(연세대)

△ 북한, 연변 자료: 소강춘(전주대)

△ 역사 문헌 자료: 김진영(경희대)

△ 구비문학 자료: 임치균(정신문화연구원)

△ 한-영 병렬 자료: 정태구(고려대)

【구축 기간】 1999년

【목적】 현대 국어, 북한/해외 한국어, 국어사 자료, 구비문학 자료에 걸친 기초 언어 말뭉치를 개발하기 위한 <21세기 세종계획>의 1999년도 국어 기초자료 구축 사업의 성과로 이루어진 것이다.

【구성 및 규모】 기존 말뭉치의 통합이 아닌, 이 프로젝트에 의해 새로이 구축된 세종 말뭉치의 구성은, 다음과 같다.

(1) 현대국어 기초 말뭉치(1,600만 어절)[15]

(2) 현대국어 형태소 분석 말뭉치(150만 어절, 127개 파일)

(3) 현대국어 구어 전사 말뭉치(50만 어절)

(4) 북한 및 해외 한국어 말뭉치(북한어 : 105만 어절, 해외(연변) 한국어 : 25만 어절)

(5) 역사 자료 말뭉치(75만 어절)

(6) 구비문학 말뭉치(130만 어절)

(7) 한·영 병렬 말뭉치(25만 어절)

【기타】 자료 파일은 <한글> 3.0의 문서 형식이며, <21세기 세종계획>의 표준 지침에 의한 TEI 헤더와 마크업이 첨부되어 있다.

15) 현대국어 문어 텍스트 말뭉치를 한자 비율 10% 이상의 국한혼용 자료 100만 어절을 포함함.

【관련 자료】 이에 관한 상세한 정보는 <21세기 세종계획 국어 기초자료 구축 1999 년도 보고서>에 기술되어 있다.

【현대국어 구어 전사 말뭉치】
(세종 말뭉치의 일부, 1999)

```
<!DOCTYPE tei.2  SYSTEM "c:\sgml\dtd\tei2.dtd" [
          <!ENTITY % TEI.corpus "INCLUDE">
          <!ENTITY % TEI.extensions.ent SYSTEM "sejong1.ent">
          <!ENTITY % TEI.extensions.dtd SYSTEM "sejong1.dtd">
]>
<tei.2>
  <teiHeader>
  <fileDesc>
     <titleStmt>
        <title>안성기 강연, 전자전사자료</title>
        <author>안성기, 대화자</author>
        <sponsor>대한민국 문화관광부</sponsor>
        <respStmt><resp>전자/표준화</resp>
              <name>연세대학교 언어정보개발연구원</name>
        </respStmt>
     </titleStmt>
     <extent>3833 어절</extent>
     <publicationStmt>
        <distributor>국립국어연구원</distributor>
        <idno>2CM00027.hwp</idno>
        <availability><p>배포 불가</p></availability>
     </publicationStmt>
     <sourceDesc>
          <p>원전 없음, 1988년 5월 27일에 연세대학교에서 열린 안성기의 강연을 일
부 녹음하여 전사함.</p>
     </sourceDesc>
  </fileDesc>
  <encodingDesc>
     <projectDesc><p>21세기 세종계획 2차연도 말뭉치 구축</p></projectDesc>
     <samplingDecl><p>녹음하여 전사</p></samplingDecl>
     <editorialDecl><p>21세기 세종계획 입말뭉치 전사 및 마크업 지침에 따름
</p></editorialDecl>
```

```
</encodingDesc>
<profileDesc>
   <creation><date>1996</date></creation>
   <langUsage>
         <language id=KO usage=99>한국어, 표준어</language>
   </langUsage>
   <particDesc>
            <person id=P1 sex=M age=?><p>사회자</p></person>
            <person id=P2 sex=M age=40s><p>안성기, 영화배우</p></person>
            <person id=P3 sex=? age=?><p>청중들</p></person>
   </particDesc>
   <settingDesc>
         <p>영화배우 안성기가 영화와 인생에 대해 강연을 하고 학생이 질문을 한
다.</p>
   </settingDesc>
   <textClass>
         <catRef scheme='SJ21' target='M2898'>구어 강연 녹음 전사</catRef>
   </textClass>
</profileDesc>
<revisionDesc>
   <change><date>1996</date>
      <respStmt>
            <resp>프로젝트책임자</resp><name>서상규</name>
      </respStmt>
      <item>음성 발화를 녹음(녹화)하여 PC로 입력</item>
   <change><date>1996-8</date>
      <respStmt><resp>보조원</resp><name>주승원</name></respStmt>
      <item>1차 전사</item>
   </change>
   <change><date>1999-9</date>
      <respStmt><resp>연구보조원</resp><name>최은희</name></respStmt>
      <item>전사 후 띄어쓰기 오류 교정, 철자 오류 교정, 태그 붙임</item>
   </change>
   <change><date>1999-9</date>
      <respStmt><resp>연구원</resp><name>안의정</name></respStmt>
      <item>태그 오류 수정, 세종 21 프로젝트 헤더 붙임, 마킹</item>
   </change>
</revisionDesc>
</teiHeader>
<text>
<u who=P1>( ) 우리 안성기씨로부터 좀 재밌는 얘기 듣기로 하고 내가 구차스럽게 더
많은 얘기를 하면 시간이 ( ). 자 박수. <vocal desc='박수소리'> ( ) 좀 조용히 얘기 들읍시
```

다.</u>
<vocal who=P3 desc='웃음'>
<u who=P2>안녕하세요. 만나 뵙게 돼서 ( ) 반갑습니다. 아.</u>
<vocal who=P3 desc='웃음'>
<u who=P1>자 그럼 좀 더 앞으로 나와요.</u>
<u who=P2>저는 저 예전부터 이 사실은 전 이 연세 대학을 굉장히 좋아했습니다. 어, 그까 대학에 들어갈 때부터 저 근데 고려 대학을 두 번 시험을 쳐 가지고 물론 두 번 다 떨어졌지만은 어, 수학을 잘 못해 가지고 연대를 포기했었습니다. 근데 연대를 정말 좋아했고 그 다음에 제가 좋아하는 분들이 또 연대를 나오신 분이 많습니다. 저희 제 저 꼬마 아들 대부를 서주신 최인훈 씨라던가 그 다음에 배창호 친구 같으신 배창호 감독 전부 어, 연대 출신이라고요. 아 그래서 연대는 또 사실 좋아하고 있습니다. 이건 알랑빵구가 아닙니다.</u>
<vocal who=P3 desc='웃음'>

【북한/연변 말뭉치】
(세종 말뭉치의 일부, 1999)

```
<!DOCTYPE tei.2 SYSTEM "c:\sgml\dtd\tei2.dtd" [
    <!ENTITY % TEI.corpus "INCLUDE">
    <!ENTITY % TEI.extensions.ent SYSTEM "sejong2.ent">
    <!ENTITY % TEI.extensions.dtd SYSTEM "sejong2.dtd">
]>
<tei.2>
<teiHeader>
<fileDesc>
<titleStmt>
<title>하늘과 땅, 전자파일</title>
<author>김영길</author>
<sponsor>대한민국 문화관광부</sponsor>
<respStmt><resp>문헌 입력, 표준화, 헤더 붙임</resp>
<name>전주대학교</name>
</respStmt>
</titleStmt>
<extent>32294 어절</extent>
<publicationStmt>
<distributor>국립국어연구원</distributor>
<idno>NE950000.hwp</idno>
<availability><p>99-05 배포 불가</p></availability>
```

```
</publicationStmt>
<notesStmt>
<note><p></p></note>
</notesStmt>
<sourcedesc>
<bibl><author>김영길</author>
<title>하늘과 땅</title>
<pubPlace>평양</pubPlace>
<publisher>문학예술종합출판사</publisher>
<date>1995</date>
</bibl>
</sourceDesc>
</fileDesc>
<encodingDesc>
<projectDesc><p>21세기 세종계획 2차연도 말뭉치 구축</p>
</projectDesc>
<samplingDecl><p>본문 전체 입력<p.>
</samplingDecl>
</encodingDesc>
<profileDesc>
<creation><date>1995</date></creation>
<langUsage>
<language id=NK usage=99>북한어, 문화어</language>
</langUsage>
<textClass>
<catRef scheme='SJ21' target=N1331> 책: 상상적(예술적)텍스트, 소설/일반</catRef>
</textClass>
</profileDesc>
<revisionDesc>
<change><date>*****1999/05</date>
<respStmt><resp>입력자</resp><name>한홍월</name></respStmt>
<item>본문 전체 입력, 헤더 붙임</item>
</change>
<change><date>*****1999/05/6</date>
<respStmt><resp>제 1교 교정자</resp><name>최홍자</name></respStmt>
<item>제 1교</item>
</change>
<change><date>1999/05/14</date>
<respStmt><resp>제 2교 교정자</resp><name>리성자</name></respStmt>
<item>제 2교 및 헤더 수정</item>
</change>
<change><date>1999/05/18</date>
```

<respStmt><resp>제 3교 교정자</resp><name>리현희</name></respStmt>
<item>제 3교 및 헤더 수정</item></revisionDesc>
</change>
<change><date>1999/10/18</date>
<respStmt><resp>제 4교 교정자</resp><name>소강춘</name></respStmt>
<item>제 4교 및 헤더 수정</item>
</change>
<item>완교</item>
</revisionDesc>
</teiHeader>
<group>
<text>
<body>
<head>김영길단편소설집</head>
<head>하늘과 땅</head>
<head>문학예술종합출판사</head>

<head>고산진언덕에서</head>

<head>1</head>

최고사령부가 전개된 고산진은 우리 나라 서북부첨단에 위치한 자그마한 산간마을이다. 마을북쪽변두리로는 물살이 급한 압록강이 여울목에 널린 들쑹날쑹한 바위들에 부딪쳐 분수와 같은 비말을 일으키며 세차게 흐르고있었다.

최고사령부가 차지한 위치의 지리학적개념으로 보아서도 나라의 정세가 초긴장상태에 이르렀다는것을 알수 있었다.

후날 력사가들과 군사전문가들은 곡절많고 신축성있는 이 전쟁행정에 대해 응당한 평가를 내릴테지만 오늘의 엄혹한 전투정황앞에서 "이제는 더 물러설 땅이 없다" 고 절망에 빠져 넉두리를 하는 사람들도 있었다.

이 시각 인민군부대들은 압록강의 급류마냥 고산진을 향해 강행군해가고있었다. 행군중에 있는 모든 장령, 전사들은 어떻게 하나 정해진 시간내에 목적지에 도달하여 최고사령관동지의 새로운 작전적방침을 실현하는 길에 한목숨 기꺼이 바치겠다는 필사의 각오외는 다른 사념이 없었다.

땅크와 포차, 후방차들로 길을 메워버린 이 철갑행렬의 틈바구니로 마치 우중충한 밀림속에 애되게 피여난 한송이 들국화마냥 너무도 색다른 파아란 승용차 한대가 경적을 연방 울리면서 앞으로 톺아나가고있었다.

승용차에 탄 외무성 부상 김명수는 류동하는 전사들과 전투기재들의 당면임무와는 거리가 먼것 같은 지구상의 평화에 관한 문제에 심취되여있었다.

성 대변인이라고 말할수 있는 김명수부상은 나이도 이제는 40줄에 들어선데다가 장대한 체구와 항상 엄숙한 얼굴표정을 짓고있는때문인지 남달리 무게가 있어보였다. 전시에도 외교일군의 체모를 잃지 않고 기름을 발라 단정스레 빗어넘긴 그의 머리에서는 아침해살이 금실처럼 반짝거렸다. 해방전에 고등교육도 받을만큼 받은 그는 험난한 세파속에 휘말려 외국에도 두루 다녀본 사람이였다.

【역사 자료 말뭉치】
(세종 말뭉치의 일부, 1999)

```
<!DOCTYPE tei.2  SYSTEM "c:\sgml\dtd\tei2.dtd" [
        <!ENTITY % TEI.corpus "INCLUDE">
        <!ENTITY % TEI.extensions.ent SYSTEM "sejong1.ent">
        <!ENTITY % TEI.extensions.dtd SYSTEM "sejong1.dtd">
]>

<tei.2>

  <teiHeader>
  <fileDesc>
     <titleStmt>
        <title>산성일기, 전자파일</title>
        <author>알수 없음</author>
        <sponsor>대한민국 문화관광부</sponsor>
        <respStmt><resp>문헌입력, 표준화, 헤더붙임</resp>
              <name>울산대학교, 한영균</name>
        </respStmt>
     </titleStmt>
     <extent>7,880어절</extent>
     <publicationStmt>
        <distributor>국립국어연구원</distributor>
        <idno>p7bc0009.hwp</idno>
        <availability><p>배포 불가</p></availability>
     </publicationStmt>
     <notesStmt>
        <note><p>이 텍스트는 산성일기의 김광순 주석본(1985년 형설출판사간)에 실려
있는 원문 영인을 이용하여 입력한 것임.</p></note>
     </notesStmt>
     <sourceDesc>
        <bibl><author>알수없음</author>
              <title>산셩일긔</title>
              <pubPlace>알수없음</pubPlace>
           <publisher>필사본임</publisher>
           <date>1636년이후로 추정</date>
        </bibl>
     </sourceDesc>
  </fileDesc>
```

```
    <encodingDesc>
        <projectDesc><p>21세기 세종계획 1차연도 말뭉치 구축</p>
        </projectDesc>
        <samplingDecl><p>직접 입력</p>
        </samplingDecl>
        <editorialDecl><p>21세기 세종계획 역사자료 말뭉치 문헌 입력 지침에 따름</p>
        </editorialDecl>
    </encodingDesc>
    <profileDesc>
        <creation><date>1998</date></creation>
        <langUsage>
            <language id=KO usage=99>한국어, 고어</language>
        </langUsage>
        <textClass>
            <catRef scheme='SJ21' target='p7bc'>필사본, 일기</catRef>
        </textClass>
    </profileDesc>
    <revisionDesc>
        <change><date>1998/11 </date>
            <respStmt><resp>프로젝트책임자</resp><name>한영균
</name></respStmt>
            <item>세종21프로젝트 헤더붙임</item>
        </change>
    </revisionDesc>
    </teiHeader>
  <text>
  <body>
  <pb n=1a>
  <head>산셩일긔 병ᄌᆞ</head>
  <p>만녁 십칠년 긔튝 구월의 노라치란 오랑캐로 뇽호쟝군을 ᄒᆞ이다</p>
  <p>노라치는 녀진쥐 ᄯᅡ 오랑캐니 / 제 한아비와 아비 다 다른 오랑캐 난의 죽고 노라치는 동방으로 ᄃᆞ라나니 / 븍녁히 모든 오랑캐를 침노ᄒᆞ야 긔셰 졈졈 듕ᄒᆞ고 / 잇다감 노략ᄒᆞᆫ 한인을 도로 보내여 듕도에 헌통ᄒᆞ더니 / 다른 오랑캐 근오십 등이 ᄉᆡ합호를 <pb n=1b> 노략ᄒᆞ고 / 지휘 뉴부를 죽이고 건쥐로 ᄃᆞ라나거ᄂᆞᆯ / 노라치 즉시 근 오십의 머리를 버혀 듕됴의 드리고 / ᄯᅩ 됴흔 물을 공ᄒᆞ고 제 한아비와 아비 듕국을 위ᄒᆞ야 병화의 죽은 곡졀을 알외여뻐 대쟝 일홈을 어드니라 / </p>
  <p>노라치 졈졈 강셩ᄒᆞ야 제 아ᄋᆞ 속기합치를 죽이고 군ᄉᆞ를 아오라 모든 오랑캐를 침노ᄒᆞ더니 / 무오 긔미 년간의 무슌셩을 함몰ᄒᆞ고 총병 댱<pb n=2a>승윤 등이 죽고 / 경냑 양효 등이 대패ᄒᆞ고 쳠ᄉᆞ 반총안 등이 다 죽다 / </p>
  <p>됴션국 도원슈 강홍닙 부원슈 김경셔 죵ᄉᆞ관 니민환 쇼쟝 김응하 등이 마가 채를 조차 삼빅 니를 드러가 도적의게 빠힌 빈 되니 / 김응하는 힘뻐 빠화 죵시 굴티 아니코 죽으니 / 이 니른 김쟝군이라 됴경이 병조판셔를 튜증ᄒᆞ고 / 강홍닙 이하는 다 항복ᄒᆞ니라 /
```

긔미의 노라치 아<pb n=2b>국을 통코져 ᄒᆞ야 홍닙은 머므러 두고 민한 등을 내여 보내니라 / 이 히 오월의 노라치 나라 일홈을 금이라 ᄒᆞ고 황의를 닙고 딤이로라 ᄒᆞ며 / 그 후의 심양을 함몰ᄒᆞ니 / 텬지 원슝환으로 경냑을 ᄒᆞ야 광녕의 가 도적을 막ᄌᆞ르게 ᄒᆞ여더니 / 병인의 노라치 원슝환의게 패ᄒᆞ야 분통ᄒᆞ야 등창 나 죽으니 ᄎᆞᄌᆞ 홍타시 납ᄒᆞ니라 / 노라치 일즉 나가 노다가 뫼 <pb n=05>녑히 ᄒᆞᆫ 겨집이 오좀을 누고 디나가거ᄂᆞᆯ 보니 / 오좀이 뫼흘 쑤러 깁희 물채가 드러가니 / 노라치 괴이히 너겨 그 겨집을 ᄃᆞ려다가 아ᄃᆞᆯ을 나흐니 이 니른바 홍타시라 / 뎡묘 경월의 강홍닙이 홍타시로 다래여 대군을 니ᄅᆞ혀 아국을 틸시 / 압녹강을 어름으로 건너 이 밤의 의쥐를 엄습ᄒᆞ니 부윤 판관이 다 죽고 / 안쥐를 함몰ᄒᆞ니 병ᄉᆞ 목ᄉᆞ들이 다 죽다 / 샹이 강화 피란ᄒᆞ시고 세ᄌᆞ는 <pb n=06>뎐쥬의 분됴ᄒᆞ시디 / </p>

【구비문학 말뭉치】
(세종 말뭉치의 일부, 1999)

```
<!DOCTYPE tei.2 SYSTEM "c:\sgml\dtd\tei2.dtd" [
    <!ENTITY % TEI.spoken "INCLUDE">
    <!ENTITY % TEI.corpus "INCLUDE">
    <!ENTITY % TEI.extensions.ent SYSTEM "sejong1.ent">
    <!ENTITY % TEI.extensions.dtd SYSTEM "sejong1.dtd">
]>
<tei.2>
<teiHeader>
<fileDesc>
  <titleStmt>
    <title>한국구비문학대계 1-7 42-43(부자 되는 법)</title>
    <author>한국정신문화연구원</author>
    <sponsor>대한민국 문화관광부</sponsor>
    <respStmt><resp>문헌 입력, 표준화, 헤더 붙임</resp>
        <name>한국정신문화연구원</name>
</respStmt>
</titleStmt>
  <extent>279 어절</extent>
  <publicationStmt>
    <distributor>국립국어연구원</distributor>
    <idno>QABG0106.HWP</idno>
    <availability><p>배포 불가</p></availability>
</publicationStmt>
  <notesStmt>
```

```
    <note></note>
</notesStmt>
  <sourceDesc>
    <bibl><author>한국정신문화연구원</author>
          <title>한국구비문학대계 1-7 42-43(부자 되는 법)</title>
          <pubPlace>서울</pubPlace>
          <publisher>고려원</publisher>
          <date>1982-12-27</date>
</bibl>
</sourceDesc>
</fileDesc>
<encodingDesc>
  <projectDesc><p>21세기 세종계획 1차연도 말뭉치 구축</p>
</projectDesc>
  <samplingDecl><p>원문대로 입력하되, 틀린부분은 주석을 달아서 표시함</p>
</samplingDecl>
  <editorialDecl><p>21세기 세종계획 문헌 입력 지침에 따름</p>
</editorialDecl>
</encodingDesc>
<profileDesc>
  <creation><date>1981-5-3</date></creation>
  <langUsage>
       <language id=KO usage=99>한국어, 방언</language>
</langUsage>
<particDesc>
      <person id='P10' sex='?' age='?'><p>성기열+최명동+김용범</p></person>
      <person id='P20' sex='남' age='64'><p>허황</p></person>
      <person id='P30' sex='?' age='?'><p>청중</p></person>
      <person id='P40' sex='?' age='?'><p>일동</p></person>
</particDesc>
<settingDesc>
<p>시종 웃음을 잃지 않으면서 이야기를 전개시켰다. 이야기 도중에 회관으로 들어온 청중 한 사람이 이 제보자의 이야기 솜씨는 이미 알고 있다는 식으로 가벼운 농담을 걸었지만 한 손으로 그 말을 제지하면서 계속 이야기를 진행시켰다. 이야기의 성격상 소를 모는 시늉이며, 싸우는 시늉을 해 가면서 동작 표정을 다양하게 사용하는 것이 이 제보자의 특징이기도 하였다.</p>
</settingDesc>
  <textClass>
       <catRef scheme='SJ21' target='M2336'>구비문학자료</catRef>
</textClass>
</profileDesc>
<revisionDesc>
```

```
<change><date>1998-6-22</date>
        <respStmt><resp>보조연구원</resp><name>윤승준,박옥주,신은경,김현진
</name></respStmt>
        <item>텍스트 입력</item>
</change>
<change><date>1998-8-10</date>
        <respStmt><resp>보조연구원</resp><name>윤승준,박옥주,신은경,김현진
</name></respStmt>
        <item>교정 및 헤더 및 태그 부착</item>
</change>
<change><date>1998-11-2</date>
        <respStmt><resp>보조연구원</resp><name>윤승준,박옥주,신은경,김현진
</name></respStmt>
        <item>교정</item>
</change>
</revisionDesc>
</teiHeader>
<text>
<body>
<head>부자 되는 법</head>
그전에 에 그전에 아주 막역한 친구가 있는데,</u> <u who=P10> 네.</u> <u who=P20>한 친구는 둘이다 똑같이 아들, 딸이 수북하고 가족이 많은데.</u> <u who=P30> 허반장이 한 마디 또 하는구료.</u> <u who=P20>아, 한 친구는 아주 그냥 큰 부잔데, 한 친구는 밑구녕이 째지게 어려워요. 밑구녕이 째지게 어련데 게 맨날 부자 친구네 가서 막역하니깐, 얻어도 먹고 그저 양식도 주먼 갖다 먹고 헌데 생전 그렇게 먹구만 살 수 없고, 한 날은 부자 친구보고,</u></p>
<p><u who=P20> "자 여보게, 거 어떡허면 부자가 되나?" </u></p>
<p><u who=P20> "자네도 부자가 되고 싶은가?" </u></p>
<p><u who=P20> "아, 이 사람아, 부자가 되고 싶지 않은 사람 있나?" </u></p>
<p><u who=P20> "그럼, 낼<notr>(내일)</notr> 아침에 일찌감치 우리 집으로 오게. 내 부자되는 법을 알으켜줄께." </u></p>
<p><u who=P20> "부자되는 법을 알으켜 주게." </u></p>
<p><u who=P20>그러니깐,</u></p>
<p><u who=P20> "그럼, 그러게." </u></p>
<p><u who=P20>아 큰아들 이름이 뭔지,</u></p>
<p><u who=P20> "야 큰 아이야." </u></p>
<p><u who=P20> "네." </u></p>
<p><u who=P20> "야, 너 저 윗간에 소 좀 저 지붕으로 치끌어라." </u></p>
<p><u who=P20> "네." </u></p>
<p><u who=P20>그러더니, 아따, 고삐, 그 아이는 고삐가 모자르니깐 좀 이어가지고 올라가서, 어이 그냥 끌어다니고 있어야 된단 말이예요?</u></p>
<p><u who=P20>자, 저의 시약시<note>(색시)</note>가 부엌에서 내다보니깐 아 저의
```

사내가 지붕에서 소를 잡아 당기고 소는 앞 뜨랑에 있거든요.</u>
<u who=P20> "아 왜 소는 그리 끌어당겨?" </u>
<u who=P20> "아 아버님이 소를 이렇게 지붕으로 끌래요." </u>
<u who=P20>그래 또 부축이 나서, 쇠 꽁메를 때리면 올라가라 부추긴단 말이예요. 그래 그냥 메누린 이를 열고 부주께<note>(부지깽이)</note>로 두둘기고, 사내는 끌어 올리고 하는데, 저의 어머니가 나오며,</u>
<u who=P20> "야! 왜 그러느냐?" </u>
<u who=P20> "아, 아버지가 소를 지붕으로 끌어 올리래요." </u>
<u who=P20> "응 그래." </u>
<u who=P20>그때 또, 저의 어머니도 그냥 문들다구두명 나가란 말이예요. 그러니깐 그 친구 어련 친구하구, 부자친구 둘이 사랑으로 내다보구 있지요.</u>
<u who=P20> "야, 이제 고만 외간으로 들이내라." </u>
<u who=P20>그랬단 말이예요. 그리고 났는데,</u>
<u who=P20> "내가 내일 아치기<note>(아침에)</note> 내가 갈 테니 자네 조반 좀 해 놓게." </u>
<u who=P20> "그럼 그러게." </u>
<u who=P20>그것만 깨주곤 부자 되는 식을 안 가르쳐줘요. 그래 그 친구네 가니깐 참 어떻게 귀한 친구가 오니까는 이젠 밥을 좀 해 놀까 하는데,</u>
<u who=P20> "밥은 누가 하며, 인젠 좀 가르쳐 주게." </u>
<u who=P20> "응. 자네 큰아들보구 저 외양간의 소 좀 지붕으로 치끌라 그러보게." </u>
<u who=P20>그러니까는 이 친구가,</u>
<u who=P20> "야! 아무개야." </u>

【현대국어 형태소 분석 말뭉치】
(세종 말뭉치의 일부, 1999)

```
<!DOCTYPE tei.2  SYSTEM "c:\sgml\dtd\tei2.dtd" [
 <!ENTITY % TEI.corpus "INCLUDE">
 <!ENTITY % TEI.extensions.ent SYSTEM "sejong1.ent">
 <!ENTITY % TEI.extensions.dtd SYSTEM "sejong1.dtd"> ]>
<tei.2>
 <teiHeader>
  <fileDesc>
   <titleStmt>
    <title>아름다운 그 시작,  형태소 분석 전자파일</title>
    <author>유기성</author>
    <sponsor>대한민국 문화관광부</sponsor>
```

```
      <respStmt>
        <resp>문헌입력, 표준화, 형태소 정보 부착</resp>
        <name>고려대학교 민족문화연구원</name>
      </respStmt>
    </titleStmt>
    <extent>17096 어절</extent>
    <publicationStmt>
      <distributor>국립국어연구원</distributor>
      <idno>2BT_001.HWP, 원본:AIDORN.HWP</idno>
      <availability><p>배포 불가</p></availability>
    </publicationStmt>
    <notesStmt>
      <note><p>원문의 일부분 ⅓을 추출하여 형태소 정보 부착</p></note>
    </notesStmt>
    <sourceDesc>
      <bibl>
        <author>유기성</author>
        <title>아름다운 그 시작</title>
        <pubPlace>서울</pubPlace>
        <publisher>삼신각</publisher>
        <date>1994</date>
      </bibl>
    </sourceDesc>
  </fileDesc>
  <encodingDesc>
    <projectDesc><p>21세기 세종계획 2차년도 형태소 말뭉치 구축</p></projectDesc>
    <samplingDecl><p>21세기 세종계획 말뭉치 파일을  형태소 정보 부착 파일로 변환
</p></samplingDecl>
    <editorialDecl><p>21세기 세종계획 형태소 분석 말뭉치 구축 지침(1999-4)에 따른 형
                태소 정보 부착</p></editorialDecl>
  </encodingDesc>
  <profileDesc>
    <creation>
      <date>1994</date>
    </creation>
    <langUsage>
      <language id=KO usage=99>한국어, 표준어</language>
    </langUsage>
    <textClass>
      <catRef scheme='SJ21' target='M1331'>책 : 상상적 텍스트, 소설/ 일반</catRef>
    </textClass>
  </profileDesc>
```

```
  <revisionDesc>
   <respStmt>
    <resp>프로젝트 책임자</resp>
    <name>김홍규</name>
   </respStmt>
   <change>
    <date>1999</date>
    <respStmt>
     <resp>연구보조원</resp>
     <name>이봉원, 김일환, 차준경, 최승철, 조진현, 김태훈</name>
    </respStmt>
    <item>21세기 세종계획 형태소 분석 말뭉치 구축 지침(1999-4)에 따른 형태소 정보
          검토, 수정</item>
   </change>
  </revisionDesc>
</teiHeader>
<text>
<body>
2BT_0010000010  1.          1/SN + ./SF
2BT_0010000020  아름다운     아름답/VA + 은/ETM
2BT_0010000030  그           그/MM
2BT_0010000040  시작         시작/NNG
2BT_0010000050  "이형,       "/SS + 이/NNP + 형/NNG + ,/SP
2BT_0010000060  잘           잘/MAG
2BT_0010000070  가시오."     가/VV + 시/EP + 오/EF + ./SF + "/SS
2BT_0010000080  석훈이       석훈/NNP + 이/JKS
2BT_0010000090  빙그르르     빙그르르/MAG
2BT_0010000100  웃으며       웃/VV + 으며/EC
2BT_0010000110  악수를       악수/NNG + 를/JKO
2BT_0010000120  청했다.      청하/VV + 았/EP + 다/EF + ./SF
2BT_0010000130  햇빛을       햇빛/NNG + 을/JKO
2BT_0010000140  받은         받/VV + 은/ETM
2BT_0010000150  석훈의       석훈/NNP + 의/JKG
2BT_0010000160  안경테가     안경테/NNG + 가/JKS
2BT_0010000170  반짝         반짝/MAG
2BT_0010000180  빛났다.      빛나/VV + 았/EP + 다/EF + ./SF
2BT_0010000190  진우는       진우/NNP + 는/JX
2BT_0010000200  말없이       말없이/MAG
2BT_0010000210  석훈이       석훈/NNP + 이/JKS
2BT_0010000220  내민         내밀/VV + ㄴ/ETM
2BT_0010000230  손을         손/NNG + 을/JKO
2BT_0010000240  잡았다.      잡/VV + 았/EP + 다/EF + ./SF
```

2BT_0010000250	따뜻했다.	따뜻/XR + 하/XSA + 았/EP + 다/EF + ./SF
2BT_0010000260	석훈의	석훈/NNP + 의/JKG
2BT_0010000270	진지한	진지/XR + 하/XSA + ㄴ/ETM
2BT_0010000280	눈동자와	눈동자/NNG + 와/JC
2BT_0010000290	마주	마주/MAG
2BT_0010000300	잡은	잡/VV + 은/ETM
2BT_0010000310	손을	손/NNG + 을/JKO
2BT_0010000320	통해	통하/VV + 아/EC
2BT_0010000330	가슴	가슴/NNG
2BT_0010000340	깊은	깊/VA + 은/ETM
2BT_0010000350	곳에	곳/NNG + 에/JKB
2BT_0010000360	묻어	묻/VV + 어/EC
2BT_0010000370	둔	두/VX + ㄴ/ETM
2BT_0010000380	서러움이	서러움/NNG + 이/JKS
2BT_0010000390	꿈틀대는	꿈틀대/VV + 는/ETM
2BT_0010000400	것	것/NNB
2BT_0010000410	같았다.	같/VA + 았/EP + 다/EF + ./SF
2BT_0010000420	손바닥의	손바닥/NNG + 의/JKG
2BT_0010000430	뜨거운	뜨겁/VA + 은/ETM
2BT_0010000440	기운,	기운/NNG + ,/SP
2BT_0010000450	힘,	힘/NNG + ,/SP
2BT_0010000460	움직임.	움직이/VV + ㅁ/ETN + ./SF
2BT_0010000470	진우는	진우/NNP + 는/JX
2BT_0010000480	새삼	새삼/MAG
2BT_0010000490	자신이	자신/NNG + 이/JKS
2BT_0010000500	살아	살/VV + 아/EC
2BT_0010000510	있음을	있/VX + 음/ETN + 을/JKO
2BT_0010000520	느꼈다.	느끼/VV + 었/EP + 다/EF + ./SF

【한영병렬 말뭉치】
(세종 말뭉치의 일부, 1999)

【한국어 말뭉치】


```
<cesDoc version=3.24>
<cesHeader version="2.3">
  <fileDesc>
      <titleStmt>
          <h.title>뉴스위크</h.title>
          <respStmt>
              <respType>문헌 입력, 표준화, 헤더붙임</respType>
              <respName>고려대학교 민족문화연구원</respName>
          </respStmt>
      </titleStmt>
      <extent>
          <wordcount>13396</wordcount>
      </extent>
      <publicationStmt>
          <distributor>국립국어연구원</distributor>
          <pubAddress>서울시 중구 정동 5-1 덕수궁내 석조전</pubAddress>
          <availability status=restricted>배포 불가</availability>
          <PubDate>1999</PubDate>
      </publicationStmt>
      <sourceDesc>
          <biblStruct>
              <monogr>
                  <h.title>뉴스위크</h.title>
                  <h.author></h.author>
                  <imprint>
                      <pubPlace>서울</pubPlace>
                      <publisher>중앙일보사</publisher>
                      <pubDate>1999</pubDate>
                  </imprint>
              </monogr>
          </biblStruct>
      </sourceDesc>
  </fileDesc>
  <encodingDesc>
```

```
        <projectDesc>21세기 세종계획 2차연도 한영 병렬 말뭉치 구축</projectDesc>
        <editorialDecl>21세기 세종계획 말뭉치 문헌 입력 지침에 따름</editorialDecl>
        <samplingDecl></samplingDecl>
        <refsDecl>본문, 구획, 문단, 문장 단위로 일련번호를 붙여 병렬텍스트와 정렬

    </refsDecl>
    </encodingDesc>
    <profileDesc>
        <creation><date>1999/07/20</date></creation>
        <langUsage>
            <language id="kr" iso639="kr"></language>
        </langUsage>
        <wsdUsage>
            <writingSystem id="iso2022-kr"></writingSystem>
        </wsdUsage>
        <textClass>
            <catRef scheme='SJ21' target="R1200">잡지: 총류</catRef>
        </textClass>
        <translations>
            <translation trans.loc="rb9901ee.hwp" lang=en wsd="iso8859-1">
            <translator></translator>
        </translations>
    </profileDesc>
    <revisionDesc>
        <change>
            <changeDate>1999/09/20<changeDate>
            <respName>정태구</respName>
            <h.item>본문 마크업</h.item>
        </change>
    </revisionDesc>
    </cesHeader>
  <text>
  <body id=1>
  <div id=1.1>
  <head id=1.1.h1>The Best Boys of Summer</head>
  <p id=1.1.p1><s id=1.1.p1.s1>내가 어렸던 시절에는 야구가 최고 인기였다.</s><s id=1.1.p1.s2> 그때 나는 날이면 날마다 당시로서는 가장 의미심장한 주제를 놓고 친구들과 입씨름을 했다.</s><s id=1.1.p1.s3> 테드 윌리엄스와 조 디마지오 중 누가 더 훌륭한 타자인가, 미키 맨틀과 윌리 메이스 중 누가 더 유능한 중견수인가, 또 메이저 리그의 모든 투수 중 가장 유능한 선수는 샌디 쿠팩스냐 봅 깁슨이냐 하는 따위였다.</s></p>
  <p id=1.1.p2><s id=1.1.p2.s1>마침내 그에 대한 해답이 나왔다.</s><s id=1.1.p2.s2> 1886년 이후의 미국 야구사를 기록해온 유력 스포츠 주간지 스포팅 뉴스가 '야구 1백 걸'
```

(Baseball' s 100 Greatest Players)을 내놓은 것이다.</s><s id=1.1.p2.s3> 그러나 최고 선수 1백 명의 서열을 대담하게 매긴 이 책은 누가 더 훌륭한 선수인가에 대한 야구계의 최대 논쟁을 2000년대로까지 이어지도록 만들고 있다.</s></p>

【영어 말뭉치】


```
<cesDoc version=3.24>
<cesHeader version="2.3">
  <fileDesc>
      <titleStmt>
          <h.title>Newsweek</h.title>
          <respStmt>
              <respType>문헌 입력, 표준화, 헤더붙임</respType>
              <respName>고려대학교 민족문화연구원</respName>
          </respStmt>
      </titleStmt>
      <extent>
          <wordcount>23028</wordcount>
      </extent>
      <publicationStmt>
          <distributor>국립국어연구원</distributor>
          <pubAddress>서울시 중구 정동 5-1 덕수궁내 석조전</pubAddress>
          <availability status=restricted>배포 불가</availability>
          <PubDate>1999</PubDate>
      </publicationStmt>
      <sourceDesc>
          <biblStruct>
              <monogr>
                  <h.title>Newsweek</h.title>
                  <h.author></h.author>
                  <imprint>
                      <pubPlace></pubPlace>
                      <publisher>Newsweek</publisher>
                      <pubDate>1999</pubDate>
                  </imprint>
              </monogr>
          </biblStruct>
      </sourceDesc>
  </fileDesc>
  <encodingDesc>
      <projectDesc>21세기 세종계획 2차연도 한영 병렬 말뭉치 구축</projectDesc>
```

```
        <editorialDecl>21세기 세종계획 말뭉치 문헌 입력 지침에 따름</editorialDecl>
        <samplingDecl></samplingDecl>
        <refsDecl>본문, 구획, 문단, 문장 단위로 일련번호를 붙여 병렬텍스트와 정렬
        </refsDecl>
    </encodingDesc>
    <profileDesc>
        <creation><date>1999/07/20</date></creation>
        <langUsage>
            <language id="en" iso639="en"></language>
        </langUsage>
        <wsdUsage>
            <writingSystem id="iso8859-1"></writingSystem>
        </wsdUsage>
        <textClass>
            <catRef scheme='SJ21' target="R1200">잡지: 총류 </catRef>
        </textClass>
        <translations>
            <translation trans.loc="rb9901ek.hwp" lang=kr wsd="iso2022-kr">
            <translator></translator>
        </translations>
    </profileDesc>
    <revisionDesc>
        <change>
            <changeDate>1999/9/20<changeDate>
            <respName>정태구</respName>
            <h.item>본문 마크업</h.item>
        </change>
    </revisionDesc>
    </cesHeader>
  <text>
  <body id=1>
  <div id=1.1>
  <head id=1.1.h1>The Best Boys of Summer</head>
  <p id=1.1.p1><s id=1.1.p1.s1>When I was a kid and baseball was king, my friends
and I whiled away the days, weeks and years discussing the great philosophical issues
of our time.</s><s id=1.1.p1.s2>Who was the better ballplayer, Ted Williams or Joe
DiMaggio?</s><s id=1.1.p1.s3>Which center fielder reigned supreme, Mickey Mantle or
Willie Mays?</s><s id=1.1.p1.s4>With all on the line, whom would you pitch, Sandy
Koufax or Bob Gibson?</s></p>
  <p id=1.1.p2><s id=1.1.p2.s1>Now, finally, on the eve of the World Series, we have
those answers (Williams, Mays, Koufax) and more.</s><s id=1.1.p2.s2>The Sporting
News, which has chronicled baseball since 1886, has produced a book, "Baseball's 100
```

Greatest Players," guaranteed to help propel the game's great debates into the next millennium.</s><s id=1.1.p2.s3>The book, sneak-previewed to NEWSWEEK last week, brazenly ranks the 100 in order.</s><s id=1.1.p2.s4>It seeks to capitalize on two American boomlets: the rebirth of affection for a game that has ceased to be the national pastime; and the nation's current fascination with summing up the century with top-100 lists.</s><s id=1.1.p2.s5>"Baseball needs more connection to its great players of the past," says Sporting News editorial director John Rawlings.</s></p>

【정렬 문서】

```
<cesAlign type=sent version=1.6>
<cesHeader version="2.3">
  <fileDesc>
      <titleStmt>
          <h.title>뉴스위크</h.title>
      <publicationStmt>
          <distributor>국립국어연구원</distributor>
          <pubAddress>서울시 중구 정동 5-1 덕수궁내 석조전</pubAddress>
          <availability status=restricted>배포 불가</availability>
          <PubDate>1999</PubDate>
      </publicationStmt>
      <sourceDesc>
          <biblStruct>
              <monogr>
                  <h.title>뉴스위크</h.title>
                  <h.author></h.author>
                  <imprint>
                      <pubPlace></pubPlace>
                      <publisher>Newsweek</publisher>
                      <pubDate>1999</pubDate>
                  </imprint>
              </monogr>
          </biblStruct>
      </sourceDesc>
  </fileDesc>
  </cesHeader>
<linkList>
<linkGrp domains="1 ; 1">
<linkGrp domains="1.1 ; 1.1">
    <link xtargets="1.1.h1 ; 1.1.h1">
<linkGrp domains="1.1.p1 ; 1.1.p1">
```

```
    <link xtargets="1.1.p1.s1 ; 1.1.p1.s1 1.1.p1.s2">
    <link xtargets="1.1.p1.s2 1.1.p1.s3 1.1.p1.s4 ; 1.1.p1.s3">
</linkGrp>
<linkGrp domains="1.1.p2 ; 1.1.p2 1.1.p3">
    <link xtargets="1.1.p2.s1 ; 1.1.p2.s1">
    <link xtargets="1.1.p2.s2 ; 1.1.p2.s2 1.1.p2.s3">
    <link xtargets="1.1.p2.s3 ; 1.1.p3.s1">
    <link xtargets="1.1.p2.s4 1.1.p2.s5 ; 1.1.p3.s2">
</linkGrp>
```

1.8. ETRI 품사태그 부착 말뭉치 (시험판 CD)

【구축자】 전자통신연구원 컴퓨터소프트웨어기술연구소 지식정보연구부

【구축 기간】 1998-1999

【목적】 본 제품은 전자통신연구원 컴퓨터소프트웨어기술연구소 지식정보연구부에서 자연어 정보처리기술 표준화 과제(1998-1999, 정보통신부)의 결과물로 만든 것이다.

【원 자료 성립 시기】

【규모 및 구성】

이 말뭉치(총 31,393어절)는, 소설류(novel) 63.8%, 비소설류(expl) 32.5%, 뉴스기사류(news, newsp) 3.7%로 구성되어 있다.

한편, 이 말뭉치는 ETRI 품사태그 부착 말뭉치 버전 2.0(약 28만 8천어절)을 원본으로 그 중에서 일부를 추출하여 시험용으로 만든 것으로, 원본의 구성은 다음과 같다.

novel	162,486어절(56%)
essay	4,575어절(2%)
expl	100,327어절(35%)
record	4,220어절(1%)
study	4,012어절(1%)
news	2,283어절(1%)
newsp	10,366어절(4%)
총	288,269어절

구성면에서 본다면, 소설류(novel, essay) 58%, 비소설류(expl, record, study) 38%, 뉴스 기사류(news, newsp) 4% 등으로 되어 있다.

【관련자료】 http://aladin.etri.re.kr/~nlu/STANDARD/(자연어 정보처리기술 표준화 웹페이지)에서 관련 정보를 볼 수 있다.

【ETRI 날말뭉치(일부)】

;<corpus_info>
;<name>ETRI raw corpus 1999</name>
;<version>2.0</version>
;<filename>expl04raw.txt</filename>
;<org_taggedfile>expl04.txt v2.0</org_taggedfile>
;<language>Korean</language>
;<textcode>KSC-5601-1988</textcode>
;<process>raw</process>
;<producer>ETRI Knowledge Information Research Department, NLU Team
</producer>
;</corpus_info>
;<source_info>
;<title>한국경제, 설 땅이 없다: 일본, 동남아연합, 한국경제의 선택과 전략</title>
;<author>이경규 [외]</author>
;<date>1993</date>
;<publisher>중앙일보사</publisher>
;<kdc>320.911</kdc>
;<ddc>330.9519</ddc>
;</source_info>
일본 장학금으로 일본으로 유학을 가든, 심지어 미국으로 유학을 가든 상관하지 않는다.
자기네들이 도와줬다는 것으로 만족한다는 느긋한 태도다.
91년 11월 런던에서는 아시아개발기금의 제6차 재원보충회의가 열렸다.
기금이 바닥 나 모두 42억달러를 모으기 위해 각국 대표들이 머리를 맞대고 회의를 거듭했다.
주요국들이 2백만 3백만달러씩을 더 내기로 했으나 그래도 목표액에는 태부족이었다.
서로 눈치만 살피면서 자기네들은 도저히 더 이상은 낼 수 없다는 표정들이었다.
이때 일본대표가 발언 신청을 했다.
"다른 나라들이 양해를 해준다면 부족한 돈 나머지를 일본이 내고 싶은 데 괜찮겠느냐"는 것이었다.
미국을 비롯한 영국 독일 프랑스 등 소위 강대국이라는 나라 대표들의 표정은 그야말로 벌레 씹은 모습들이었다고 한다.
이렇게 해서 총액 기준으로 따져서 미국은 6억 8천만달러를 낸데 비해 일본은 자그마치 15억 8천만달러를 냈던 것이다(한국은 대만과 함께 1천 5백만달러를 냈다).
그러잖아도 일본의 이니셔티브 속에 운영되어 왔던 아시아개발은행(ADB)은 이 일을 계기로 일본의 위상이 어떤 것인지를 확실하게 보여준 셈이었다. <이하 생략>

【ETRI 품사태그 부착 말뭉치(일부)】

;<corpus_info>
;<name>ETRI tagged corpus 1999</name>
;<version>2.0</version>
;<filename>novel01.txt</filename>
;<old_filename>etri99_kaist_corpus_97_mis-novel_pos_aca001.txt</old_filename>
;<language>Korean</language>
;<textcode>KSC-5601-1988</textcode>
;<process>pos</process>
;<copyright>copyright by ETRI (Electronics and Telecommunications Research Institute) 1999</copyright>
;<producer>ETRI Knowledge Information Research Department, NLU Team</producer>
;</corpus_info>
;<source_info>
;<title> 보이지 않는 손</title>
;<author>강형원</author>
;<date>1994</date>
;<publisher>남도</publisher>
;<kdc>813.6</kdc>
;<ddc>895.734</ddc>
;</source_info>

101000010	"얼른	"/s+얼른/mag
101000020	다녀오겠어요."	다녀오/pv+겠/ep+어/ef+요/jx+./s+"/s
101000030	이영후는	이영후/nc+는/jx
101000040	이렇듯	이렇/pa+듯/ec
101000050	손님을	손님/nc+을/jc
101000060	집으로	집/nc+으로/jc
101000070	끌고	끌/pv+고/ec
101000080	와	오/pv+아/ec
101000090	밤새도록	밤/nc+새/pv+도록/ec
101000100	놀면서	놀/pv+면서/ec
101000110	아내에게	아내/nc+에게/jc
101000120	그런	그런/mm
101000130	심부름까지	심부름/nc+까지/jx
101000140	시키는	시키/pv+는/etm
101000150	것이	것/nb+이/jc
101000160	몹시	몹시/mag

101000170	미안한	미안/nc+하/xsm+ㄴ/etm
101000180	듯	듯/nb
101000190	그녀가	그녀/np+가/jc
101000200	나가는	나가/pv+는/etm
101000210	뒷모습을	뒷모습/nc+을/jc
101000220	물끄러미	물끄러미/mag
101000230	바라보았다.	바라보/pv+았/ep+다/ef+./s
101000240	그러나	그러나/maj
101000250	그녀는	그녀/np+는/jx
101000260	언제나	언제나/mag
101000270	싫은	싫/pa+은/etm
101000280	소리	소리/nc
101000290	한번	한/nn+번/nb
101000300	하지를	하/pv+지/ec+를/jc
101000310	않았다.	않/px+았/ep+다/ef+./s
101000320	"지점장님.	"/s+지점장/nc+님/xsn+./s
101000330	술은	술/nc+은/jx
101000340	무슨	무슨/mm
101000350	술이에요.	술/nc+이/co+어/ef+요/jx+./s
101000360	이만큼	이/np+만큼/jc
101000370	마셨으면	마시/pv+었/ep+으면/ec
101000380	됐지....."	되/pv+었/ep+지/ef+...../s+"/s
101000390	술은	술/nc+은/jx
101000400	혼자	혼자/mag
101000410	다한	다하/pv+ㄴ/etm
101000420	듯	듯/nb
101000430	코가	코/nc+가/jc
101000440	빨갛게	빨갛/pa+게/ec
101000450	달아오른	달아오르/pv+ㄴ/etm
101000460	박진성이	박진성/nc+이/jc
101000470	말했다.	말/nc+하/pv+었/ep+다/ef+./s

<이하 생략>

1.9. 말뭉치 서비스 기관 : 21세기 세종 계획 자료실

【설명】 1998년부터 시작된 '21세기 세종계획'에서 구축된 각종 말뭉치와, 말뭉치 처리를 위한 각종 프로그램을 "일반 국민을 비롯하여 국어 연구자 및 응용 분야의 산업에 국어 연구를 위한 기반 자료와 방법론을 제시"하기 위해서 개설된 국내 최초의 공개적인 말뭉치 서비스 기관이다.

【주소】 http://www.sejong.or.kr

【이용 방법】 : 위의 주소로 접속하여, 자료 사용의 원칙[16]에 동의하여, 자료실에 사용자 등록을 하고, 아이디와 비밀번호를 발급받으면, 자료실의 자료를 모두 이용할 수 있다.

【말뭉치】

(1). 국어기초 자료 구축

○ 현대국어

신소설로부터 1920년대를 중심으로 한 주요 소설 작품 전문, 약간의 형태소 분석 자료가 제공되고 있음. 2000년 10월 현재, 모두 47개 표본(형태소 분석 표본 2개 포함).

16) 국립국어연구원에서는 자료실을 통하여 국립국어연구원과 21세기 세종계획의 사업 결과물을 공개하고 있습니다. 이는 일반 국민을 비롯하여 국어 연구자 및 응용 분야의 산업에 국어 연구를 위한 기반 자료와 방법론을 제시하여 한국어의 연구 기반을 튼튼히 하기 위한 것입니다.

따라서 본 자료실에 등록되어 있는 자료들은 무엇보다도 국어 연구를 위한 연구용 목적으로 사용할 때만 이용이 가능합니다. 자료실에 등록되어 있는 자료를 무단으로 복제하거나 프로그램의 일부를 수정없이 추출하여 상업용 목적으로 판매하는 것은 금지합니다. 자료실의 자료를 국립국어연구원의 의도와 다른 목적으로 사용할 경우는 국립국어연구원 및 해당 연구자의 동의를 얻어야 합니다.

자료실의 자료들은 국어정보화 관련 산업의 발전을 위한 기초 연구 자료로 활용해 주십시오 1999년 8월 12일 국립국어연구원

○ 국어사

15세기로부터 1900년경에 이르는 국어사 자료의 원문 입력 자료가 제공되고 있음. 2000년 10월 현재, 348개 표본.

국어사

검색된 게시물 : 349 / 349 , 쪽번호 : 23 / 24

번호	◎	제 목	성 명	작성일	조회
18		속삼강행실도(중간본)	국립국어연구원	99/06/02	69
17		속삼강행실도(원간본)	국립국어연구원	99/06/02	78
16		주역언해	국립국어연구원	99/06/02	74
15		텬로력뎡 권2	국립국어연구원	99/05/25	59
14		텬로력뎡 권1	국립국어연구원	99/05/25	65
13		오륜행실도	국립국어연구원	99/05/25	64
12		동의보감(탕액편)	국립국어연구원	99/05/25	62
11		삼강행실도(동경대본)	국립국어연구원	99/05/25	76
10		내훈	국립국어연구원	99/05/25	95
9		훈민정음언해본	국립국어연구원	99/05/25	291
8		월인석보 권15 [호암미술관 소장]	국립국어연구원	99/05/25	94
7		월인천강지곡(상)	국립국어연구원	99/05/25	88
6		석보상절(권 3)	국립국어연구원	99/05/25	135
5		바미나경언해	국립국어연구원	99/05/25	67
4		용비어천가	국립국어연구원	99/05/25	173

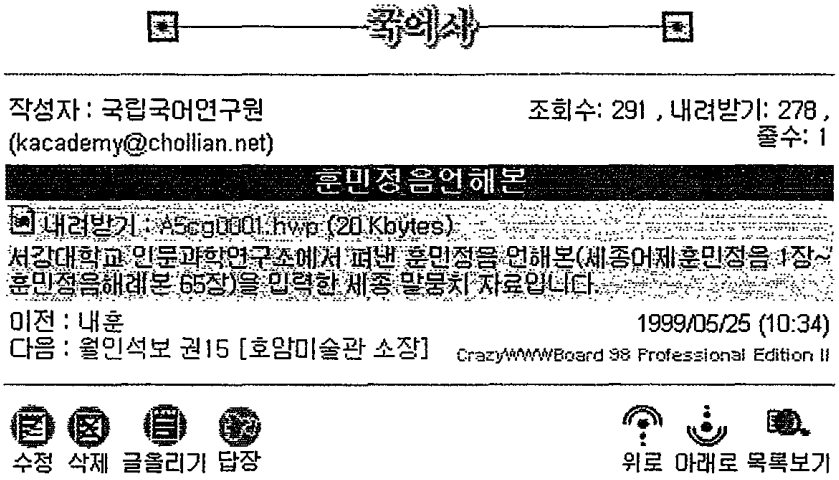
[1] [Prev] - [21] [22] [23] [24]

전체목록

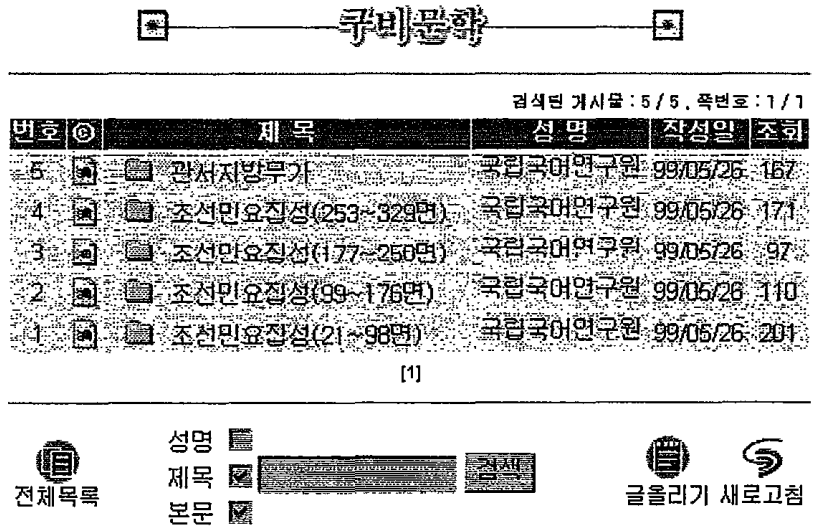
성명
제목
본문

검색

위로 아래로 글올리기 새로고침


국어사
작성자 : 국립국어연구원
(kacademy@chollian.net)
조회수: 291 , 내려받기: 278 , 줄수: 1
훈민정음언해본
내려받기 : A5cg0001.hwp (20 Kbytes)
서강대학교 인문과학연구소에서 펴낸 훈민정음 언해본(세종어제훈민정음 1장~훈민정음해례본 65장)을 입력한 세종 말뭉치 자료입니다.
이전 : 내훈
다음 : 월인석보 권15 [호암미술관 소장]
1999/05/25 (10:34)
CrazyWWWBoard 98 Professional Edition II
수정 삭제 글올리기 답장
위로 아래로 목록보기

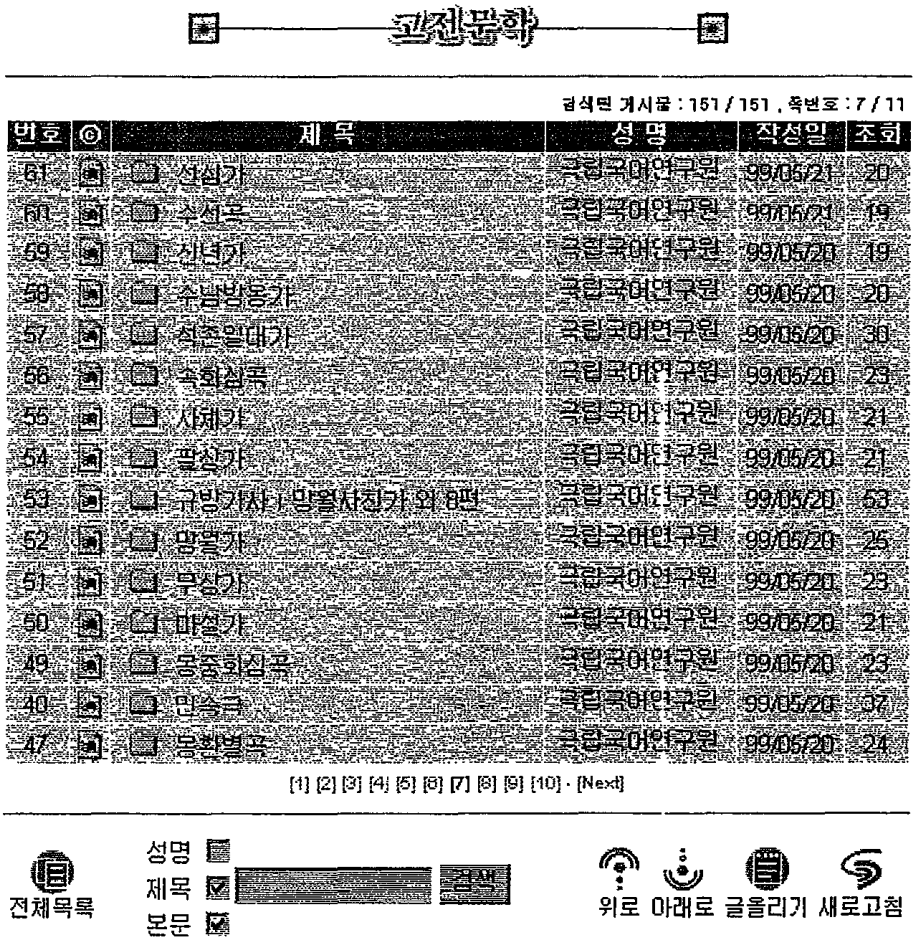
○ 구비문학

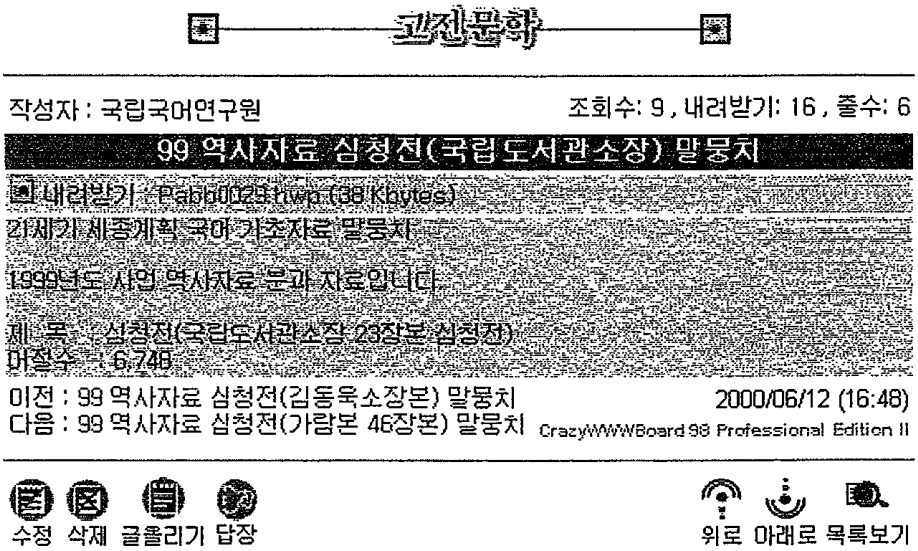

구비문학
검색된 게시물 : 5 / 5 , 쪽번호 : 1 / 1
번호 제 목 성 명 작성일 조회
5 관서지방무가 국립국어연구원 99/05/26 167
4 조선민요집성(253~329면) 국립국어연구원 99/05/26 171
3 조선민요집성(177~250면) 국립국어연구원 99/05/26 97
2 조선민요집성(99~176면) 국립국어연구원 99/05/26 110
1 조선민요집성(21~98면) 국립국어연구원 99/05/26 201
[1]
전체목록
성명
제목
본문
검색
글올리기 새로고침

조선민요집성, 관서지방무가 등의 자료가 제공되고 있음. 2000년 10월 현재, 5개 표본.

○ 고전문학

심청전, 춘향전 등 소설 및 가사 등의 필사본과 구활자본 등의 고전문학 자료의 전문이 제공되고 있음. 2000년 10월 현재, 151개 표본.





○ 신문

- 독립신문: 1896년 4월 7일의 창간호 이후의 신문 입력 자료가 제공됨 251개 표본

【날말뭉치(세종계획 홈페이지)】
(독립신문 일부)

```
<!DOCTYPE tei.2  SYSTEM "c:\sgml\dtd\tei2.dtd" [
    <!ENTITY % TEI.corpus "INCLUDE">
    <!ENTITY % TEI.extensions.ent SYSTEM "sejong1.ent">
    <!ENTITY % TEI.extensions.dtd SYSTEM "sejong1.dtd">
]>
<tei.2>
  <teiHeader>
  <fileDesc>
  <titleStmt>
  <title>독립신문, 전자 파일</title>
  <author>독립신문 편집부</author>
```

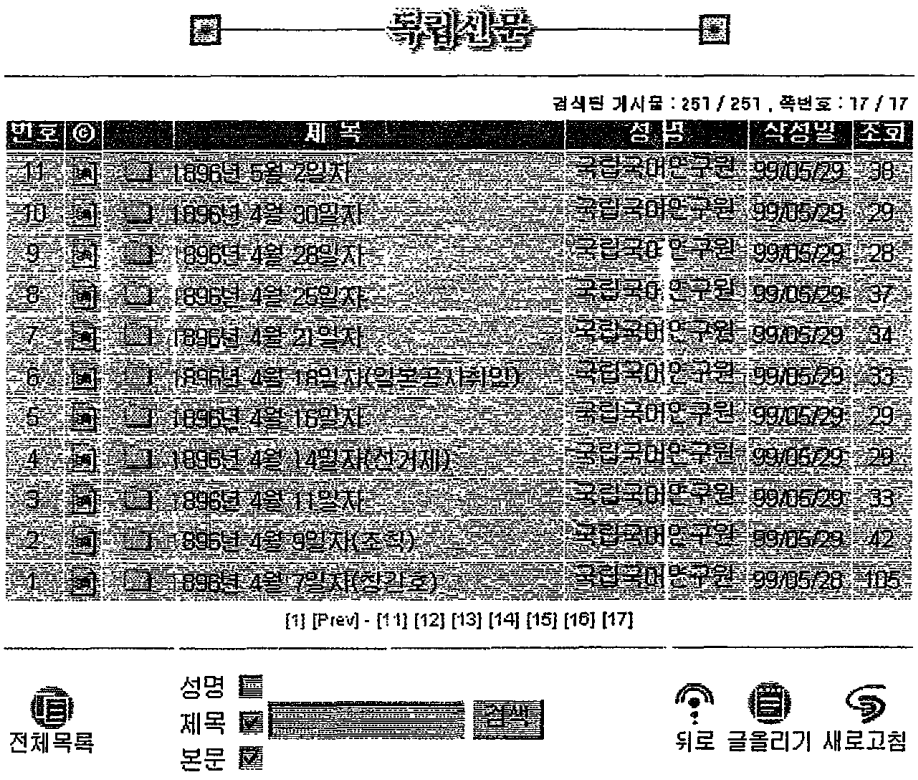
```
<sponsor>대한민국 문화관광부</sponsor>
<respStmt><resp>국립국어연구원 말뭉치 입수, 표준화, 헤더 붙임</resp>
  <name>국립국어연구원</name>
</respStmt>
</titleStmt>
<editionStmt>
  <edition><date>1995/09/20</date>전산입력</edition>
</editionStmt>
 <extent>865 어절</extent>
<publicationStmt>
<distributor>국립국어연구원</distributor>
<idno>aa000374.hwp(Pp964070.hwp)</idno>
<availability><p>배포 불가</p></availability>
</publicationStmt>
<notesStmt>
<note><p>이 텍스트는 국립국어연구원 작업 내용 유지</p></note>
</notesStmt>
<sourceDesc>
<bibl><author>독립신문 편집부</author>
  <title>독립신문</title>
  <pubPlace></pubPlace>
  <publisher>독립신문사</publisher>
  <date>1896/04/07</date>
</bibl>
</sourceDesc>
</fileDesc>
<encodingDesc>
<projectDesc><p>21세기 세종계획 1차년도 말뭉치 구축</p>
</projectDesc>
<samplingDecl><p>파일 변환 정보 없음</p>
</samplingDecl>
<editorialDecl><p>21세기 세종계획 말뭉치 문헌 입력 지침에 따름</p>
</editorialDecl>
</encodingDesc>
<profileDesc>
<creation><date>1896/04/07</date></creation>
<langUsage>
  <language id=KO usage=99>한국어</language>
</langUsage>
<textClass>
  <catRef scheme='SJ21' target='M1100'>신문: 총류</catRef>
</textClass>
```

```
</profileDesc>
<revisionDesc>
  <change>
      <date>1995/09/20</date>
      <respStmt>
          <resp>입력자</resp><name>서정목</name>
      </respStmt>
      <item></item>
  </change>
  <change>
      <date>1995/09/30</date>
      <respStmt>
          <resp>교정자</resp><name>서정목</name>
      </respStmt>
      <item>교정</item>
  </change>
  <change>
      <date>1998/10</date>
      <respStmt>
          <resp>프로젝트 책임자</resp><name>김흥규</name>
          <resp>연구원</resp><name>차준경</name>
          <resp>프로그래머</resp><name>이애란</name>
      </respStmt>
      <item>파일 변환, 세종 21 프로젝트 헤더 붙임, 마킹</item>
  </change>
</revisionDesc>
</teiHeader>

<text>
<body>
<p><PP964070,01><note>- PP964070.HWP -</note></p>
<head>논셜</head>
<p>우리가 독닙신문을 오늘 처음으로 출판ᄒᆞᄂᆞᆫᄃᆡ 조션 속에 잇ᄂᆞᆫ ᄂᆡ외국 인민의게 우리 쥬의를 미리 말ᄉᆞᆷᄒᆞ여 아시게 ᄒᆞ노라</p>
<p>우리는 첫지 편벽되지 아니ᄒᆞᆫ 고로 무ᄉᆞᆷ 당에도 상관이 업고 상하 귀천을 달니 ᄃᆡ졉 아니ᄒᆞ고 모도 죠션 사ᄅᆞᆷ으로만 알고 죠션만 위ᄒᆞ며 공평이 인민의게 말ᄒᆞᆯ 터인ᄃᆡ 우리가 셔울 빅셩만 위ᄒᆞᆯ 게 아니라 죠션 젼국 인민을 위ᄒᆞ여 무ᄉᆞᆷ 일이든지 ᄃᆡ언ᄒᆞ여 주라 ᄒᆞᆷ 경부에셔 ᄒᆞ시ᄂᆞᆫ 일을 빅셩의게 젼ᄒᆞᆯ 터이요 빅셩의 경셰을 경부에 젼ᄒᆞᆯ 터이니 만일 빅셩이 경부 일을 자세이 알고 경부에셔 빅셩에 일을 자세이 아시면 피ᄎᆞ에 유익ᄒᆞᆫ 일 만히 잇슬 터이요 불평ᄒᆞᆫ ᄆᆞᄋᆞᆷ과 의심ᄒᆞᄂᆞᆫ 싱각이 업서질 터이옴 우리가 이 신문 출판ᄒᆞ기ᄂᆞᆫ 취리ᄒᆞ랴ᄂᆞᆫ 게 아닌 고로 갑슬 헐허도록 ᄒᆞ엿고 모도 언문으로 쓰기ᄂᆞᆫ 남녀 상하 귀천이 모도 보게 ᄒᆞᆷ이요 ᄯᅩ 귀졀을 ᄲᅢ여 쓰기ᄂᆞᆫ 알어보기 쉽도록 ᄒᆞᆷ이라 우리ᄂᆞᆫ 바른 ᄃᆡ로만 신
```

문을 홀 터인 고로 정부 관원이라도 잘못ᄒᆞᄂᆞᆫ 이 잇스면 우리가 말홀 터이요 탐관오리들을 알면 세샹에 그 사ᄅᆞᆷ의 힝적을 폐일 터이요 ᄉᆞᄉᆞ빅셩이라도 무법ᄒᆞᆫ 일 ᄒᆞᄂᆞᆫ 사ᄅᆞᆷ은 우리가 차저 신문에 셜명홀 터이옴 우리ᄂᆞᆫ 죠션

대군쥬 폐하와 됴션 정부와 죠션 인민을 위ᄒᆞᄂᆞᆫ 사ᄅᆞᆷ드린 고로 편당 잇ᄂᆞᆫ 의ᄂᆞᆫ 이든지 ᄒᆞᆫ 쪽만 싱각코 ᄒᆞᄂᆞᆫ 말은 우리 신문샹에 업실 터이옴 ᄯᅩ ᄒᆞᆫ 쪽에 영문으로 기록ᄒᆞ기는 외국 인민이 죠션 ᄉᆞ경을 자셰이 몰은즉 혹 편벽된 말만 듯고 죠션을 잘못 싱각홀까 보아 실샹 ᄉᆞ경을 알게 ᄒᆞ고져 ᄒᆞ여 영문으로 조곰 긔록홈 그리ᄒᆞᆫ즉 이 신문은 쪽 죠션만 위홈을 가히 알 터이요 이 신문을 인연ᄒᆞ여 너외 남녀 샹하 귀쳔이 모도 죠션 일을 서로 알 터이옴 우리가 ᄯᅩ 외국 사경도 죠션 인민을 위ᄒᆞ여 간간이 긔록홀 터이니 그걸 인연ᄒᆞ여 외국은 가지 못ᄒᆞ드리도 죠션 인민이 외국 사경도 알 터이옴 오날은 처음인 고로 대강 우리 쥬의만 세샹에 고ᄒᆞ고 우리 신문을 보면 죠션 인민이 소견과 지혜가 진보홈을 밋노라 ᄂᆞᆫ셜 ᄭᅳᆺ치기 젼에 우리가 대균쥬 폐하씌 송덕ᄒᆞ고 만세을 부르ᄂᆞ이다



- 대한매일신보: 1904년에서 1910년까지의 입력 자료 149개 표본이 제공되고 있음.

【날말뭉치(세종계획 홈페이지)】
(대한매일신보 일부)

```
<!DOCTYPE tei.2  SYSTEM "c:\sgml\dtd\tei2.dtd" [
    <!ENTITY % TEI.corpus "INCLUDE">
    <!ENTITY % TEI.extensions.ent SYSTEM "sejong1.ent">
    <!ENTITY % TEI.extensions.dtd SYSTEM "sejong1.dtd">
]>
<tei.2>
  <teiHeader>
  <fileDesc>
  <titleStmt>
  <title>대한매일신보, 전자 파일</title>
  <author>대한매일신보 편집부</author>
  <sponsor>대한민국 문화관광부</sponsor>
  <respStmt><resp>국립국어연구원 말뭉치 입수, 표준화, 헤더 붙임</resp>
    <name>국립국어연구원</name>
  </respStmt>
  </titleStmt>
  <editionStmt>
    <edition><date>1995/09/01</date>전산입력</edition>
  </editionStmt>
   <extent>1937 어절</extent>
  <publicationStmt>
  <distributor>국립국어연구원</distributor>
  <idno>aa000111.hwp(Pn075230.hwp)</idno>
  <availability><p>배포 불가</p></availability>
  </publicationStmt>
  <notesStmt>
  <note><p>이 텍스트는 국립국어연구원 작업 내용 유지</p></note>
  </notesStmt>
  <sourceDesc>
  <bibl><author>대한매일신보 편집부</author>
    <title>대한매일신보</title>
    <pubPlace></pubPlace>
    <publisher>대한매일신보사</publisher>
```

```
  <date>1907/05/23</date>
</bibl>
</sourceDesc>
</fileDesc>
<encodingDesc>
<projectDesc><p>21세기 세종계획 1차년도 말뭉치 구축</p>
</projectDesc>
<samplingDecl><p>파일 변환 정보 없음</p>
</samplingDecl>
<editorialDecl><p>21세기 세종계획 말뭉치 문헌 입력 지침에 따름</p>
</editorialDecl>
</encodingDesc>
<profileDesc>
<creation><date>1907/05/23</date></creation>
<langUsage>
  <language id=KO usage=99>한국어</language>
</langUsage>
<textClass>
  <catRef scheme='SJ21' target='M1100'>신문: 총류</catRef>
</textClass>
</profileDesc>
<revisionDesc>
  <change>
      <date>1995/09/01</date>
      <respStmt>
          <resp>입력자</resp><name>서정목</name>
      </respStmt>
      <item></item>
  </change>
  <change>
      <date>1995/10/01</date>
      <respStmt>
          <resp>교정자</resp><name>서정목</name>
      </respStmt>
      <item>교정</item>
  </change>
  <change>
      <date>1998/10</date>
      <respStmt>
          <resp>프로젝트 책임자</resp><name>김홍규</name>
          <resp>연구원</resp><name>차준경</name>
          <resp>프로그래머</resp><name>이애란</name>
```

```
        </respStmt>
        <item>파일 변환, 세종 21 프로젝트 헤더 붙임, 마킹</item>
    </change>
  </revisionDesc>
  </teiHeader>

<text>
<body>

<pb n='01'>
<head>사셜</head>
<head>국문신보 발간</head>
<p>본긔ᄌᆞ 이 한국사ᄅᆞᆷ을 ᄃᆡᄒᆞ야 ᄒᆞᆫ마ᄃᆡ 말노 질문코져 ᄒᆞ노니 대져 샴천리 강토와 이천만 인구로 ᄌᆞ쥬독립ᄒᆞ지 못ᄒᆞᆯ 걱정이 업거ᄂᆞᆯ 무ᄉᆞᆷ 연고로 오ᄂᆞᆯ날에 나라 권셰를 온젼히 일코 사ᄅᆞᆷ의 권리가 젼혀 업서져 무궁히 비참ᄒᆞᆫ 경우에 빠졋ᄂᆞ뇨 그 원인을 의론컨ᄃᆡ ᄌᆞ리로 한국인이 편리ᄒᆞᆫ 국문을 바리고 편리치 못ᄒᆞᆫ 한문을 숭샹ᄒᆞᄂᆞᆫ 폐막으로 말미암이라 ᄒᆞ노니 모든 한문가에셔ᄂᆞᆫ 혹 이 말에 ᄃᆡᄒᆞ야 노여ᄒᆞ며 괴이히 넉이ᄂᆞᆫ 자도 잇ᄉᆞ려니와 이것ᄉᆞᆫ 한국ᄂᆡ에 큰 마귀의 저희인 즉 일쟝 셜명ᄒᆞ야 벽과치 아니치 못ᄒᆞᆯ지로다</p>
<p>대져 세계 렬국이 각기 졔나라 국문과 국어(나라방언)로 제 나라 정신을 완젼케 ᄒᆞᄂᆞᆫ 긔초를 삼ᄂᆞᆫ 것이어ᄂᆞᆯ 오직 한국은 졔나라 국문을 ᄇᆞ리고 타국의 한문을 숭샹홈으로 졔나라 말ᄭᆞ지 일허ᄇᆞ린 쟈가 만흐니 엇지 능히 졔나라 정신을 보존ᄒᆞ리오 그 국문을 ᄇᆞ리고 한문을 숭샹ᄒᆞᆫ 폐막을 대강 말ᄒᆞ랴면 여러가지라</p>
<p>한가지ᄂᆞᆫ 국문을 비호지 안코 한문만 비홈으로 말고 글이 ᄒᆞᆫᄀᆞᆯᄌᆞ지 못ᄒᆞ야 공부ᄒᆞᄂᆞᆫᄃᆡ 시히 어려우니 만일 평싱 젼문가이 아니면 사ᄅᆞᆷ마다 비호지 못ᄒᆞᄂᆞᆫ 고로 국민의 보통 지식을 기발ᄒᆞᄂᆞᆫ 길이 심히 좁고 ᄯᅩ 한가지ᄂᆞᆫ 비호기 쉽고 쓰기 편ᄒᆞᆫ 국문은 ᄇᆞ리고 비호기 어렵고 쓰기 불편ᄒᆞᆫ 한문을 괴로히 공부홈으로 청츈브터 장을 치고 빅슈가 되로록 경셔를 궁리ᄒᆞ되 혜두가 더욱 막혀가고 실효가 더욱 업셔져셔 졔 집안의 경졔도 ᄒᆞ기 어렵거든 엇지 부국강병ᄒᆞᆯ 능력이 잇스리오 지식이 막히고 실업이 쇠ᄒᆞ고 렴치가 업셔진 거시다 일노 말미암이오 ᄯᅩ 한가지ᄂᆞᆫ 졔나라 국문은 쳔히 넉이며 경히 넉이고 남의 나라 한문은 귀히 넉이며 중히 넉이ᄂᆞᆫ 고로 졔나라를 제가 업수히 보고 남의 나라를 쳐다보ᄂᆞᆫ 노예의 셩품을 양셩ᄒᆞ고 독립의 명의ᄂᆞᆫ 도모지 아지도 못ᄒᆞ니 엇지 독립ᄉᆞ샹이 잇ᄉᆞ리오</p>
```

2. 말뭉치 활용 도구

이 장에서는, 말뭉치 구축을 위한 텍스트의 개념을 이해하고, 이를 다룰 수 있는 텍스트 편집기에 대해 살펴보고, 말뭉치를 가공하기 위한 자동품사 주석기와 그 활용, 말뭉치를 이용한 용례의 검색, 말뭉치에서 언어정보를 계량화하기 위한 유틸리티와 그 활용 방법에 대해 설명한다.

이와 같은 내용의 실습 내용을 구성할 각 프로그램에 대해서는, 1). 프로그램의 개요(프로그램의 개발 목적, 개발 기관 또는 개인 소개, 개발 시기 등, 관련 사이트 정보)를 소개한 뒤, 기본적인 사용법과 옵션에 대해 설명한다.

특히, 제3부에 소개될 '실습용 말뭉치'를 사용하여 각종의 사용법을 쉽게 익힐 수 있도록 한다.

한편, 각 내용에 필요한 프로그램은, 21세기 세종 계획 등을 통하여, 일반 공개되어 있는 각종 공개 프로그램을 비롯하여, 각 연구 기관에서 개발하여 사용중인 프로그램 가운데에서 각 기능별로 대표적인 몇 가지의 도구로 실습을 할 수 있도록 할 것이다.

2.1 말뭉치 구축을 위한 텍스트의 개념 이해와 텍스트 편집기

이 장에서는 말뭉치의 처리에 있어서 가장 기본적인 개념인 '텍스트 파일'의 개념을 이해하고, 실제로 각종 문서 편집기 등을 활용하여, 목적에 맞는 형태의 텍스트 파일을 생성하고 활용할 수 있게 한다.

2.1.1 텍스트 파일이란?

대량의 텍스트 자료를 분석하여 원하는 결과를 얻어내기 위해서는 컴퓨터의 사용이 절대적으로 필요하다. 0과 1만을 인식하는 컴퓨터가 인간의 언어로 되어 있는 텍스트의 구조와 내용을 분석하기 위해서는 종이 문서나 녹음 테이프 등의 형태가 아닌, 컴퓨터가 판독할 수 있는 부호체계로의 변환이 가능한 텍스트가 제공되어야 한다. 이러한 기계 가독형 텍스트는, 주로 문서 편집기(워드프로세서나 텍스트 편집기, 웹문서 에디터 등)으로 입력되고 작성된 텍스트[17] 파일을 들 수 있다. 이들은 외부적으로는 자연 언어(문자)로 표현되지만 내부적으로는 컴퓨터 코드로 구성되며 따라서 문서 편집 이외의 전산 처리가 가능한 기계 가독형 텍스트이다.

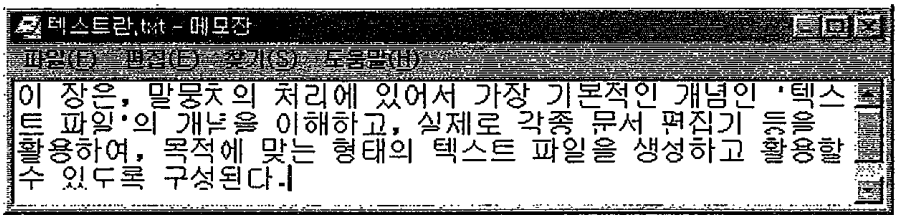
2.1.2 텍스트 파일의 편집

말뭉치로 활용하기 위해 각종 자료를 입력 또는 편집하려는 경우, 대개 문서 작성기나 텍스트 전용 편집기를 이용하게 된다.[18] 텍스트 파일을 다룰 때에는 굳이

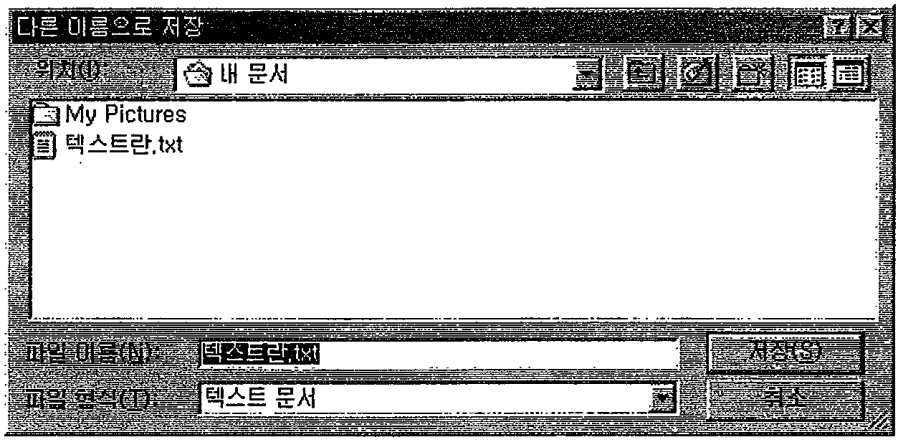
17) 언어학적으로는, 텍스트는 "A text is a sequence of paragraphs that represents an extended unit of speech."로 정의되지만, 여기서 말하는 텍스트는, 파일의 저장 방식에 따른 명칭을 가리킨다. 참고로, "표준국어대사전"(1999)에 의하면, 텍스트(text)란 "①주석, 번역, 서문 및 부록 따위에 대한 본문이나 원문. ②(언어) 문장보다 더 큰 문법 단위. 문장이 모여서 이루어진 한 덩어리의 글을 이른다."로 정의되어 있다.

18) 간혹 공개되어 배포되는 말뭉치를 내려받거나(예컨대, 국립국어연구원의 세종 계획 홈페이지인 http://www.sejong.or.kr의 자료실), 컴퓨터 통신이나 인터넷 게시물 등의 문서를 내려받아 말뭉치로 구성하는 경우가 있는데, 이럴 경우에는 단순 입력 작업은 생

서식이나 글꼴 등의 정보는 필요 없는 경우가 많기 때문에, 아래와 같이 가장 단순한 형태의, 윈도의 메모장을 활용할 수 있다. 다만, 텍스트의 크기가 메모장이 다룰 수 있는 크기를 넘어설 경우에는, 별도의 문서작성기나 텍스트 전용 편집기[19]를 활용하게 된다.



<그림 18> '메모장'을 이용한 텍스트 편집



<그림 19> 메모장의 파일 저장 방식

☼연습 문제☼

(1). 통신망 또는 인터넷에서 텍스트 편집기를 내려받아 설치하여 보자.

(2). 텍스트 편집기를 이용하여, 실습용 말뭉치의 텍스트 파일을 열어, 말뭉치의

략할 수 있어 편리하기는 하나, 대개는 교정이나 각종의 가공 단계에서 문서작성기나 텍스트 편집기를 사용해야 하는 점은 마찬가지이다.

19) 텍스트 전용 에디터는, 각 통신망이나 인터넷에서 손쉽게 구해서 사용할 수 있다. 한 예로, www.acrosoft.pe.kr에서는 AcroEdit 0.9.0.3 버전이 공개되어 있다.(2000년 11월 현재)

구조를 이해하자.

(3). 텍스트 파일과 바이너리 파일의 차이에 대해 알아보자.

2.2 말뭉치를 가공하기 위한 자동품사 주석기와 그 활용

2.2.1 품사 주석이란?

일반적으로 날말뭉치를 통해서 얻을 수 있는 정보는 한정되어 있다. 자소나 음절, 어절 등의 사용 실태를 파악하는 것과 같은 경우에는 날말뭉치를 분석함으로써도, 그다지 어렵지 않게, 필요한 정보나 통계를 얻을 수 있다.[20]

그러나 한층 높은 단계의 언어 정보를 얻기 위해서는 날말뭉치를 일정한 원칙에 따라서 가공하여 분석하게 된다. 특정한 어휘나 형태의 사용 빈도, 사용례의 탐색, 두 개 요소 이상의 연접 빈도 등의 다양한 정보를 필요로 하는 경우, 날말뭉치에 일정한 정보를 부가하는 과정을 거침으로써 더욱 정확하고도 효율적으로 얻을 수 있기 때문이다.

말뭉치의 주석(annotation)이란, 날말뭉치(raw corpus)에 언어학적 정보를 첨가하여 주석된 말뭉치(annotated corpus)를 산출해내는 과정이라고 할 수 있다. 이 때, 주석에 의해서 덧붙여지는 언어학적 정보는, 음성·음운, 형태소, 어휘 단위의 정보로부터, 구문, 대화 분석, 화용, 오류 등 매우 다양한 정보를 가리키며, 이들 분석의 단위는 주석의 목적에 따라서 달리 선택하게 된다.

말뭉치 주석은, 정확한 언어 정보의 추출을 위해서뿐 아니라, 주석 말뭉치를 토대로 한 다음 상위 단계에서의 재활용이라는 목적에 의해서도 이루어지게 된다.

한국어 말뭉치의 주석 단계에서 일반적으로 가장 널리 이루어진 것은 '품사 정보 주석'(part of speech tagging : POS tagging) 또는 '문법 정보 주석'(Grammatical annotation)이다.[21]

이것은 주로 띄어쓰기를 기준으로 한 어절단위에 문법적 정보를 부가하는 작업인데, 그 주석 방식 또는 외적인 형태에 따라서 수평적 형식(horizontal format)의 주석과 수직적 형식(vertical format)의 주석이 있다.

---

20) 이 책의 2.4장에서 다룰 각종 도구를 참고 바람.

21) 여기에서는 이를 줄여서 '품사 주석'이라고 부르기로 한다. 실제로는 어절을 단위로 하여, 그 구성 요소를 분석하고, 그 하나하나에 대해서 문법 정보(품사 표지)를 붙여 나가는 작업을 뜻하므로, '어절 분석', '형태소 분석' 등의 여러 가지 용어로 섞어 쓰기도 한다.

문명/ncn+의/jcm 발달/ncpa+은/jxc 현대인/ncn+에게/jca 쉽/paa+고/ecq 편하/paa+ㄴ/etm 삶/ncn+을/jco 많이/a 가르치/pvg+어/ecx 주/px+었/ep+다/ef ./s. 사람/ncn+들/xn+은/jxc 이해/ncpa+하/xpv+기/exn 복잡/ncps+하/xsm+거나/ecc 번거롭/paa+은/etm 것/nbn+을/jco 멀리/a 하/pvg+고/ecc 음식/ncn+이나/jcp 음료/ncn+도/jxc 간편/ncps+하/xpa+고/ecc 깨끗하/pa+ㄴ/etm 것/nbn+을/jco 좋아하/pv+게/ecx 되/pv+었/ep+다/ef ./s.

수평적 형식의 말뭉치의 예

000008200	풍수의	풍수/ncn+의/jcm
000008300	기본	기본/ncn
000008400	원리는	원리/ncn+는/jxc
000008500	땅속에	땅속/ncn+에/jca
000008600	살아	살/pvg+아/ecx
000008700	꿈틀대는	꿈틀대/pvg+는/etm
000008800	정기가	정기/ncn+가/jcs
000008900	있으며	있/paa+으며/ecs

수직적 형식의 주석 말뭉치의 예

2.2.2 자동 품사 주석기와 형태소 분석기

형태소 분석이란, 자연어 처리의 출발점이라고 할 만큼 말뭉치의 이용의 첫 단계에서 중요한 일이라고 말할 수 있다. 형태소 분석은 어절을 해석하여 형태소 단위로 정확히 구분하는 것이 목적이라고 할 수 있다.

예를 들어, '사과를'이라는 어절을 '사과/명사 + 를/조사'로 분석한다든가, '나는'이라는 형태의 어절을 그 쓰임에 따라서 "나/명사 + 는/조사, 나/대명사 + 는

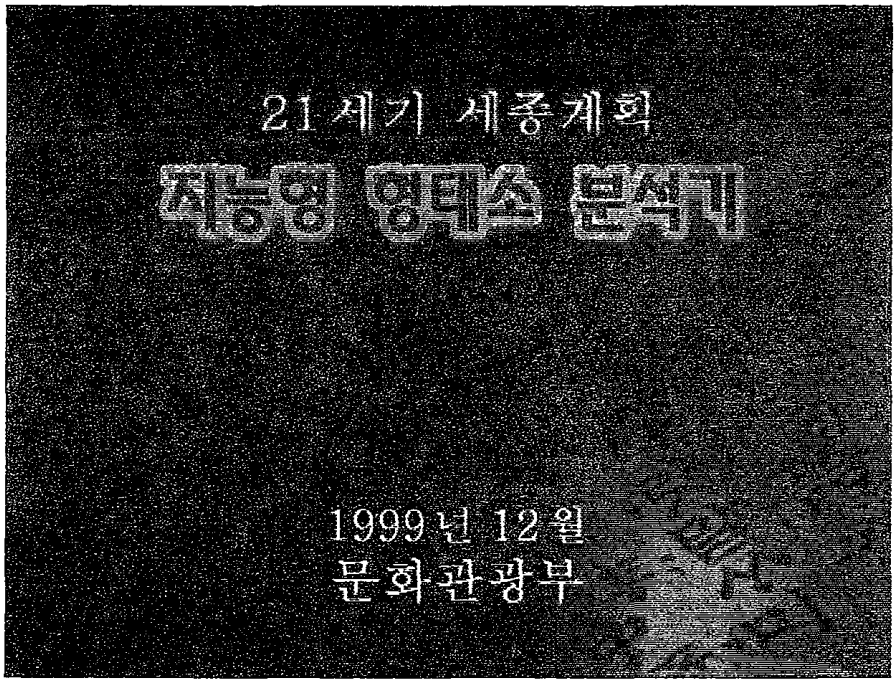
/조사, 나/동사 + 는/어미, 날/동사 + 는/어미" 등으로 분석하는 것을 뜻한다.

이러한 일을 아주 작은 규모의 말뭉치에 적용하는 경우에는 사람이 일일이 언어학적으로 분석하여 그 결과를 써 넣을 수 있겠지만, 대규모의 말뭉치를 대상으로 하는 경우에는 프로그램에 의존할 수밖에 없게 된다. 이와 같은 목적으로 개발된 프로그램에는 형태소 분석기와 자동 품사 주석기(tagger)가 있다.

일반적으로, 형태소 분석기는 해석 가능한 모든 유형을 결과로 제시하는 데 반해, 자동 품사 주석기는 분석 가능한 모든 예들 중에서 가장 가능성 높은 하나를 결과로 제시한다.

또한 태거는 보통 형태소 분석기를 기반으로 하여, 확률 통계적 기법과 같은 다양한 정보를 활용하여, 말뭉치에 주석을 붙이게 된다.

2.2.3 지능형 형태소 분석기



【소개】 "지능형 형태소 분석기"는 단어에 가능한 모든 형태론적 구조를 생성하고 좌우 문맥 정보를 효과적으로 활용하여 올바른 하나의 형태론적 구조를 선택하여 제시할 수 있는 프로그램이다.
【사양】 윈도우즈 95, 98, NT
【개발자】 고려대학교 컴퓨터학과 자연어처리연구실(개발 책임자: 임해창 교수)
【개발시기】 1999년
【관련 사이트】 http://www.sejong.or.kr의 자료실
【설치 방법】

1) CD-rom에 있는 setup.exe를 실행하여, 설치할 폴더를 지정하면, 자동으로 설치된다.
2). 실행 프로그램 :
품사부착도구.exe
규칙획득도구.exe
사전관리도구.exe

【사용법】

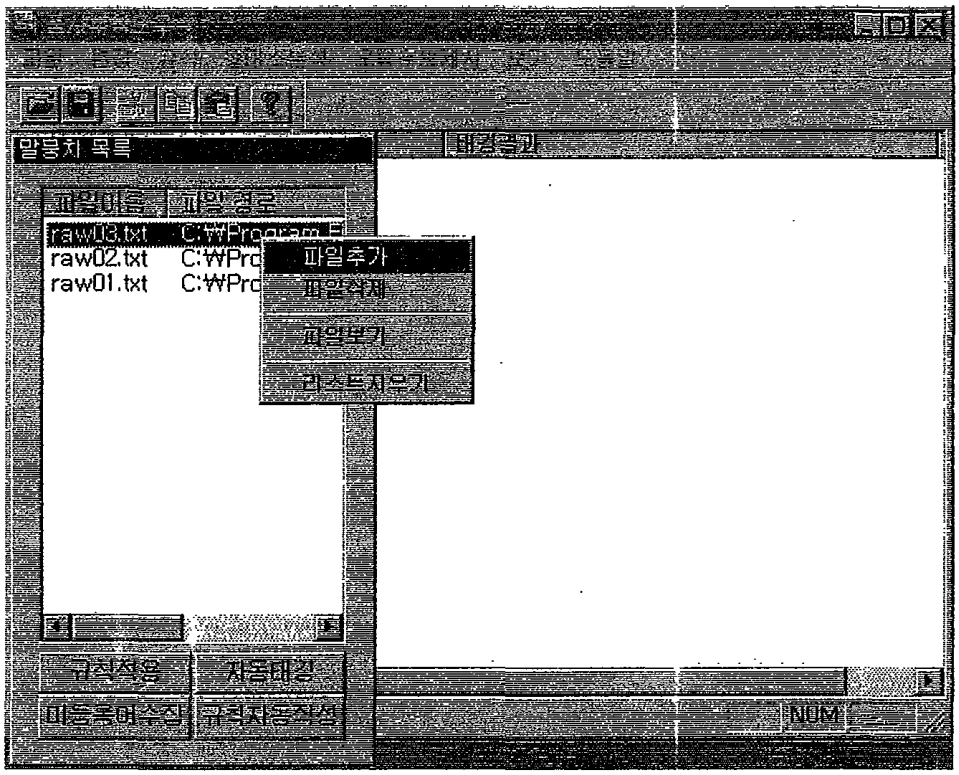
○ 품사 부착 말뭉치 구축 과정

한국어 품사 부착 말뭉치의 구축 과정은 크게 "문서 선택, 전처리, 자동 품사 부착, 오류 교정, 문서 저장"의 과정으로 이루어진다.

① "품사부착도구.exe"를 실행한다.

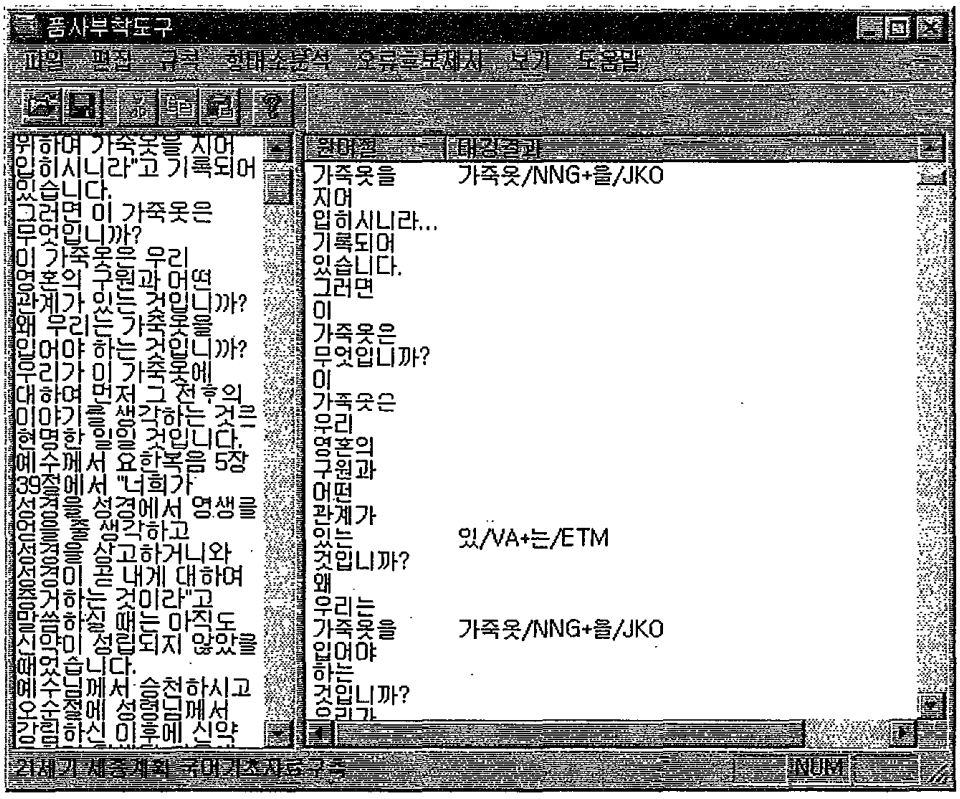
② 먼저 품사를 부착하고자 하는 문서들을 말뭉치 목록에 삽입한다.
- '파일' 메뉴의 '원시 말뭉치 추가'를 선택하거나 툴바의 'open' 버튼을 클릭하여 품사를 부착하고자 하는 원시 말뭉치들을 선택하면 다음과 같이 말뭉치 목록 대화상자에 표시된다.
- 말뭉치 목록 대화상자에서 마우스 오른쪽 클릭시 나타나는 팝업 메뉴를 이용하여 말뭉치를 추가, 삭제, 파일의 내용 보기 등을 할 수 있다. (☞ 입력 가능한 문서 파일 포맷은 완성형 텍스트 문서이다. 훈글 파일이나

워드 파일의 경우 제대로 처리할 수 없으므로, 완성형 텍스트 문서로 변환한 후에 추가하여야 한다.)



<말뭉치 목록 추가>

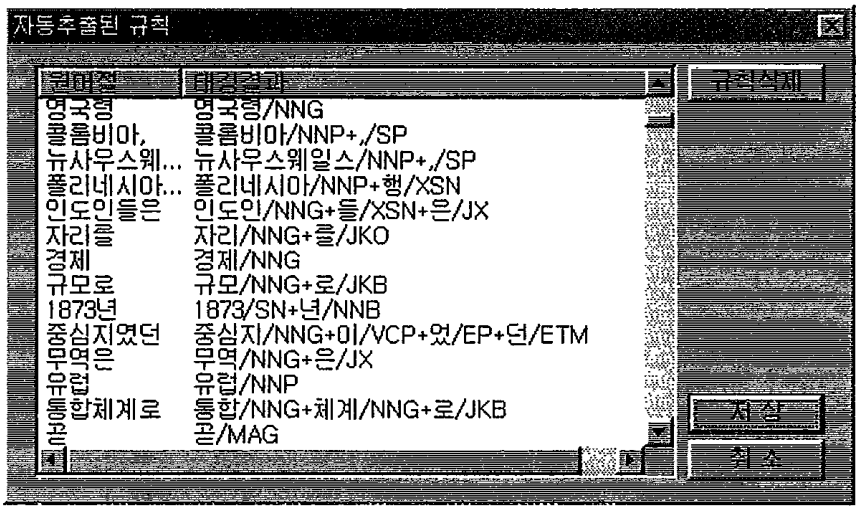
③ 입력된 문서들에 대해 이미 획득되어 있는 규칙을 이용하여 품사를 부착하기 위해 말뭉치 목록 대화상자의 [규칙적용] 버튼을 클릭한다. 규칙에 의해 품사가 부착된 어절을 확인하기 위해서는 말뭉치 목록 대화상자의 파일 이름을 더블클릭하거나 마우스 오른쪽 클릭시 나타나는 팝업 메뉴에서 '파일보기'를 선택한다. 그러면 화면 오른쪽의 리스트 콘트롤에 원어절과 품사 부착된 결과가 출력된다.



<[규칙적용]을 실행한 결과>

④ 규칙이 적용되지 않은 어절들을 대상으로 입력된 문서에 포함된 미등록어 포함 어절들을 수집하기 위해 말뭉치 목록 대화상자의 [미등록어수집] 버튼을 클릭한다. 미등록어 목록을 저장할 파일 이름을 설정하면 말뭉치에 포함된 미등록어 포함 어절들을 수집하기 시작한다.

⑤ 제시된 미등록어 포함 어절 목록을 이용하여 미등록어를 사전에 추가하거나 미등록어 포함 어절을 올바르게 품사 부착할 수 있는 규칙을 작성한다. 미등록어를 사전에 추가하기 위해서는 '사전관리도구'를 이용하고 미등록어 포함 어절에 대한 규칙을 획득하기 위해서는 '규칙획득도구'를 이용한다.



<형태소 분석 결과가 1개인 어절>

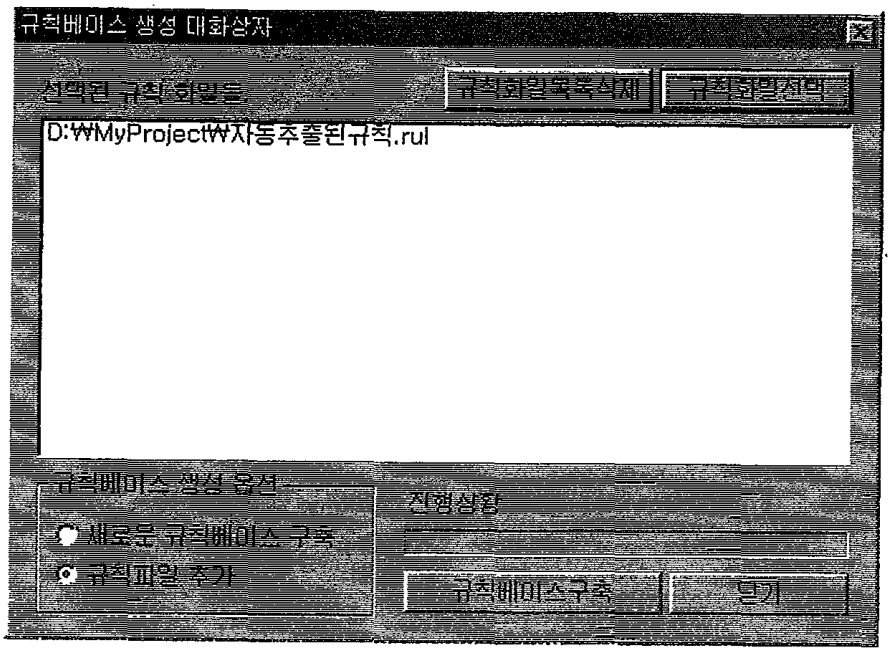
⑥ 형태소 분석 결과가 1개인 어절들은 좌우 문맥에 관계없이 품사 부착이 가능하며, [규칙자동작성] 버튼을 클릭하면 이러한 어절들의 목록을 확인할 수 있다.

- 사용자는 이들 목록에서 잘못 분석된 어절들을 삭제할 수 있다.
- 사용자의 확인을 거쳐 남은 어절들에 대해 규칙을 만들기 위해서 '저장' 버튼을 클릭하고 규칙을 저장할 외부 파일을 지정한다.
- 이들 규칙을 규칙베이스에 추가하기 위해서는 품사부착도구의 '규칙' 메뉴에서 '규칙베이스 구축' 메뉴를 선택한다.

- [규칙화일선택] 버튼을 클릭하여 획득된 규칙 파일을 지정하면 화면 중앙의 리스트에 선택된 규칙 파일이 표시된다. 생성 옵션을 '규칙파일추가'에 맞추어 놓은 상태에서 '규칙베이스구축' 버튼을 클릭하면 작성된 규칙들이 규칙베이스에 삽입된다.
- '규칙베이스 생성 대화상자'를 닫고 다시 한번 [규칙 적용] 버튼을 클릭하여 파일의 내용을 확인하면 새로이 획득된 규칙에 의해 품사가 부착된다.

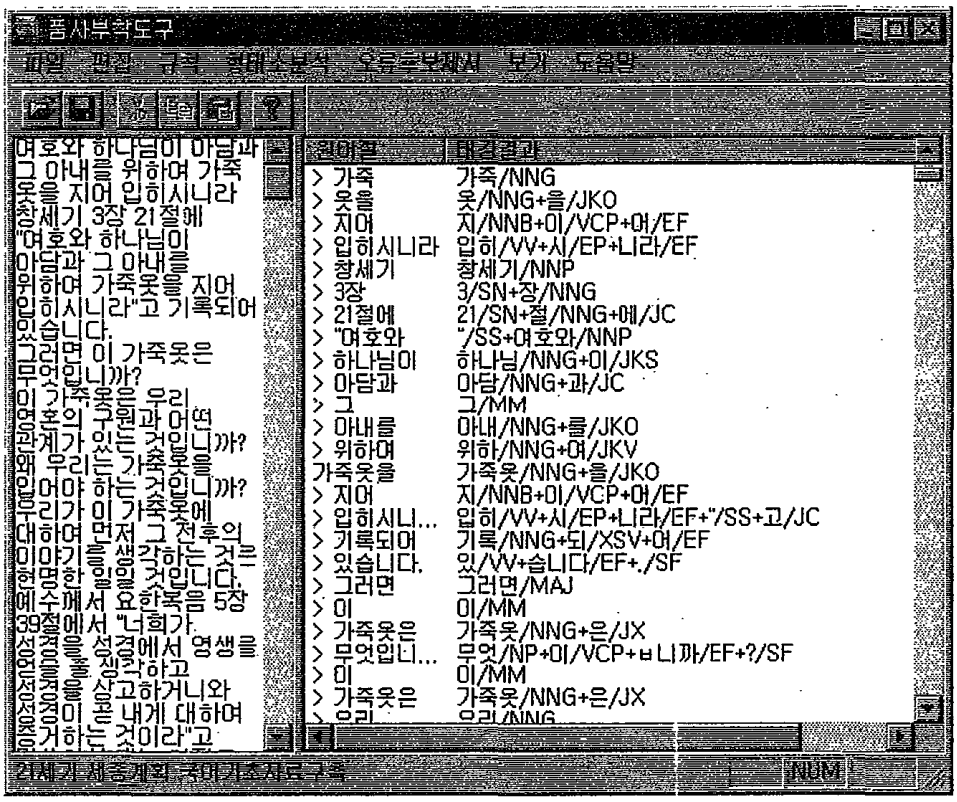
⑦ 규칙이 적용되지 않아 품사가 부착되지 않은 어절들을 통계 기반의 자동 태거

를 이용하여 입력된 문서의 모든 어절에 품사를 부착할 수 있다.



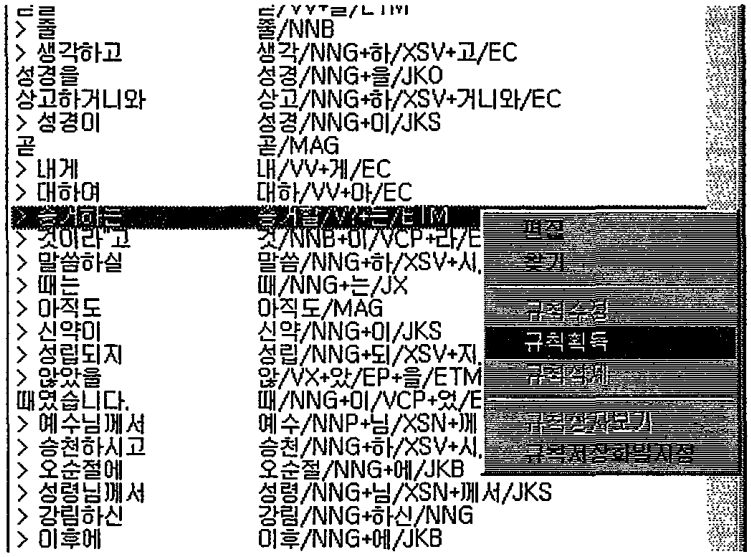
<획득된 규칙을 규칙베이스에 추가>

- 말뭉치 목록 대화상자의 [자동태깅] 버튼을 클릭하면 자동 품사 부착이 진행된다.
- 잠시 후 자동 태깅이 끝났음을 알리는 메시지가 나타나면 파일 이름을 더블클릭하거나 마우스 오른쪽 클릭시 나타나는 팝업 메뉴에서 '파일보기'를 선택하여 자동 품사 부착된 결과를 확인할 수 있다.
- 규칙에 의해 품사가 부착된 어절들과 달리 통계 태거에 의해 품사가 결정된 어절들은 원어절 표시란에 '>' 표시가 자리하게 된다. 따라서 규칙의 정확성이 매우 높다고 가정할 경우 '>' 표시가 붙은 어절들만을 대상으로 오류 수정 작업을 진행하는 방법도 가능하다.



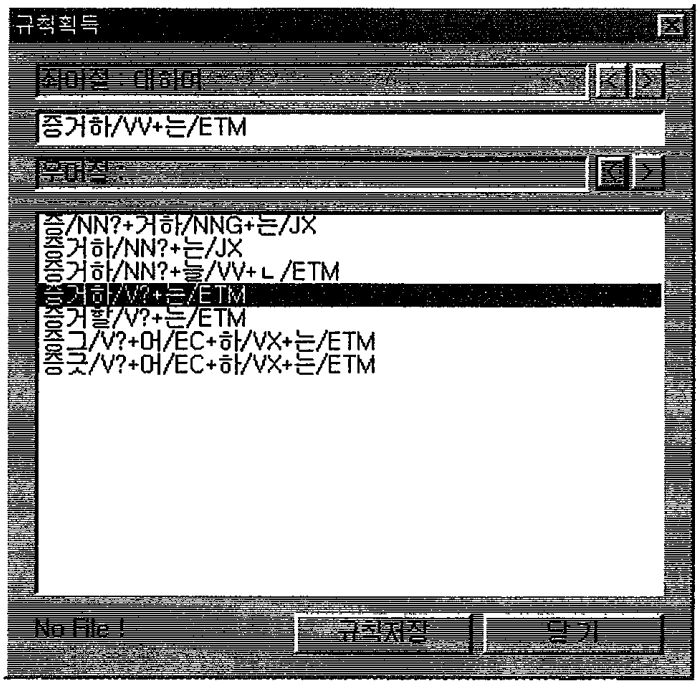
<[자동태깅]을 실행한 결과>

⑧ 오류 교정 작업을 수행한다. 오류 교정 작업은 잘못 태깅된 어절에 대해 이를 올바르게 태깅할 수 있는 규칙을 획득하거나 직접 결과를 편집하는 과정으로 이루어진다. 사용자 편리성을 위해 수정로그를 이용하여 오류 가능 어절을 자동으로 제시해 준다.



<오류 어절에 대한 규칙 획득>

- 품사 부착된 결과 가운데 오류 어절이 발견되면 마우스 오른쪽 클릭을 통해 나타나는 팝업 메뉴에서 '규칙 획득'을 선택한다.
- 이 오류 어절을 올바르게 교정하고 추후 동일한 문맥상에서 동일한 품사 부착 오류가 반복되는 것을 방지하기 위해 규칙을 획득하여 저장한다. 이를 위해 '규칙 획득 대화상자'에서 품사 부착시 고려해야 할 좌우 문맥을 결정하고 올바른 품사 부착 결과를 지정한 후에 '저장' 버튼을 클릭한다.
- 이때 획득된 규칙은 규칙베이스에 바로 저장이 되며 규칙 적용 버튼을 클릭하여 동일한 문맥상에 존재하는 동일한 오류 어절들을 자동으로 올바르게 수정할 수 있다.



<규칙획득 대화상자>

⑨ 입력된 말뭉치의 각 어절과 이들에 대한 형태소-품사쌍열을 외부 파일로 저장한다.

- 품사 부착된 말뭉치를 외부 파일로 저장하기 위해서는 '파일' 메뉴의 [품사태깅된말뭉치저장]이나 툴바의 [Save] 버튼을 클릭한다.
- 외부 파일 이름을 저장하면 현재 화면의 오른쪽 리스트콘트롤의 내용이 다음과 같은 형태로 저장된다.

```
벨기에는          벨기에/NNP+는/JX
콩고를   콩고/NNP+를/JKO
차지했고,         차지/NNG+하/XSV+았/EP+고/EC+,/SP
포르투갈은        포르투갈/NNP+은/JX
앙골라와          앙골라/NNP+와/JC
모잠비크에        모잠비크/NNG+에/JKB
자리를   자리/NNG+를/JKO
잡았다.  잡/VV+았/EP+다/EF+./SF

중국은   중국/NNP+은/JX
유럽인들에        유럽인/NNG+들/XSN+에/JKB
의해서   의하/VV+아서/EC
광동(廣東)에서   광동/NNP+(/SS+廣東/SH+)/SS+에서/JKB
톈진까지의        톈진/NNP+까지의/JKG
지역에서          지역/NNG+에서/JKB
대륙의   대륙/NNG+의/JKG
항구를   항구/NNG+를/JKO
강제로   강제/NNG+로/JKB
양도해야          양도/NNG+하/XSV+아야/EC
했고,    하/VX+았/EP+고/EC+,/SP
연안지방의        연안/NNG+지방/NNG+의/JKG
토지권을          토지권/NN?+을/JKO
포기해야만        포기/NNG+하/XSV+아야만/EC
했다.    하/VX+았/EP+다/EF+./SF

철도     철도/NNG
부설권과          부설권/NNG+과/JC
무역     무역/NNG
```

<품사 표지가 부착된 말뭉치>

2.2.4 반자동 품사 주석기 '꼬리달개'

【소개】 '꼬리달개'는 자동 태거인 HAM_KTS를 통해 품사 주석 텍스트의 오류를 교정하기 위한 프로그램이다. '꼬리달개'의 가장 큰 특징은, 어절에 대한 분석 결과를 데이터베이스에 저장하였다가, 다음에 똑같은 어절이 나온 경우 DB를 검색하여 분석을 수정할 수 있으므로 분석에 대한 일관성을 유지할 수 있다는 것이다.

【사양】 윈도우9x용으로 개발되었다.

【구성 파일 및 설치】 프로그램 설치프로그램은 C:\Program Files\꼬리달개 에 다음과 같은 파일들을 만들게 된다.

(1). 꼬리달개.exe : 실행파일. 프로그램 설치를 완전히 마치더라도 예상했던 '꼬리달개' 라는 그룹이 시작 메뉴에 생기지 않으므로, 직접 실행 파일을 찾아가서 실행을 해야 하는데, 단축 아이콘을 만들어 놓는 것이 좋다.

(2). Eojul.DB, Bunsuc.DB : 한 번 나온 어절과 그에 따른 분석을 저장하는 데이터 파일이다. 처리를 하면 할수록 파일의 크기가 커지고, 가지고 있는 내용이 충실해진다.

【개발자】 연세대학교 언어정보개발연구원

【개발 시기】 1999년

【사용 방법】

○ 파일 읽기

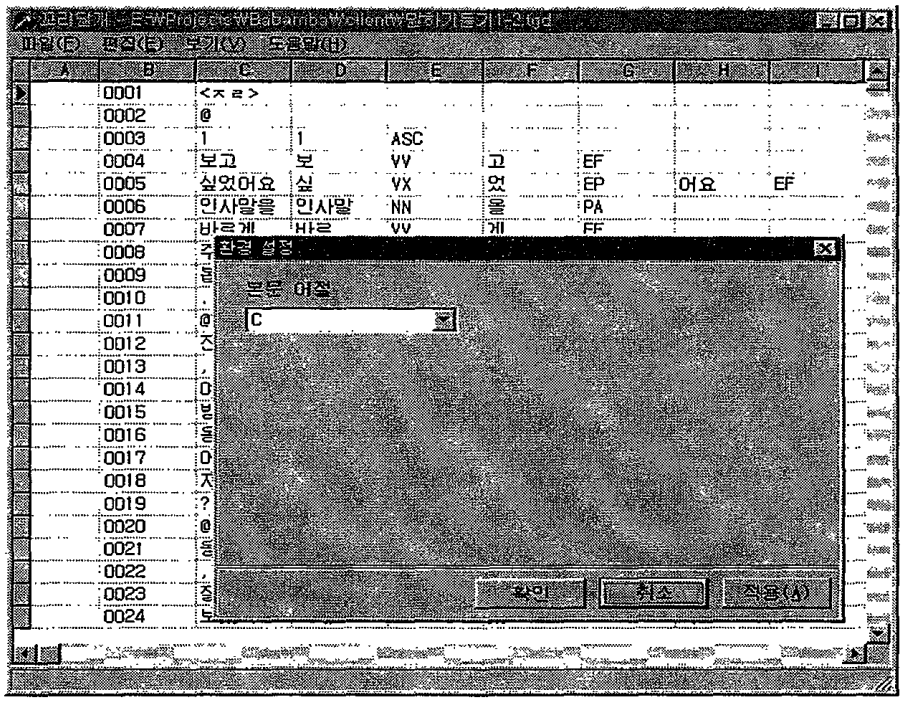
(1). 주석할 말뭉치를 불러오는 것으로 시작한다. 파일 메뉴에서 열기를 선택하거나 단축키 Ctrl-O를 사용하여도 좋다.

(2). 불러 올 파일 형식은 현재 버전에서는 텍스트 파일만을 읽을 수 있다. 텍스트 파일은 구분자가 tab이어야만 한다.

○ 환경 설정

작업을 위한 파일을 불러온 뒤, 먼저 환경 설정을 해야 한다. 환경 설정에서는 기

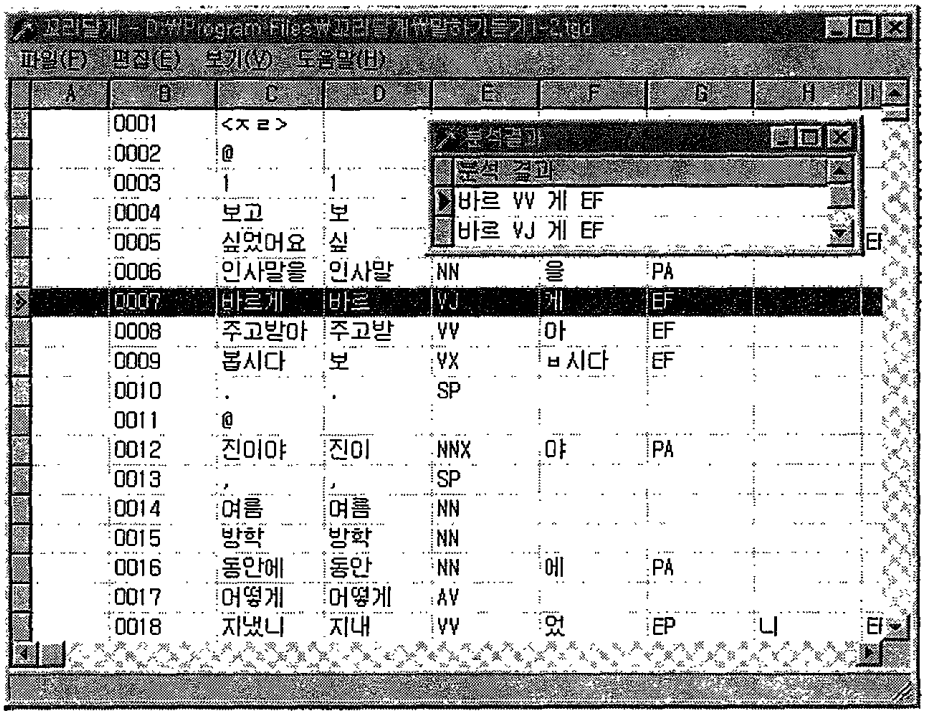
준이 되는 본문 어절이 몇 번째 칸인가를 정해 준다. 아래 그림의 예에서는 C번 칸이 기준이 되는 어절이다.



○ 분석 결과 창

분석 결과 창은 지금까지 해당 어절에 대한 분석을 어떻게 했나를 보여 주는 창이다.

아래의 그림에서는 '바르게'라는 어절이, 지금까지 '바르 VV 게 EF'와 '바르 VJ 게 EF'의 두 가지로 분석이 되었음을 나타내며, 이 중 올바른 것을 선택하면 작업 파일에 분석 결과로써 기입된다.



(1). 분석 결과 창 나타내기

분석 결과 창을 나태내기 위해서는 메뉴의 '보기'-'분석결과'를 클릭하면 된다. 한 번 더 클릭하면 분석 결과 창이 사라진다.

(2). 숫자로 결과 바꾸기

위의 그림에서 분석 결과를 '주고받 VV 아 EF'로 하려면 Ctrl키와 함께 숫자 1을

누르면 된다. '주 VX 고 EF 받 VX 아 EF'로 하려면 Ctrl키와 함께 숫자 2를 누르면 된다.

(3). Ctrl-Tab으로 포커스 바꾸기

사실 숫자를 입력하는 것이 그리 쉬운 일은 아니다. 그리고, 키보드 상에서 Ctrl 키와 숫자가 있는 곳은 굉장히 멀다. 또한 대부분의 사람이 숫자를 오른 쪽에 따로 마련되어있는 키패드를 사용하기 때문에 작업할 때 쓸데없는 손의 이동이 생기게 마련이다. 이런 이유로 Ctrl키와 숫자를 누르기가 싫은 사람은 Ctrl + Tab을 사용하자. 그러면 편집 창에서 분석 결과 창으로 초점(focus)이 옮겨진다. 그리고 나서는 화살표 키(↑ ↓)를 이용하여 원하는 결과를 선택할 수 있다.

(4). 분석 결과 추가하기

원하는 결과가 분석 창에 없는 경우는 두 가지로 볼 수 있다.

첫째는 편집 창에는 분석이 제대로 되어 있는데, 분석 결과 창은 비어 있거나, 다른 결과만 있는 경우이다. 이럴 때는 편집 창에서 Enter↵키를 한 번 쳐 주면 편집 창의 결과가 분석 결과 창에 입력이 된다.

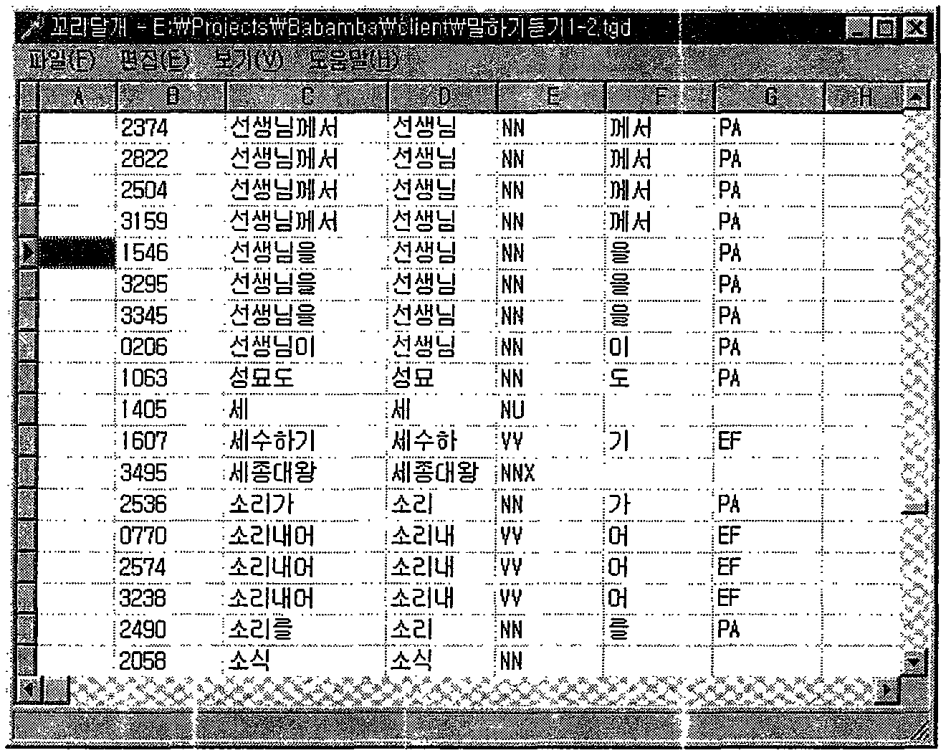
둘째는 편집 창에도 제대로 돼 있지 않고 분석 결과 창에도 올바른 분석이 없을 때인데, 이럴 때는 편집 창에서 엑셀에서처럼 셀을 하나 하나 고치고 나서 Enter↵를 쳐 주는 방법과, 분석 결과 창에서 'E'를 치고 나서 '분석 수정' 창이 생기면 스페이스를 구분자로 해서 입력을 하는 방법이 있다. 두 가지 다 자동으로 한글/영어를 전환해 주니 한/영 전환키를 누를 필요는 없다. 두 가지 방법 중에 편한 것을 사용하면 될 것이다.

○ 찾기, 바꾸기

편집 메뉴에서 찾기, 바꾸기를 선택하면 된다. 혹은 'Ctrl + F', 'Ctrl + R' 단축키를 사용해도 된다. 사용방법은 아래아 한글을 비롯한 대부분의 워드프로세서와 비슷하다.

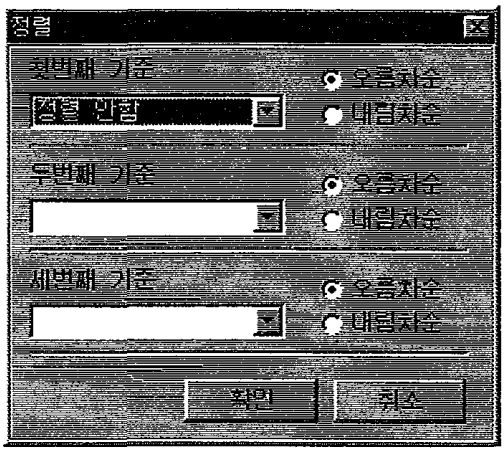
○ 정렬하기

편집 메뉴에서 정렬을 선택하면 된다. 정렬된 결과를 가지고 일괄적으로 분석 결과를 고칠 수도 있다.



A	B	C	D	E	F	G	H
	2374	선생님께서	선생님	NN	께서	PA	
	2822	선생님께서	선생님	NN	께서	PA	
	2504	선생님께서	선생님	NN	께서	PA	
	3159	선생님께서	선생님	NN	께서	PA	
	1546	선생님을	선생님	NN	을	PA	
	3295	선생님을	선생님	NN	을	PA	
	3345	선생님을	선생님	NN	을	PA	
	0206	선생님이	선생님	NN	이	PA	
	1063	성묘도	성묘	NN	도	PA	
	1405	세	세	NU			
	1607	세수하기	세수하	VV	기	EF	
	3495	세종대왕	세종대왕	NNX			
	2536	소리가	소리	NN	가	PA	
	0770	소리내어	소리내	VV	어	EF	
	2574	소리내어	소리내	VV	어	EF	
	3238	소리내어	소리내	VV	어	EF	
	2490	소리를	소리	NN	를	PA	
	2058	소식	소식	NN			

원본 형태로 돌아가기 위해서는 '정렬 안함'을 선택하면 된다.



○ 저장하기

작업을 마치면 저장을 한다. 파일 형식은 구분자를 탭으로 사용하는 텍스트 파일이다.

○ 품사 표지 부착 말뭉치 보기

0001	<ㅈㄹ=말하기/듣기 1학년 1학기>				
0002	@				
0003	1	1	ASC		
0004	우리	우리	NP		
0005	@				
0006	자연스러운	자연스럽	VJ	은	EF
0007	자세로	자세	NN	로	PA
0008	말하여	말하	VV	여	EF
0009	봅시다	보	VX	ㅂ시다	EF
0010	.	.	SP		
0011	@				
0012	친구에게	친구	NN	에게	PA
0013	나에	나	NP	에	PA
0014	대하여	대하	VV	여	EF
0015	자연스러운	자연스럽	VJ	은	EF
0016	자세로	자세	NN	로	PA

2.3 말뭉치를 이용한 용례의 추출과 활용

2.3.1 말뭉치와 용례 추출이란?

용례 추출(concordance)이란, 말뭉치에서 연구자 자신이 원하는 용례를 찾아내는 일을 가리킨다. 특정한 목적 아래 구축된 말뭉치에서 찾고자 하는 용례를 효과적으로 찾아내는 일이야말로, 말뭉치 자료 분석의 가장 기초가 되는 작업이라 할 수 있다.

용례 추출은, 텍스트에서 목적 문자열에 맞는 부분을 찾아내는 일이라 할 수 있는데, 그 결과를 최대한 사용자에게 편리하도록 일정하게 가공해서 보여 주는 것이 일반적이다. 이러한 용례 추출 결과를 보여 주는 방법으로는 KWIC(Keyword In Context)과 KWAL(Keyword And Line), KWOC(Keyword Out of Context)와 같은 다양한 방법이 사용되고 있다.

KWIC은 '주요어 문맥 색인'으로 불리는데, 검색어(Keyword)를 중심으로 앞뒤의 문맥을 볼 수 있도록 배열하여 보여 주는데, 그 결과를 다양하게 응용할 수 있어, 가장 널리 사용되는 방법이다.

2.3.2 용례 검색기 '글잡이 1.0'(윈도용)

【소개】 '글잡이 1.0'은 한국어 용례 추출기로서, 품사 부착 말뭉치로부터 용례 및 빈도 정보를 획득하는 작업을 지원하는 도구이다. 문화관광부 지원으로 이루어진 '21세기 세종 계획'의 기초언어자료 분과에서 1998년 과제의 일부로서 개발한 용례 검색 프로그램이다.

【사양】 Microsoft Windows 95/98, MS-DOS 6.2 이상의 버전, 16M DRAM, 56.6MBytes HDD (예제 실행이 가능한 최소 공간), 마우스

【개발자】 고려대학교 컴퓨터학과 자연어처리연구실(개발 책임자: 임해창 교수)

【개발시기】 1998년

【관련 사이트】 http://www.sejong.or.kr의 자료실

【설치 방법】

1) CD-rom에 있는 setup.exe를 실행하면 창이 하나 나오고, 창에는 설치할 디렉토리를 입력할 수 있게끔 되어 있다. 이 곳에 설치할 디렉토리를 입력한 후, "설치" 버튼을 누르면 설치된다.
2) 설치 후에, 다른 설정의 변경 없이 바로 사용할 수 있다.
3) 설치가 끝나면, 설치한 디렉토리 밑에 "Dos용"과 "Windows용"이라는 디렉토리와, "manual.hwp"와 "liner.exe" 화일이 생기게 된다.
4). 실행 프로그램 : "Windows용"이라는 하위 디렉토리 밑의 "geul.exe"가 윈도용 실행 파일이다.[22]

【사용법】 이 프로그램은, 원하는 말뭉치를 대상으로 미리 색인을 하고 이를 대상으로 용례를 검색할 수 있도록 하였으며, 용례와 함께 각종 빈도 정보 제공일반 연구자들을 위한 다양한 검색 옵션 제공하고 있다.

여기에서는, 첨부된 예제 말뭉치[23]를 이용하여, 기본적인 사용법을 익히도록

---

22) dos용으로 쓸 수 있는 프로그램의 이름과 기능은 다음과 같다.

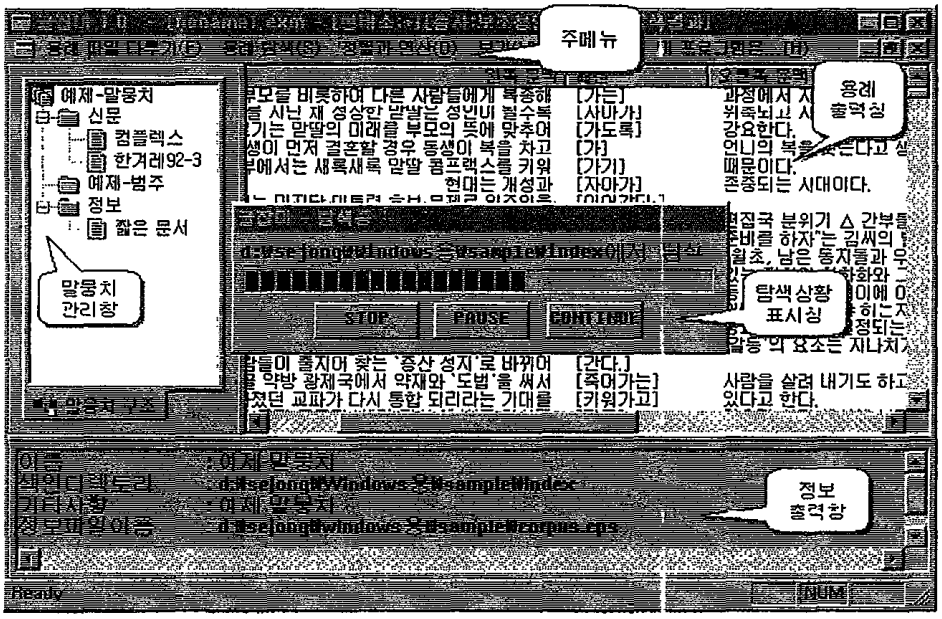
hantags.exe : 문서 태깅
pre.exe : 문서 전처리
index.exe : 문서 색인
ee.exe : 색인된 문서에서 어절용례 탐색
me.exe : 색인된 문서에서 형태소용례 탐색
fre.exe : 색인된 문서에서 빈도정보 추출
eed.exe : 색인되지 않은 문서에서 어절용례 탐색
med.exe : 색인되지 않은 문서에서 형태소용례 탐색
fred.exe : 색인되지 않은 문서에서 빈도정보 추출
select.exe : 탐색된 용례에 대한 연산 수행
sort.exe : 탐색된 용례에 대한 정렬 수행
viewpos.exe : 품사태그와 이름에 관한 정보

23) 예제 말뭉치의 이름은 "예제-말뭉치"이다. 여기에는, 각각 "신문"과 "정보"라는 이름의 범주가 있고, 신문" 범주 아래에는 "한겨레92-3"이라는 문서가 하나 등록되어 있으며, "정보" 범주 아래에는 "컴플렉스"라는 문서가 하나 등록되어 있다.

"한겨레92-3"과 "컴플렉스" 문서는 본 프로그램과 함께 제공되는 문장분리기(liner.exe)의 과정을 거친 후에 각각의 범주 아래에 등록이 되었으며, 품사 표지가 부착된 문서들이며, 색인 정보도 아울러 가지고 있다.

한다.

○ 전체 화면 설명



<그림 35> 전체 화면 구성

- 주메뉴 : 용례 파일 관리 등 프로그램 사용에 필요한 메뉴를 말한다.
- 말뭉치 관리 창 : 현재 작업중인 말뭉치를 표시하며 이 창에서 오른쪽 마우스를 클릭하면 본 프로그램에서 제공하는 거의 모든 작업을 할 수 있도록 메뉴가 화면에 나타난다.
- 용례 출력 창 : 사용자가 입력한 질의어에 해당하는 용례를 찾아 보여준다.
- 탐색 상황 표시 창 : 현재의 용례 탐색 및 연산, 정렬 상황을 진행막대그래프로 표시해 주고, '중단(STOP)/일시정지(PAUSE)/계속(CONTINUE)' 기능이 가능하도록 하였다.
- 정보 출력 창 : 각 말뭉치, 범주, 문서에 대한 정보나 찾아진 용례의 문맥을 보여준다.

○ 말뭉치 전처리 : 헤더-마크업 제거 및 문장 분리기 (liner.exe)

용례 추출기에서는 기본적으로 입력 파일이 문장 단위로 Enter(↵)가 입력된 텍스트 파일임을 가정한다. 이 프로그램은 용례 추출기(글잡이 - guel.exe)에서 사용할 수 있도록 말뭉치 파일을 변환하는 데 사용되는 프로그램으로 CD롬에 함께 저장되어 있으며, 설치할 때 용례 추출기가 설치되는 디렉토리에 자동적으로 복사된다.

따라서 21세기 세종계획을 통해 구축된 말뭉치의 경우는 우선 한글(hwp)파일을 텍스트로 저장한 후, 본 프로그램을 사용하여 헤더-마크업을 제거하고 문장단위로 Enter가 입력되도록 변환한 후에 변환된 파일(.mst)을 사용해야 한다.

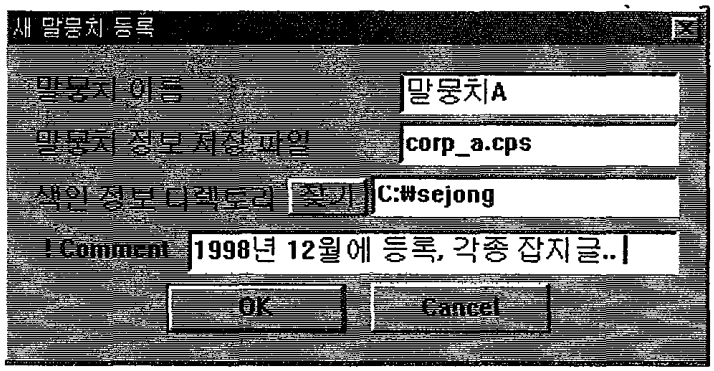
* 사용 방법 *

1). liner.exe 를 실행시킨다.
2). 파일 추가(Alt+A)를 선택하여 변환하려는 텍스트 파일들을 선택한다.
(파일 선택 상자에서는 여러 개의 파일을 동시에 선택할 수 있다.)
3). 선택한 파일들이 리스트 박스에 등록된다.
4). 작업 수행(Alt+D)을 선택하면, 화면 하단에 상태 진행 바가 움직이면서 파일 변환 작업이 시작된다.
5). 파일 변환이 종료되면, 리스트 박스에 등록되었던 파일들이 삭제된다.

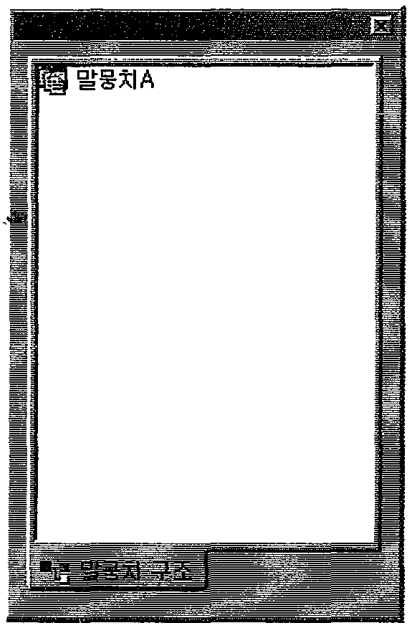
○ 말뭉치 등록과 작업 말뭉치 변경

말뭉치를 등록하려면 말뭉치 관리 창에서 오른쪽 마우스를 클릭하고 [말뭉치 등록]을 선택하면 된다. 그러면 대화 상자가 화면에 나타나는데, 말뭉치 이름은 사용자가 원하는 이름으로 입력하고 말뭉치 정보 파일은 확장자를 .cps로 하여 원하는 이름으로 입력한다.[24]

24) 이 파일은 등록하는 말뭉치가 어떤 범주와 문서를 갖고 있으며 색인 정보 디렉토리는 무엇인지 등에 대한 정보가 기록될 파일로서 실행파일이 있는 디렉토리의 하위디렉토리인 Data 디렉토리에 저장된다. 색인정보 디렉토리는, 지금 등록하는 말뭉치를 색인할 경우, 그 색인정보가 위치할 디렉토리로서 [찾기] 단추를 사용하여 특정 디렉토리를 지정하면 된다. !Comment는 말뭉치에 대한 각종 정보를 사용자가 원하는 방식으로 입력하



새 말뭉치 등록하기



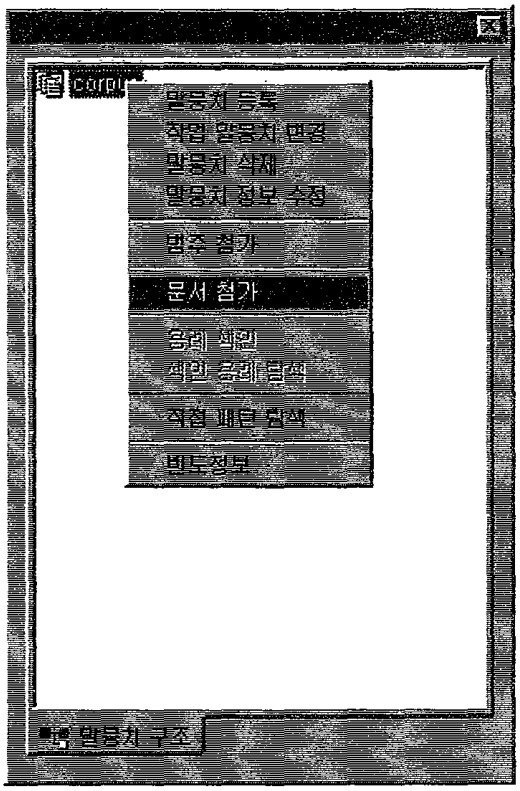
<그림> 말뭉치 등록 후의 말뭉치 관리 창

말뭉치가 새로 등록되면 화면에 등록한 말뭉치가 말뭉치 관리 창에 나타난다.

면 된다.

이로써, 말뭉치에 범주나 문서를 등록하여 용례를 검색할 수 있게 되는 것이다.[25]

이 프로그램에서 사용하기 위해, 곧바로 문서 첨가하기 위해서는, 말뭉치 아이템(최상위 루트노드)에 오른쪽 마우스를 클릭하면 아래 그림과 같은 메뉴가 뜨는데 이때 [문서 첨가]를 선택한다. 그러면 문서 등록 대화 상자가 화면에 나타난다.[26]



오른쪽 마우스를 클릭했을 때 나타나는 메뉴

25) 프로그램이 시작될 때 말뭉치 관리 창에 나타나는 말뭉치는 최근에 프로그램을 종료했을 당시 작업했던 말뭉치이며, 프로그램을 설치한 후 처음 사용할 경우에는 예제-말뭉치가 말뭉치 관리 창에 나타난다.

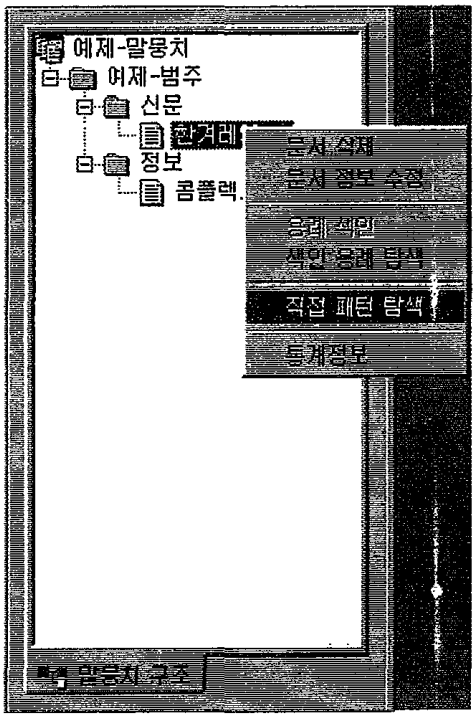
26) 말뭉치 관리 창의 사용법이나, [작업 말뭉치 변경] 등에 관해서는, 프로그램에 부속된 사용 설명서를 참조하면 된다.

프로그램을 설치 후 geul.exe를 실행시키면 화면 왼쪽에 예제 말뭉치 구조가 나타나게 된다. 이제부터 예제 말뭉치를 통해 본 프로그램을 사용하여 용례를 검색하는 방법에 대해 알아보도록 한다.

○ 용례 검색

(1). 직접 용례 탐색

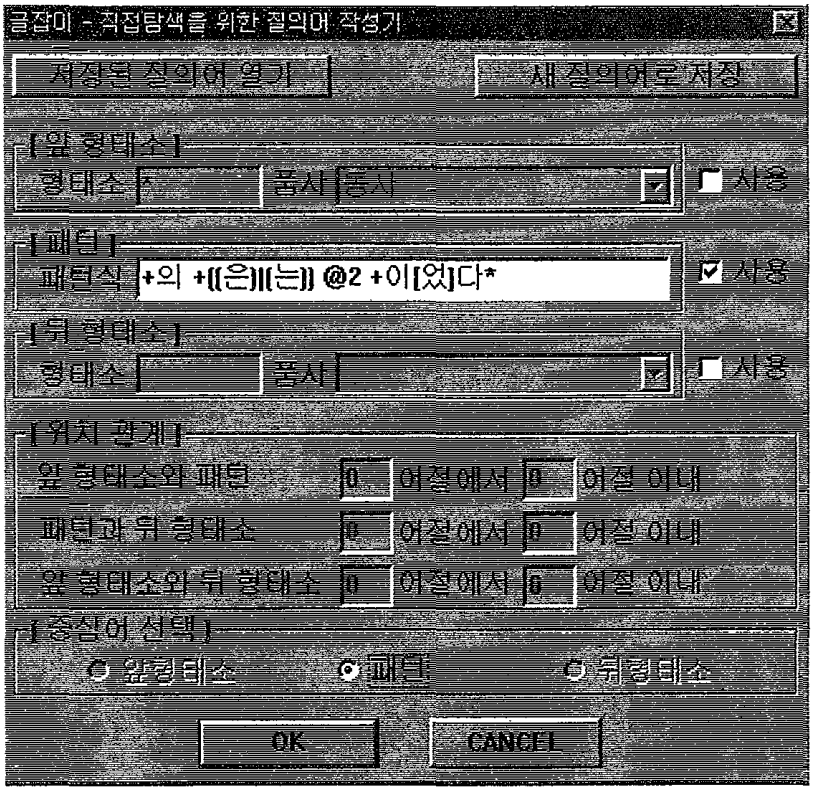
직접 용례 탐색을 사용하여 용례를 탐색하려면 말뭉치 관리 창에서 말뭉치('한겨레 92-3' 문서)에 오른쪽 마우스를 클릭하고 그림과 같이 [직접 패턴 탐색]을 선택한다.



<직접 패턴 탐색 선택>

직접 패턴 탐색을 선택하면 아래의 그림과 같은 질의어 작성 대화 상자가 화면

에 나타나는데, 화면에 그림에서와 같이 입력하고 OK를 클릭한다. 이 질의는 문서 "한겨레92-3"에서, '-의 -은' 또는 '-의 -는' 패턴 뒤에 '-이다' 또는 '-이었다'가 2어절 이내에 나오는 패턴을 탐색하도록 하는 질의이다.



<직접 패턴 탐색의 질의어 입력 상자>

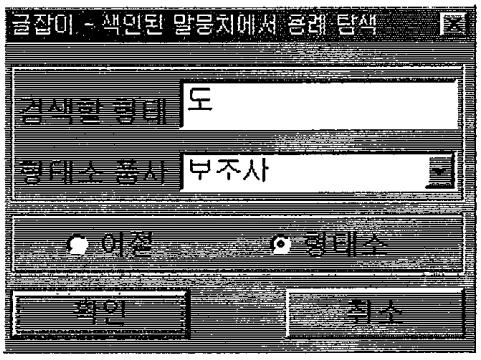
아래의 그림은, 위의 질의에 대한 결과로 화면 우측의 용례 출력 창에 나타나는 용례를 보여주고 있다.

<직접 패턴 탐색의 결과>

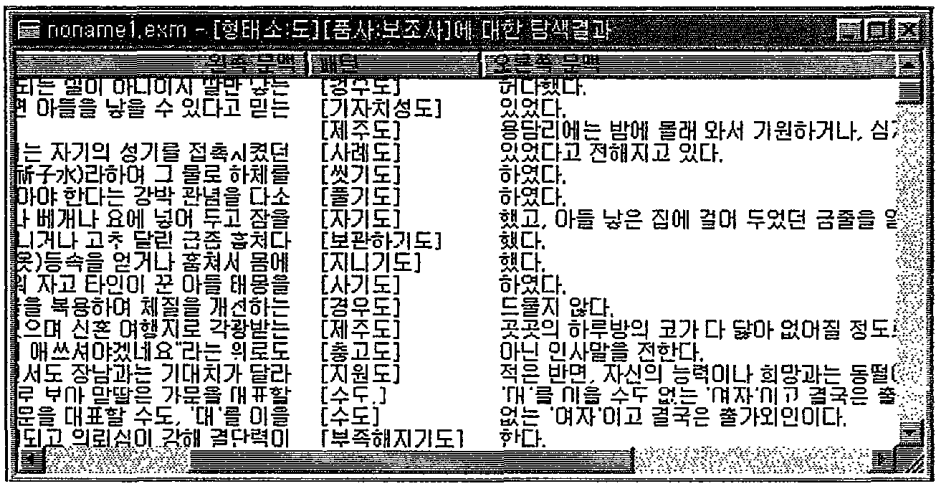
(2). 색인 탐색

색인 용례 탐색은 색인 정보를 활용하여 용례를 탐색하는 작업이다. 이번에는 말뭉치 관리 창에서 '예제-말뭉치'에 오른쪽 마우스를 클릭하고 [색인 용례 탐색]을 선택한다. 이렇게 하면 말뭉치 전체에서 용례를 추출할 수 있게 된다.

[색인 용례 탐색]을 선택하면 아래의 그림과 같은 질의어 입력 대화 상자가 나타난다. 그림에서는 "예제-말뭉치"에서 보조사 '도' 형태소에 대한 용례를 탐색하기 위한 질의어 입력과 그 결과를 보여주고 있다.



<질의어 입력 상자>



noname1.exm - [형태소:도][품사:보조사]에 대한 탐색결과

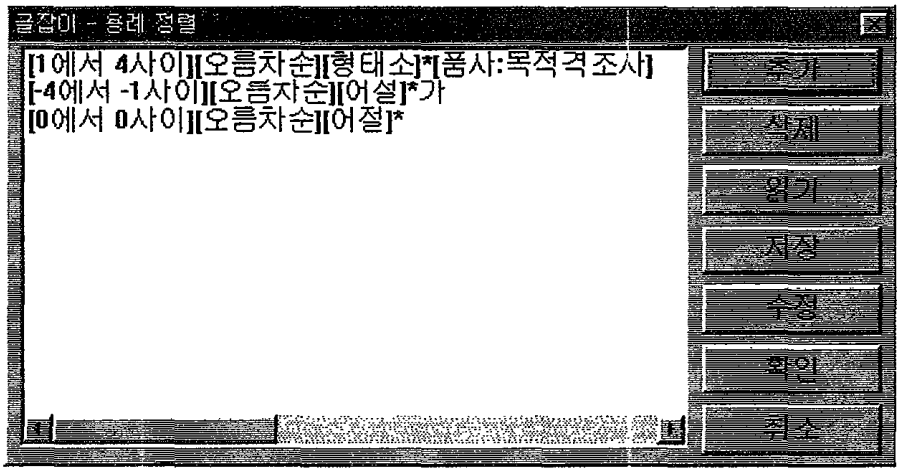
왼쪽 문맥	패턴	오른쪽 문맥
되는 일이 아니이지 말만 낳는	[경우도]	허다했다.
면 아들을 낳을 수 있다고 믿는	[기자치성도]	있었다.
	[제주도]	용담리에는 밤에 몰래 와서 기원하거나, 심
는 자기의 성기를 접촉시켰던	[사례도]	있었다고 전해지고 있다.
祈子水)라하여 그 물로 하체를	[씻기도]	하였다.
아야 한다는 강박 관념을 다소	[줄기도]	하였다.
나 베개나 요에 넣어 두고 잠을	[자기도]	했고, 아들 낳은 집에 걸어 두었던 금줄을 일
니거나 고추 달린 금줄 훔쳐다	[보관하기도]	했다.
옷)등속을 얻거나 훔쳐서 몸에	[지니기도]	했다.
위 자고 타인이 꾼 아들 태몽을	[사기도]	하였다.
을 복용하여 체질을 개선하는	[경우도]	드물지 않다.
으며 신혼 여행지로 각광받는	[제주도]	곳곳의 하루방의 코가 다 닳아 없어질 정도로
애쓰셔야겠네요"라는 위로도	[충고도]	아닌 인사말을 전한다.
서도 장남과는 기대치가 달라	[지원도]	적은 반면, 자신의 능력이나 희망과는 동떨어
루 부마 말딸은 가문을 대표할	[수도,]	'대'를 이을 수도 없는 '여자'이고 결국은 출
문을 대표할 수도, '대'를 이을	[수도]	없는 '여자'이고 결국은 출가외인이다.
되고 의리심이 강해 결단력이	[부족해지기도]	한다.

<색인 용례 탐색 결과>

○ 용례 정렬

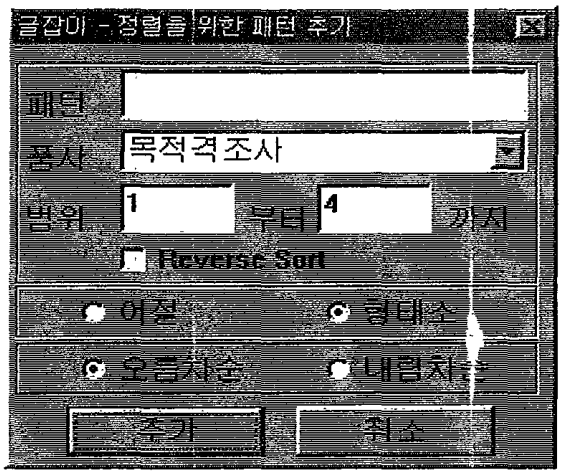
용례 정렬은 용례 탐색을 통해 찾아진 용례에서 사용자가 원하는 정보를 얻어내기 위해 가나다순으로 정렬하는 기능을 한다. 다음의 순서로 질의를 작성한다.

1) "정렬과 연산" 메뉴 아래의 [용례 정렬]을 선택하면 그림과 같은 상자가 나타난다.



<용례 정렬 입력 상자>

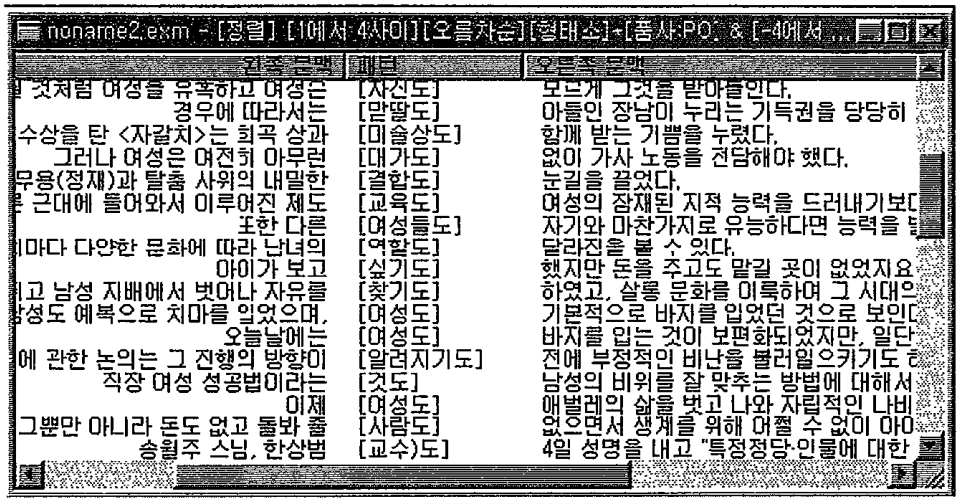
2) 정렬 질의를 입력하기 위해 '추가' 버튼을 클릭하면, 다음 그림과 같은 '단어 추가 상자'가 화면에 나타난다. 이 용례 정렬 단어 추가 상자에 그림과 같이 입력하고 '추가'를 누르면, 해당하는 질의가 '용례 정렬 입력 상자'에 추가된다.



<용례 정렬 단어 추가 상자>

3) 질의를 생성한 후 '확인'을 클릭하면 아래 그림과 같은 용례 정렬 결과를 얻을 수 있다. 이렇게 하면 '색인 용례 탐색'에서의 결과에 대해서, 중심어 뒤 1에서 4어절 사이에 목적격조사가 있는 어절, 중심어 앞 -1에서 -4어절 사이에 "~가"가 있는 어절, 그리고 중심어에 대해서 정렬을 한다.

정렬의 우선 순위는 목적격조사가 있는 어절, "~가", 그리고 중심어의 순으로 된다. 목적격 조사를 가지는 단어가 동일한 어절인 경우에, "~가"를 가진 어절에 대해서 정렬하고, "~가"를 가진 어절 역시 동일한 어절인 경우에는 중심어('[]'로 표시된 어절)에 대해서 정렬을 하게 된다.



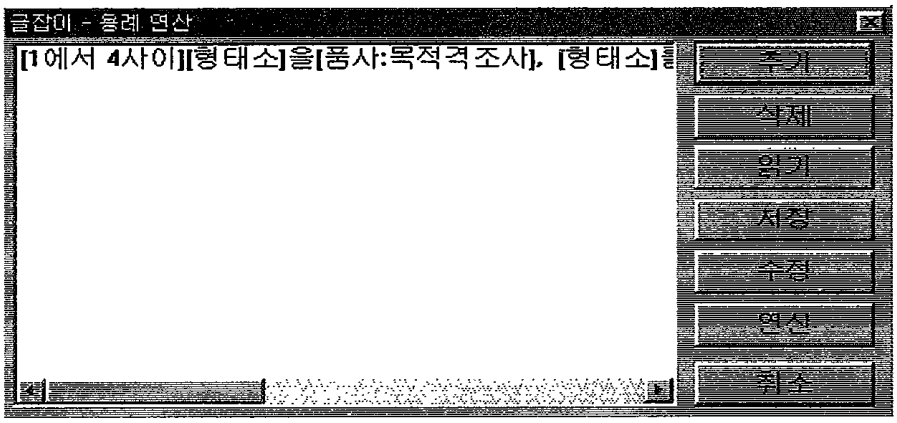
<용례 정렬 결과>

○ 용례 연산

용례 연산은 용례 탐색을 통해 찾아진 용례에서, 사용자가 원하는 정보만을 얻어내기 위한 작업이다. 다음의 순서로 질의를 작성하도록 한다.

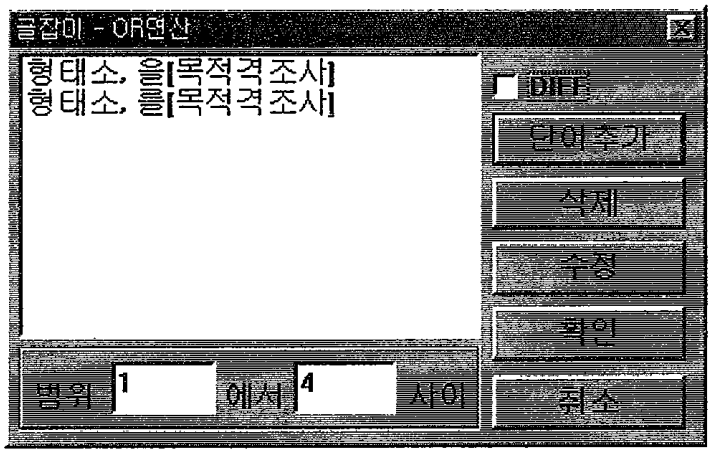
1) "정렬과 연산" 메뉴 아래의 [용례 연산]을 선택하면 그림과 같은 상자가 나타난다.[27]

27) 아래에 설명할 과정을 거쳐 입력한, 그림의 질의어는 "이미 추출된 용례 검색의 결과



<용례 연산 입력 상자>

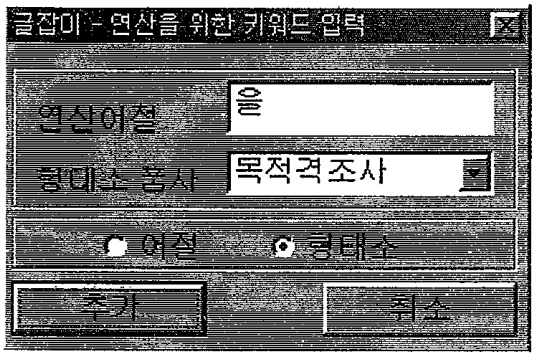
2) 앞의 그림의 형태로 내용을 채우기 위해서 오른쪽의 "추가" 버튼을 누르면 아래의 용례 연산 상자가 나타난다.



<용례 연산 OR 연산 상자>

여기에서 "단어 추가" 버튼을 누르면 단어 추가 상자가 나타난다.

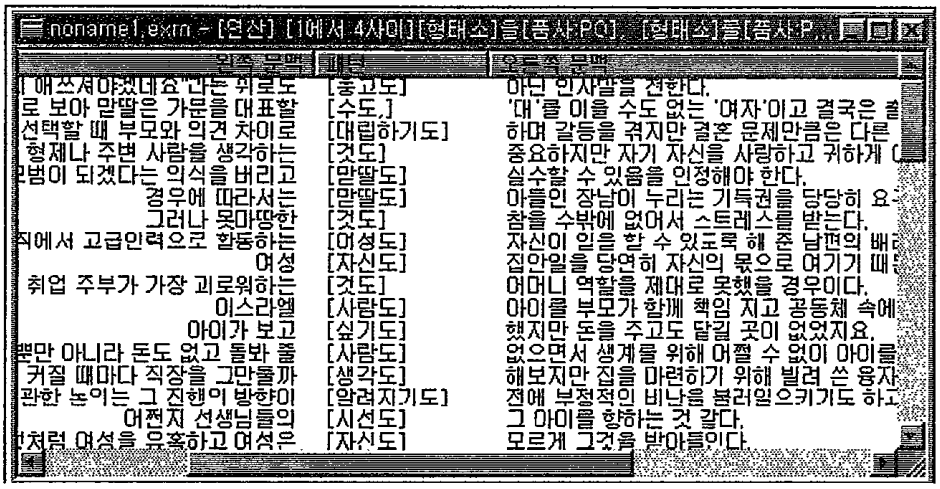
에 대해서, 중심어 뒤 1어절에서 4어절 사이에 목적격조사 '을'이나 '를'이 있는 어절만을 가려내는" 연산을 의미한다.



<단어 추가 상자>

3) 단어 추가 상자에 연산 조건을 입력하고 "추가" 버튼을 누르면, 용례 연산 입력 상자에 한 줄이 추가가 되게 된다. 이런 식으로 "단어 추가" 버튼을 사용해서 연산 조건을 입력한다. 이런 식으로 여러 줄을 추가할 수 있다.
4) 연산 조건이 모두 결정된 후에, "연산" 버튼을 누르면 연산을 하게 되면, 각 줄의 연산 조건이 모두 성립되는 용례만이 출력된다.

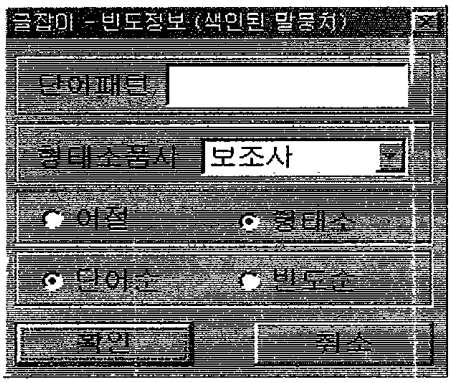
다음 그림은 위의 용례 연산에 대한 결과이다.



<용례 연산 결과>

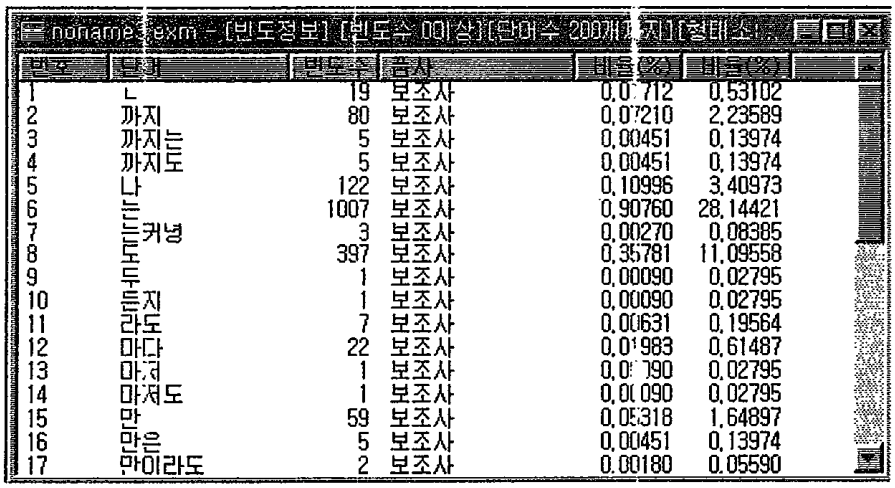
○ 빈도 정보 탐색

용례의 빈도 정보를 추출하려면 '예제-말뭉치'에 오른쪽 마우스를 클릭하고 [빈도정보]를 선택한다. 그러면 그림과 같은 입력 상자가 화면에 나타난다.



<빈도 정보 입력 상자>

이 상자에서 그림과 같이 입력하고 '확인'을 누르면 형태소 중에서 보조사에 대한 빈도 정보가 아래의 그림과 같이 화면에 출력된다.



번호	단어	빈도수	품사	비율(%)	비율(%)
1	ㄴ	19	보조사	0.0[illegible]712	0.53102
2	까지	80	보조사	0.07210	2.23589
3	까지는	5	보조사	0.00451	0.13974
4	까지도	5	보조사	0.00451	0.13974
5	나	122	보조사	0.10996	3.40973
6	는	1007	보조사	0.90760	28.14421
7	는커녕	3	보조사	0.00270	0.08385
8	도	397	보조사	0.35781	11.09558
9	두	1	보조사	0.00090	0.02795
10	들지	1	보조사	0.00090	0.02795
11	라도	7	보조사	0.00631	0.19564
12	마다	22	보조사	0.01983	0.61487
13	마저	1	보조사	0.0[illegible]090	0.02795
14	마저도	1	보조사	0.0[illegible]090	0.02795
15	만	59	보조사	0.05318	1.64897
16	만은	5	보조사	0.00451	0.13974
17	만이라도	2	보조사	0.00180	0.05590

<빈도 정보 결과>

2.3.3 '글잡이 1.0' 부속 도스용 용례 추출기

【소개】 '글잡이 1.0'은 한국어 용례 추출기로서, 품사 부착 말뭉치로부터 용례 및 빈도 정보를 획득하는 작업을 지원하는 도구이다. 문화관광부 지원으로 이루어진 '21세기 세종 계획'의 기초언어자료 분과에서 1998년 과제의 일부로서 개발한 용례 검색 프로그램이다.

【사양】 Ms-Dos

【개발자】 고려대학교 컴퓨터학과 자연어처리연구실(개발 책임자: 임해창 교수)

【개발시기】 1998년

【관련 사이트】 http://www.sejong.or.kr의 자료실

【설치 방법】

1) CD-rom에 있는 setup.exe를 실행하면 창이 하나 나오고, 창에는 설치할 디렉토리를 입력할 수 있게끔 되어 있다. 이 곳에 설치할 디렉토리를 입력한 후, "설치" 버튼을 누르면 설치된다.
2) 설치 후에, 다른 설정의 변경 없이 바로 사용할 수 있다.
3) 설치가 끝나면, 설치한 디렉토리 밑에 "Dos용"과 "Windows용"이라는 디렉토리와, "manual.hwp"와 "liner.exe" 화일이 생기게 된다.
4). 실행 프로그램 : "Dos용"이라는 하위 디렉토리 밑의 각 실행 파일이다.
   dos용으로 쓸 수 있는 프로그램의 이름과 기능은 다음과 같다.

   hantags.exe : 문서 태깅
   pre.exe: 문서 전처리
   index.exe : 문서 색인
   ee.exe : 색인된 문서에서 어절용례 탐색
   me.exe: 색인된 문서에서 형태소용례 탐색
   fre.exe : 색인된 문서에서 빈도정보 추출
   eed.exe : 색인되지 않은 문서에서 어절용례 탐색
   med.exe : 색인되지 않은 문서에서 형태소용례 탐색
   fred.exe : 색인되지 않은 문서에서 빈도정보 추출
   select.exe : 탐색된 용례에 대한 연산 수행
   sort.exe : 탐색된 용례에 대한 정렬 수행

viewpos.exe : 품사태그와 이름에 관한 정보

○ 말뭉치 태깅(hantags.exe)

【소개】 날말뭉치에 품사를 부착해서 품사 부착 말뭉치를 만드는 프로그램
【사용 방법】 hantags -c <디스크립트 파일>

- c는 옵션으로, 본 시스템에서 사용되는 품사 부착 말뭉치를 만드는 데 반드시 사용해야 하는 옵션이다. <디스크립트 파일>은 hantags에서 사용되는 환경 설정 파일과 품사 부착하고자 하는 날말뭉치, 그리고 품사 부착된 말뭉치를 저장할 파일이름을 가진다. 형식은 다음과 같다.

```
hantag.ini
raw corpus file
save file
```

- 태깅에 사용되는 디스크립트 파일의 형식 : "hantag.ini"는 hantags.exe에서 사용하는 환경변수가 저장된 파일이다. "raw corpus file"은 품사를 부착할 날말뭉치를 가지고 있는 파일이며, "save file"은 품사 부착된 말뭉치를 저장할 파일이다. 반드시 한 줄에 하나의 파일 이름이 적혀 있어야 된다. 이는, 파일 이름에 공백 사용이 가능한 윈도우95이상의 파일 이름 형식을 지원하기 위해서이다.

사용예) hantags -c hantag.sct

```
hantag.ini
corpus.txt
corpus.tag
```

<hantag.sct>

○ 전처리(pre.exe)

【소개】 품사 부착 말뭉치에서 각 문장의 위치를 있도록 하기 위해서, 색인 전에

말뭉치에 위치 정보를 기록하는 프로그램이다. 그러므로, pre.exe는 반드시 품사 부착 이후에 실행되어야 된다.

【사용 방법】 pre <품사 부착 말뭉치>

'품사 부착 말뭉치'는 품사 부착이 된 문서가 저장되어 있는 파일을 의미한다.

사용예) pre donga.tag

"동아일보.tag"이라는 품사 부착된 말뭉치 파일에 대해서 전처리를 실행한다.

○ 용례 색인(index.exe)

【소개】 품사 부착과 전처리 과정을 거친 파일(말뭉치)에 대해서 용례를 색인하는 프로그램이다.

【사용 방법】 index <말뭉치 디렉토리> <용례 정보 디렉토리>

<말뭉치 디렉토리>는 말뭉치가 있는 디렉토리를 의미한다. 색인은 <말뭉치 디렉토리>안에 있는 모든 문서들에 대해서 색인을 하게 된다. 말뭉치 디렉토리 안에는 반드시 품사 부착과 전처리를 과정을 거친 파일들만이 있어야 된다. 그렇지 않을 경우, 잘못된 정보가 색인될 수 있다. <용례 정보 디렉토리>는 색인 정보가 있는 디렉토리를 의미한다. <용례 정보 디렉토리>가 없는 경우, 디렉토리를 만든다.

사용예) index donga_c donga_i

"donga_c"라는 디렉토리 밑에 있는 모든 파일들에 대해서 색인을 한다. 색인시에 사용되는 모든 용례 정보 파일은 "donga_i"라는 디렉토리에 저장되게 된다.

○ 색인 용례 탐색

색인이 된 문서에 대해서, <용례 정보 디렉토리>에 있는 색인정보를 사용해서 용례를 탐색하게 된다.

(1). 어절 색인 용례 탐색

【소개】 어절에 대한 용례를 탐색하게 된다.
【사용 방법】 ee <어절> <용례 정보 디렉토리> <저장할 파일>
【설명】 <어절>은 찾고자 하는 어절을 의미한다. 탐색 <어절>은 와일드문자(wild character)인 "*"를 지원한다. 가령, "어머니"라는 단어를 탐색하고자 하는 경우, <어절>에 "어머니"만을 입력하는 경우, "어머니"가 하나의 어절로 분리되어 있는 문장을 찾는 반면에, "어머니*"를 입력하면, "어머니"뿐 아니라, "어머니는", "어머니가", "어머니를" 등의 어절을 가진 문장까지 탐색하게 된다. <용례 정보 디렉토리>는 색인 정보가 있는 디렉토리를 의미한다. <저장할 파일>은 탐색된 용례들을 저장할 파일을 의미한다.

사용예) ee "어머니*" donga_i 어머니.exm

어절 "어머니*"를 용례 정보 디렉토리 "donga_i"를 참조해서 탐색한 후에, 탐색된 결과를 "어머니.exm"파일에 저장한다.

효철은 [어머니가] 다녀간 건 아닐까 생각했다.
있다 다 저녁 때 나가는 걸 [어머니가] 본다면 필시 못 나가게 하거나 나가게
아버지와 [어머니가] 진우의 과거·현재·미래를 캐물을 것이 분
근데 [어머니가] 편찮으시다면서 왜 병간호는 안 하고
르고 누나가 집 나간 얘기며 [어머니가] 아파 누워 있단 소리까지 아무렇지 않게
시골에 계신 [어머니는] 몸이 편찮으셔서 몇년째 앓고 계십니다."

<ee "어머니*" donga_i 어머니.exm>

(2). 형태소 색인 용례 탐색(me.exe)

【소개】 형태소에 대한 용례를 탐색하게 된다.
【사용 방법】 me <형태소> <품사> <용례 정보 디렉토리> <저장할 파일>
【설명】 <형태소>는 찾고자 하는 형태소를 의미한다. <품사>는 찾고자 하는 형태소의 품사를 의미한다. <용례 정보 디렉토리>는 색인 정보가 있는 디렉토리를 의미한다. 형태소 용례 탐색에서는 "*"를 지원하지 않는다. 이는 형태소의 수가 어절의 수보다 적고, 일반적으로 형태소의 길이가 길지 않기 때문이다.

사용예1) me "가" "동사" donga_i 가_동사.exm

품사가 "동사"인 형태소 "가"를 용례 정보 디렉토리 "donga_i"를 참조해서 탐색한 후에, 탐색된 결과를 "가_동사.exm" 파일에 저장한다.

깊었지만 효철은	[갈] 곳이 없었다.
는 신경질적으로 발딱 일어나 부엌으로	[갔다.]
	[가면] 간다고 말을 하고 가야지."
땜에 챙피해서 어디	[가서] 등생 있단 소리도 못해.
해 광릉 수목원으로 진우와 함께 산책을	[가기로] 한 날이었다.
점점 더 그런 추세로	[가고] 있기 때문이다.

<me "가" "동사" donga_i 가_동사.exm>

사용예2) me "가" * donga_i 가_.exm

형태소 "가"에 대해서 품사를 구별하지 않고 모든 용례를 찾는다. 즉, 품사가 "동사", "주격조사"등의 모든 품사에 대해서 형태소 "가"의 용례를 찾게 된다.

사용예3) me * "동사" donga_i _동사.exm

품사가 "동사"인 모든 형태소의 용례를 찾는다.

○ 직접 패턴 탐색

크기가 작은 말뭉치에 대해서 색인의 과정을 거치지 않고, 바로 어절, 형태소를 검색하고자 개발된 프로그램이다. 품사 부착과 전처리의 과정을 거친 말뭉치에 대해서 직접 패턴 탐색이 가능하다.

(1). 어절 직접 패턴 탐색

【소개】 어절에 대한 용례를 탐색하게 된다.
【사용 방법】 eed <어절> <태깅된 문서 파일> <저장할 파일>
【설명】 <어절>은 찾고자 하는 어절을 의미한다. 탐색 <어절>은 단일 어절로서 윈도우즈에서 지원했던 모든 기호를 사용할 수 있다. (단, 위치연산 @ 기호는 제외). <태깅된 문서 파일>은 탐색을 수행할 파일을 의미한다. <저장할 파일>은 탐색된 용례들을 저장할 파일을 의미한다.

사용예) eed "+에((도)|(만))" a.tag out.exm

어절 "+에((도)|(만))"를 a.tag에서 탐색한 후에, 탐색된 결과를 "어머니.exm" 파일에 저장한다.

(2). 형태소 직접 패턴 탐색

【소개】 형태소에 대한 용례를 탐색하게 된다.
【사용 방법】 med <형태소> <품사> <태깅된 문서 파일> <저장할 파일>
【설명】 <형태소>는 찾고자 하는 형태소를 의미한다. <품사>는 찾고자 하는 형태소의 품사를 의미한다. <태깅된 문서 파일>은 탐색을 수행할 파일을 의미한다.

사용예1) med "*" "동사" a.tag 동사.exm

품사가 "동사"인 모든 형태소를 a.tag에서 탐색한 후에, 그 결과를 "out.exm"

파일에 저장한다.

사용예2) med "도" * a.tag 도.exm

형태소 "도"에 대해서 품사를 구별하지 않고 모든 용례를 찾는다. 즉, 품사가 "보조사", "보통명사"등의 모든 품사에 대해서 형태소 "도"의 용례를 찾게 된다. (탐색 결과는 색인 용례 탐색과 거의 유사하므로 생략한다.)

○ 용례 연산(select.exe)

【소개】 탐색된 용례 중에서, 사용자에게 진정으로 필요한 것을 추려내는 기능을 가진다.

【사용 방법】 select <디스크립트 파일> <탐색된 용례 파일> <연산된 결과 저장 파일>

【설명】 <디스크립트 파일>은 용례 연산을 위해 사용될 정보들을 저장하고 있는 파일이다. 디스크립트 파일의 형태는 다음과 같다.[28)]

28) 디스크립트 파일에서 "morpheme : xxx xxx xxx"에서 "xxx"는 형태소를 의미한다. 형태소 간에는 OR연산이 수행된다. 형태소에 대해서 "*"를 사용할 수 있다. "category : yyy yyyy yyyy"에서 "yyy"는 형태소의 품사를 의미한다. 위의 형태소의 개수가 3개인 경우, 품사의 개수 또한 3개이어야 된다. "*"인 형태소의 품사가 "*"인 경우, 이 형태소에 대한 연산은 무시된다. "position : nn ~ nn"에서 nn은 중심어에서의 상대적인 위치를 의미한다. "difference : on/off"에서 "on"인 경우, 형태소나 어절이 없는 경우 연산조건을 만족하는 것으로 간주되며, "off"인 경우에는 있는 경우 연산조건을 만족하는 것으로 한다.

```
M orpheme e : xxx xxx xxx
  category : yyy yyyy yyyy
  position : nn ~ nn
  difference : on/off
end
eojeol : xxx xxxx xxx
  position : nn ~ nn
  difference : on/off
end
```

<용례 연산을 위한 디스크립트 파일의 형식>

<탐색된 용례 파일>은 탐색 후에 저장된 파일을 의미한다. <연산된 결과 저장 파일>은 연산 결과를 저장할 파일을 의미한다.

사용예) ee "*을" donga_i 을.exm
select select.sct 을.exm 을.sel

먼저, "*을"에 대해 어절 탐색의 결과로 생긴 "을.exm"을 대상으로 용례 연산을 한다. 디스크립트 파일은 select.sct이다. 저장할 파일은 "을.sel".

```
eojeol : 한 국 *
position : 1 ~ 4
difference : off
end
morpheme : 가
position : -4 ~ -1
difference : on
category : 주 격 조 사
end
```

<디스크립트 파일 select.sct>

위의 select.sct는, 어절 "한국*" (eojeol : 한국*)이 중심어 뒤 1~4(position : 1~4)의 위치에 있는 문장이면서, 품사가 주격조사(category : 주격조사)인 형태소 "가" (morpheme : 가)가 중심어 앞 1~4(position -4 ~ -1)의 위치에 없는 문장만을 가려내도록 하는 디스크립트이다.

, 누가회 < CMF >, 라브리, 목산교회, 새생명사랑회,	[생명을] 아끼는 모임, 예수원, 한국기독
위해 기독교는 사랑을, 불교는	[자비심을] 가르쳤으며, 한국의 천도교는
같은	[종후군들을] 나열해서 보더라도 한국정치

<select select.sct 을.exm 을.sel>

○ 용례 정렬(sort.exe)

【소개】 탐색된 용례를 정렬하는 명령어이다.

【사용 방법】 sort <디스크립트 파일> <탐색된 용례 파일> <정렬된 결과 저장 파일>

【설명】 <디스크립트 파일>은 용례 정렬을 위해 사용될 정보를 저장하고 있는 파일이다. 디스크립트 파일의 형태는 다음과 같다.[29] <탐색된 용례 파일>은 탐색 후에 저장된 파일을 의미한다. <정렬된 결과 저장 파일>은 정렬후의 결과를 저장할 파일을 의미한다.

29) <디스크립트 파일>에서 "morpheme : xxxx"에서 xxxx는 형태소를 의미한다. "category : yyyy"에서 yyyy는 형태소의 품사를 의미한다. "position : nn ~ nn"에서 nn은 중심어에 대한 상대적인 위치를 나타낸다. "order : [sSiI]"에서 's'는 오름차순정렬, 'S'는 내림차순정렬, 'i'는 뒤집힌 형태의 단어에 대한 오름차순정렬, 'I'는 뒤집힌 형태의 단어에 대한 내림차순정렬을 의미한다. 여기에서 뒤집힌 형태의 단어란, 단어 "우리는"에 대한 뒤집힌 형태의 단어는 "는리우"가 된다. 뒤집힌 형태의 정렬을 함으로써 어절에 나타나는 조사에 의한 정렬이 가능하게 된다.

디스크립트 파일에서 앞에 있는 조건일수록 우선 순위가 높다. 정렬에서 우선 순위가 높은 단어에 대해서 정렬을 하다가, 단어가 같게 되는 경우, 그 다음 높은 우선 순위의 단어로 정렬을 하게 된다. 여기에서는 "morpheme : xxxx"이 "eojeol : xxxx"보다 우선 순위가 높게 된다.

```
morpheme : xxxx
    category : yyyy
    position : nn ~ nn
    order    : [sSiI]
    end
eojeol : xxxx
    position : nn ~ nn
    order    : [sSiI]
    end
```

<용례 정렬을 위한 디스크립트 파일 형식>

사용예) me "을" "목적격조사" donga_i 을_목.exm
sort sort.sct 을_목.exm 을_목.srt

먼저, "을"에 대해 형태소 탐색의 결과로 생긴 "을_목.exm"을 대상으로 용례 연산을 한다. 디스크립트 파일은 sort.sct이다. 저장할 파일을 "을_목.srt".

```
eojeol : * 는
position : -3 ~ -1
order : s
end
morpheme : 가
category : 주 격 조 사
position : 1 ~ 3
order : S
end
```

<디스크립트 파일 sort.sct>

위의 sort.sct는 중심어 앞 1~3(position : -3 ~ -1)의 위치에 있는 어절 "*는"(eojeol : *는)과, 중심어 뒤 1~3(position : 1 ~ 3)의 위치에 있는 품사가 주격조사(category : 주격조사)인 형태소 "가" (morpheme : 가)에 대해서 정렬을 한다. 어절 "*는"이 형태소 "가"보다 우선 순위가 높다.

```
따라서 이들에게는 자율정신의 [중요성을] 배울 필요가 있다.
 바가 있지만 이제는 지방에서 [서울을] 거치지 않고도 세계가 보이도록 지방
과 능력을 키울 수 있는 재질과 [적성을] 발전시키도록 도와줄 필요가 있다는
로운 매니지먼트}는 굳어 있는 [조직을] 창의가 넘치는 새로운 조직으로 탈
기가 살고 있는 지방의 역사와 [현황을] 모르는 경우가 많았으며, 교과과정
```

<sort sort.sct 을_목.exm 을_목.srt>

○ 빈도 정보 추출

어절과 형태소의 빈도를 추출한다.

(1). 직접 빈도 정보 추출(fred.exe)

【소개】 문서를 처음부터 읽으면서 어절과 형태소의 빈도 정보를 추출한다. 문서를 처음부터 끝까지 다 읽어야 하기 때문에 많은 시간을 필요로 하는 단점이 있다. 그러므로, 이 명령어는 색인되지 않은 문서나 문서의 크기가 작은 문서에 사용하는 것이 적절하다.

【사용 방법】 fred <어절 | 형태소> <빈도 정보 추출 파일> <빈도 정보를 저장할 파일>

【설명】 <어절 | 형태소>은 어절 빈도 정보와 형태소 빈도 정보를 구별하기 위해 사용된다. 어절 빈도 정보를 위해서는 0, 형태소 빈도 정보를 위해서는 1로 두면 된다. <빈도 정보 추출 파일>은 빈도 정보를 추출하고자 하는 파일을 의미한다. <빈도 정보를 저장할 파일>은 추출된 빈도 정보를 저장할 파일을 의미한다.

사용예) fred 0 donga_92.txt bindo.txt

"donga_92.txt"라는 파일에서 어절에 대한 빈도 정보를 추출한다.

64	보았다.	49	0.111078
65	보며	48	0.108811
66	들었다.	47	0.106545
67	더	47	0.106545
68	입을	47	0.106545
69	있던	47	0.106545
70	보였다.	46	0.104278

<fred 0 donga_92.txt bindo.txt>

(2). 색인 빈도 정보 추출

【소개】 색인된 문서들에 대해서 빈도 정보를 추출한다. 색인된 문서에 대해서는 어절과 형태소의 빈도 정보가 이미 사전에 등록된 상태이기 때문에, 사전의 탐색으로 빈도정보를 구할 수 있게 된다.

【사용 방법】 fre <어절 | 형태소> <패턴> <용례 정보 디렉토리> <빈도 정보를 저장할 파일>

【설명】 <어절 | 형태소>는 어절 빈도 정보와 형태소 빈도 정보를 구별하기 위해 사용된다. 어절 빈도 정보를 위해서는 0, 형태소 빈도 정보를 위해서는 1로 두면 된다. <패턴>은 어절에서는 어절의 형태, 가령 "*는"을 의미하며, 형태소의 경우에는 형태소 혹은 품사를 의미한다. 형태소만 주어진 경우, 형태소의 품사별 빈도수가 나타나며, 품사만 주어진 경우는 형태소별 빈도수가 나타난다. <용례 정보 디렉토리>는 색인 정보가 있는 디렉토리를 의미한다. <빈도 정보를 저장할 파일>은 추출된 빈도 정보를 저장할 파일을 의미한다.

사용예) fre 1 "주격조사" donga_i i_bindo.txt

"donga_i"라는 용례 정보 디렉토리에서 "가"라는 형태소의 빈도 정보를 추출한다.

가_PS	2718	1.256826%
까지가_PS	1	0.000462%
께서_PS	1	0.000462%
란_PS	18	0.008323%
로서안이_PS	1	0.000462%
만이_PS	44	0.020346%
부터가_PS	1	0.000462%

<fre 1 "주격조사" donga_i i_bindo.txt>

【참고】 직접 패턴 탐색에서의 질의어 작성 및 탐색 유형

원하는 용례를 찾아내기 위해 질의어를 작성하는 과정은 직접 패턴 탐색에서 특히 중요하다. 찾아내기 원하는 패턴에 가장 적합한 질의어를 만들기 위해서는 사용되는 여러 가지 기호를 숙지할 필요가 있다. 아래 표는 직접 패턴 탐색에서 사용되는 기호를 설명하고 있다.

기 호	설 명
*	0 또는 그 이상의 길이를 가진 문자열
+	1 또는 그 이상의 길이를 가진 문자열
?	1음절
[A]	A가 0 또는 1회 출현
((A)|(B))	A또는 B가 1회 출현
A B	어절 A와 B가 이어서 출현
A @ B	A와 B사이에 0개 이상의 어절이 존재
A @n B	A와 B사이에 n개 이하의 어절이 존재
A @m,n B	A와 B사이에 m개 이상 n개 이하의 어절이 존재

위의 기호들을 잘 조합하면 사용자가 원하는 패턴의 어절이나 문맥을 찾아낼 수 있다.

2.3.4 패턴매치를 이용한 용례 추출기 hgrep

【소개】 HGREP.EXE는 '호글' 2.0 이상으로 작성, 저장된 자료 파일(*.HWP:압축 파일 제외)을 대상으로 한 문자열 검색 프로그램이다.[30]
【개발자】 국립국어연구원(용역 개발자 : (주)셈틀로소프트)
【개발시기】 1998년 7월
【관련 사이트】 http://www.sejong.or.kr의 자료실
【사용법】 [31]

(1). OS상에서 다음과 같은 명령문을 타자하여 실행시킨다.

hgrep [옵션] [명령파일]

[옵션]
/c 검사 결과를 파일로 저장하지 않고 일치하는 개수만 표시한다.
/o 출력 결과 파일을 덮어쓸 것인지 물어보지 않고 무조건 저장한다.
/? 도움말을 보여 준다.

[명령파일]
hgrep의 명령어가 저장된 명령파일을 지정한다. 명령파일은 '호글' 2.0 이상으로 작성되어야 한다. 명령파일의 확장자를 생략하면 .CMD가 자동으로 붙는다. 명령파일 자체를 생략했을 때는 HGREP.CMD가 현재

30) 그러나 '호글'의 내부 2바이트 코드 체계가 워낙 방대하기 때문에, 자체적인 사용자 인터페이스를 갖지 못하고, '호글'을 이용해 미리 작성된 명령어 파일을 읽어 그 파일에 기술된 명령어를 해석하여 작업을 수행한다. 이와 같이 특정한 작업을 위한 명령어가 기술된 파일을 스크립트(script) 파일이라고 한다. 우리가 흔히 접할 수 있는 스크립트 파일로는 MS-DOS의 배치(batch) 파일이 있다.

31) 이후의 설명에서 나오는 다음과 같은 기호는 특별한 의미로 가지므로 그에 따라서 올바르게 사용해야 한다.
[항목] ☞ []로 묶인 항목은 생략할 수 있다.
<항목> ☞ <>로 묶인 항목은 생략할 수 없다.
"항목" ☞ ""로 묶인 항목은 꼭 따옴표로 묶어 입력해야 한다.
(항목) ☞ ()로 묶인 항목은 꼭 괄호로 묶어 입력해야 한다.

디렉토리에 있는지 찾아서 있으면 그 파일을 실행하고, 없으면 에러 메시지를 표시하면서 끝낸다.

(2). 명령파일 구성

명령파일에는 HGREP의 명령어를 기술한다. HGREP에는 다음과 같은 세 가지 명령어가 있다.

◆ 검색파일 <파일지정> [옵션]
hgrep으로 검색할 대상 파일을 지정한다.

◆ 출력형식 "형식" [옵션]
검색한 결과를 출력할 때 알아 보기 쉽도록 다양한 형식을 지정한다.

◆ 검색내용 "정규식" <출력파일> [옵션]
실제 지정한 내용을 검색하여 결과를 출력 파일에 저장하는 명령이다.

한 명령문은 한 단락에 기술해야 한다. 만약 여러 개의 명령어를 동일한 단락에 기술했거나, 한 명령어를 여러 단락에 걸쳐 기술하면 에러 메시지를 내면서 검색이 중지된다.

한 명령파일에 기술할 수 있는 명령어의 수에는 제한이 없다. 예를 들어 '검색파일' 명령어를 여러 개 주면 HGREP은 지정한 순서 대로 여러 파일을 검색한다. 또한 '검색내용' 명령을 여러 개 주면 HGREP을 여러 번 돌려서 결과를 만들 작업을 한 번 작업으로 처리할 수 있게 된다.

(3). '검색파일' 명령문 구성

형식 : 검색파일 <파일지정> (새검색파일 <파일지정>) [옵션]

<파일지정>
검색할 대상 파일을 지정한다. 검색할 파일은 '한글' 2.0 이상으로 작

성된 것이어야 한다. 파일 이름에는 '?', '*'와 같은 아무 문자가 사용될 수 있다. 한 명령파일에 검색파일 명령어가 여러 개 지정되었을 때는 지정한 순서 대로 지정한 파일에 대해 모두 검색이 수행된다. 그러나 "새검색파일" 명령어가 지정되었을 때는 현재까지 지정된 검색파일 명령이 취소되고, 새로운 검색파일의 지정을 시작한다.

[옵션]

/하위디렉토리 지정된 파일이 있는 위치와 그 하위디렉토리를 모두 검색한다.

(4). '출력형식' 명령문 구성

형식 : 출력형식 "형식" [옵션]

"형식"

"표제어:n,출전:n,예문:n"의 형식으로 지정한다. 각 항목의 순서는 마음대로 바꿀 수 있으며, 항목을 생략하면 그 항목은 아예 출력되지 않는다. 각 항목의 출력 폭은 n으로 지정한다. n으로 지정하는 값은 '호글'에서의 문자 단위 폭이다. 초기값은 "표제어:30,출전:30,예문:72"이다. 그 사이는 1칸씩 비기 때문에 실제로는 "표제어:31,출전:31,예문:72"라고 할 수 있다. 예문의 최대 폭은 300이다.

[옵션]

/어절강조 검색에 일치된 부분을 예문에서 '{}' 부호로 묶어 강조한다.

/어절제시 예문을 제시할 때 경계에 걸린 어절의 중간을 자르지 않고 어절 전체를 아예 제외한다.

/예문표제:n 예문을 제시할 때 검색에 일치된 내용을 표시할 위치를 지정한다. 초기값은 37이다.

/어절왼쪽 예문을 제시할 때 발견된 내용을 왼쪽 맞춤으로 정렬한다. 왼쪽 맞춤을 지정했을 때 '예문표제:n'으로 지정한 값은

왼쪽 끝을 나타낸다.

/어절오른 예문을 제시할 때 발견된 내용을 오른쪽 맞춤으로 정렬한다. 오른쪽 맞춤을 지정했을 때 '예문표제:n'으로 지정한 값은 오른쪽 끝을 나타낸다.

/단락구분 예문을 제시할 때 단락이 바뀌면 다음 단락의 내용은 보여주지 않는다. 즉 모든 단락 뒤에 '^'가 있는 것처럼 검색된다.

2.4 언어정보를 계량화하기 위한 처리 도구와 활용

2.4.1 국어정보화 프로그램(21세기 세종 계획)

국어정보화 프로그램

검색된 게시물 : 13 / 13 , 페이지 : 1 / 1

번호	◎	제 목	성 명	작성일	조회
13		말뭉치 분석 도구용 프로그램 추가(rtm.exe)	국립국어연구원	00/06/07	209
12		한자 DB 검색 프로그램	국립국어연구원	00/03/25	411
11		말뭉치 분석 도구용 프로그램	국립국어연구원	00/03/23	600
10		[국어기초자료구축]지능형 형태소 분석기 1.0	국립국어연구원	00/03/23	403
9		문자열, 사전 검색 프로그램 소스	국립국어연구원	99/10/22	1057
8		★문자열/사전 검색 프로그램 판올림	국립국어연구원	99/07/13	1793
7		문자열, 사전 검색 프로그램(hgrep, hdb)	국립국어연구원	99/05/14	1564
6		[한민족언어정보화]한글맞춤법검색시스템#4	국립국어연구원	99/05/13	885
5		[한민족언어정보화]한글맞춤법검색시스템#3	국립국어연구원	99/05/13	576
4		[한민족언어정보화]한글맞춤법검색시스템#2	국립국어연구원	99/05/13	639
3		[한민족언어정보화]한글맞춤법검색시스템#1	국립국어연구원	99/05/13	1234
2		[한민족언어정보화] 남북한언어비교사전	국립국어연구원	99/05/13	607
1		[국어기초자료구축] 한국어용례추출기 프로그램	국립국어연구원	99/05/13	966

[1]

2.4.2 21세기 세종계획 기초언어자료 처리 도구

【소개】 21세기 세종 계획의 기초언어자료 분과에서 1998년 과제의 일부로서 개발한 말뭉치 처리 도구 패키지임.[32)]

- ▲ 문서표준형 점검 프로그램
- ▲ 자소 · 음절 빈도 조사 프로그램
- ▲ 대규모 말뭉치 대상의 어절 빈도 조사 프로그램
- ▲ 한자 · 한자어 빈도 조사 프로그램
- ▲ <훈글> 내부코드 자료 순차배열 프로그램
- ▲ 교정목록에 의한 말뭉치 자동교정 프로그램
- ▲ 말뭉치 계량 및 분포 측정 프로그램

구체적으로는 다음과 같은 프로그램이 포함되어 있다.

H2B.EXE	HWP 파일을 <훈글>의 내부코드 파일(H2B)로 변환.[33)]
H2K.EXE	HWP 파일과 KSSM 코드 파일의 상호 변환.
WDcount.EXE	HWP, H2B, 텍스트 파일들의 어절 수 측정.
WDC_TEI.EXE	21세기 세종계획 표준형 문서의 어절 수 측정.
CWORD.EXE	어절, 한자, 한자어 빈도 조사.[34)]
Jamo.EXE	한글 음절 사용 빈도 및 자소별 전이빈도 조사.[35)]
Hsort_H.EXE	HWP, H2B 파일 자료의 순차배열 및 KSSM 코드 출력.

32) 아래의 프로그램들의 사용법 및 각종 설명은, 1998년도 21세기 세종 계획 기초자료구축분과(연구책임자:김홍규 교수)의 최종보고서에 따른다. 이 내용은, tools98.hwp라는 파일로 배포되고 있는 프로그램 패키지 속에 포함되어 있다.

33) H2B.EXE의 핵심이 되는 <훈글> 파일 처리 부분은 한글과컴퓨터社로부터 오브젝트 코드 형태의 라이브러리를 제공받아 사용했다.

34) 설계 및 프로그래밍: 김홍규. CWord.EXE는 김홍규 · 강범모 공동 제작.

35) 이 프로그램은 고려대 심리학과의 이만영 교수가 구현 요망사항을 제시하고, 김홍규 · 강범모 교수의 자문에 의해 고려대 민족문화연구소 자연언어 데이터베이스실 프로그래머인 윤석일 씨가 제작했다.

Correct.EXE　　교정 어절 목록에 의한 대량의 텍스트 자동 교정.

Chk_TEI.EXE　　TEI 문서 형태(헤더와 마크업)의 기본 규칙 오류 검사.

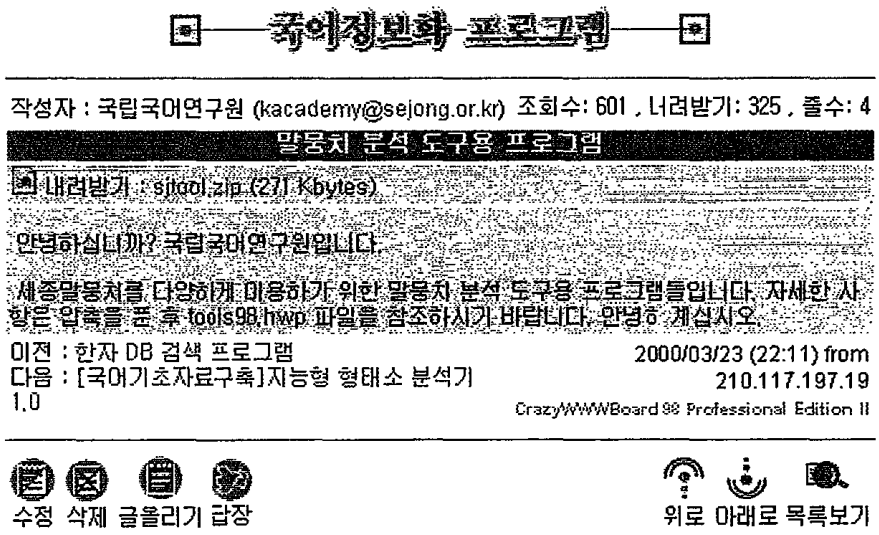
【사양】 사용 환경으로는 IBM-PC 호환기종의 DOS와 윈도95의 MS-DOS가 모두 가능하도록 했다. 다만 CWORD.EXE는 대규모 자료 처리를 위한 확장 메모리의 요구가 크기 때문에 주기억장치가 16MB 또는 그 이상이 되는 하드웨어 환경이 바람직하다.

【개발자】 고려대학교 민족문화연구원(연구 책임 : 김홍규 교수)

【개발시기】 1998년

【관련 사이트】 http://www.sejong.or.kr의 자료실

【사용법】



○ 프로그램의 설치

위의 도구 프로그램과 그 부속파일들은 전달 매체의 \SJTOOL이라는 디렉토리에 들어 있으며, 다음과 같은 지침에 따라 설치되어야 한다.

1) H2B.EXE는 C:\SEJONG 디렉토리에 설치해야 한다. 다른 디렉토리에 설치하더

라도 그 자체의 사용에는 지장이 없으나, 나머지 도구 프로그램들이 <훈글> 파일 변환을 위해 내부적으로 H2B.EXE를 호출할 때는 C:\SEJONG에서 찾기 때문이다.

2) 부속 파일로 함께 들어 있는 RTM.EXE는 CWORD.EXE와 같은 디렉토리, 또는 경로(path)가 지정된 디렉토리에 있어야 한다.[36]

3) 나머지 프로그램들은 임의의 디렉토리에 설치해도 무방하나, 향후의 여러 가지 가능성을 고려하여 \SJTOOL이라는 디렉토리를 사용할 것을 권장한다.

H2B.EXE

○ 기능 : HWP 파일을 <훈글>의 내부코드 파일(H2B)로 바꾼다.

○ 용법

(1). 다음과 같이, 도스 명령어행의 입력 방식으로 실행한다.

H2B <원본파일명> [목적파일] [/Q]

(2). 목적 파일 지정을 생략하면 자동적으로 '원본파일명.2B'로 목적파일이 설정된다.

(3). 원본 파일 지정에 아무개 문자들(*, ?)을 쓸 수 있다. 이 경우, 그 디렉토리의 해당 원본파일들에 대해 순차적으로 H2B 파일들을 만든다. 이 기능을 사용할 때는 목적파일을 따로 지정할 수 없으며, '원본파일명.2B'로 목적파일들이 설정된

36) RTM.EXE는 Borland Pascal로 작성된 프로그램의 확장 메모리 사용을 지원하는 보조 프로그램으로서, Inprise사에 저작권이 있다.

다.

(3). 해당 디렉토리에 이미 같은 이름의 목적파일이 있으면 겹쳐쓸 것인가를 묻는다.

(4). 이런 질문을 받지 않고 무조건 목적파일을 생성하려면 '/Q' 옵션을 사용한다.

(5). 프로그램 이름만을 입력하면 다음과 같은 도움말 화면이 나온다.

H2B, v3.1, 1997-1998, H.Kim, Korea Univ., Thanks to HNC.
Syntax: H2B <source.hwp> [/Q]
[/Q] = Overwrites Target without Question

H2K.EXE

○ 기능

HWP 파일을 KSSM 코드 파일로, KSSM 파일을 HWP 코드 파일로 바꾼다.

○ 용법

(1). 다음과 같이, 도스 명령어행의 입력 방식으로 실행한다.

H2K <옵션> <원본파일> [목적파일 꼬리] [/Q]

(2). 옵션의 종류는 다음과 같다.

/HK HWP 파일을 KSSM 파일로 변환

/KH KSSM 파일을 HWP 파일로 변환

(3). <원본파일> 지정에 아무개 문자들(*, ?)을 쓸 수 있다. 이 경우 그 디렉토리의 해당 원본파일들에 대해 순차적으로 변환 파일들을 만든다.

(4). [목적파일 꼬리] 파라미터로는 목적파일 이름꼬리만을 줄 수 있으며, '원본파일명.꼬리이름'으로 목적파일들이 설정된다.

(5). [목적파일 꼬리]를 지정하지 않을 경우, 다음과 같이 자동 설정된다.

HWP -> KSSM .TXT

KSSM -> HWP .HWP

(6). 해당 디렉토리에 이미 같은 이름의 목적파일이 있으면 겹쳐쓸 것인가를 묻는다. 이런 질문을 받지 않고 무조건 목적파일을 생성하려면 '/Q' 옵션을 사용한다.

(7). 프로그램 이름만을 입력하면 다음과 같은 도움말 화면이 나온다.

H2K: HWP <-> KSSM Text, H.Kim, Korea Univ, 1998, v1.2

Syntax : H2K </HK[n] or /KH[n]> sourceMask [targetExtension]
Default Target Extensions are 'TXT' for KSSM and 'HWP' for HWP'
Option + 'q' for Blind OverWrite Mode

WDcount.EXE

○ 기능

특정 디렉토리와 그 이하에 있는 파일들의 어절 수를 헤아린다.

○ 용법

(1). 이 프로그램은 도스 명령어행의 입력 방식으로 실행한다.

(2). 프로그램 이름만을 주면 다음과 같은 타이틀과 도움말이 나타난다.

WDcount, v1.6, for <21st Century Sejong Project>

H.Kim, Center for Electronic Texts, IKC, Korea Univ., 1998

Syntax: WDcount <Dir\File Mask> <TargetFile> [/option]

- Counts specified files' words in/under a DIR.
- Source files may be in HWP/H2B, KSSM, KSC5601 state.
  H2B files are assumed to have '.2B?/.H2B/.HWP' extension.

(3). 조건 입력 방식은 다음과 같다.

WDCOUNT <자료디렉토리\파일지정> [목적파일] [/옵션]

(4). 원본파일 지정에 아무개 문자들(*, ?)을 쓸 수 있다.

(5). <자료디렉토리\파일지정>은 다음의 사례를 참고한다.

(예 1) D:\TEXT*.*
D:\TEXT라는 디렉토리와 그 이하 디렉토리들에 있는 모든 파일의 어절 수를 센다.

(예 2) E:\TEXT\LAW*.HWP
E;\TEXT\LAW라는 디렉토리와 그 이하 디렉토리들에 있는 모든 *.HWP 파일의 어절 수를 센다.

(예 3) D:\TEXT*.*
D: 드라이브의 루트 디렉토리와 그 이하 디렉토리들에 있는 모든 TEXT*.* 파일의 어절 수를 센다.

(6). 대상 파일은 HWP, H2B, 텍스트 파일 모두가 된다. 텍스트 파일의 경우에는 KSC-5601-1992 코드(조합형 한글 및 한자) 코드와 KSC-5601-1987(완성형) 코드

를 자동적으로 식별한다. 텍스트 파일은 코드 종류를 식별하여 '!!KSSM' 또는 '!!5601'로 기록한다.

(7). 프로그램은 지정된 디렉토리와 그 아래에 있는 모든 서브디렉토리에 대해 파일 지정 형식에 해당하는 파일을 찾아서 어절 수를 헤아리고, 최종 보고 파일에 기록한다. 조사 보고는 일반 텍스트 파일로 기록된다.

(8). HWP 파일들에 대해서는 조사가 이루어지는 단계에서 자동으로 임시 H2B 파일을 생성했다가 프로그램 종료 전에 지워 준다.

(9). <훈글> 내부코드 파일은 파일 이름의 꼬리 부분이 '.HWP', '.H2B' 또는 '.2B?'(?는 임의의 글자)이어야 한다.

(10). TEI 방식과 국립국어연구원 방식의 파일 헤더 부분은 어절 수 계산에 포함하지 않고 건너뛴다. 프로그램은 이 두 종류의 헤더가 파일 서두에 있는지를 검사하여, 있을 경우 그것을 건너뛴 다음 부분만을 어절 수 계산 대상으로 삼는다. 세 번째 파라미터로서 '/S'를 주면 헤더가 있는지 검사하지 않고 파일의 서두에서부터 어절 수를 헤아린다.[37]

CWORD.EXE

○ 기능

지정된 위치(들)의 파일에 있는 어절들을 그 빈도와 함께 수집한다.

37) TEI 헤더의 존재 유무는 파일에 아무런 실질 문자가 나오기 전에 '<TEIHEADER>'라는 문자열이 있고, 그 아래의 200행 이내에 '</TEIHEADER>'라는 문자열이 있는 경우로 간주한다. '<TEIHEADER>'는 나왔는데 그 뒤의 200행 이내에 '</TEIHEADER>'가 없으면 헤더 에러로 간주한다. 국어연구원 자료 헤더의 존재 유무는 파일에 아무런 실질 문자가 나오기 전에 '/*'라는 문자열이 있고, 그 아래의 25행 이내에 '*/'라는 문자열이 있는 경우로 간주한다. '/*'는 나왔는데 그 뒤의 25행 이내에 '*/'가 없으면 헤더 에러로 간주한다. 헤더 에러가 있으면 경고음과 함께 그 사실을 화면에 알리고 작업은 계속한다. 이 경우 어절 수 계산의 정확성은 보증할 수 없다.

○ 특성 및 사용 환경[38)]

(1). 이 프로그램은 매우 빠르고 강력하다. 단적인 예로, 펜티엄 200 기종에서 3000만 어절 규모의 말모둠 전체를 대상으로 한 어절 통계를 5~6시간 정도에 완료한다.

(2). 이 프로그램은 30만개 내외의 어절을 넣을 수 있는 메모리 바구니를 사용하여, 대규모 자료 처리 도중 하드 디스크에 중간 결과를 기록하거나 통합하는 시간을 최소화함으로써 매우 빠른 속도로 수백만 개 이상의 어절들을 조사한다.

(3). 이 프로그램의 기능을 충분히 활용하기 위해서는 RAM이 16메가 바이트 또는 그 이상이어야 한다.

○ 용법

(1). 이 프로그램은 도스 명령어행의 입력 방식으로 실행한다.

(2). 조건 입력 방식은 다음과 같다.

CWORD <원본파일 또는 @목록파일> [목적파일] [/옵션]

(3). 원본파일 지정에 아무개 문자들(*, ?)을 쓸 수 있다.

(4). 원본파일은 어느 드라이브, 디렉토리에 있더라도 경로(path)가 붙어 있으면 처리된다.

(4). 여러 디렉토리에 있는 원본파일들을 선별적으로 지정해서 어절 조사를 하려면 목록파일에 그 사항을 기록하여 '@목록파일' 형식으로 표시할 수 있다. 목록파일의 형식을 참고로 예시하면 다음과 같다.

목록파일의 성격이나 용도를 기술

38) - 이 프로그램에 구현된 기능의 개발 분담은 다음과 같다.
△ 트리 소트 엔진 : 강범모
△ 확장 메모리 사용, <훈글> 자료 직접 처리, 각종 선택기능 구현 : 김홍규

\text\novel*.hwp
\corpus\nature\a*.hwp
\corpus\nature\m*.*
E:\temp\diction\ch*.txt
F:\last\paper*.h*

(5). 대상 파일은 HWP, H2B, 텍스트 파일 모두가 된다. 텍스트 파일의 경우에는 KSC-5601-1992 코드(조합형 한글 및 한자) 코드와 KSC-5601-1987(완성형) 코드를 자동적으로 식별한다.[39)]

(6). HWP 파일들에 대해서는 조사가 이루어지는 단계에서 자동으로 임시 H2B 파일을 생성했다가 프로그램 종료 전에 지워 준다.

(7). <훈글> 내부코드 파일은 파일 이름의 꼬리 부분이 '.HWP', '.H2B' 또는 '.2B?'(?는 임의의 글자)이어야 한다. 목적파일 이름을 주지 않으면 'CWORD.OUT'로 지정된다. 같은 이름의 파일이 있었다면, 묻지 않고 겹쳐쓰기가 된다.

(8). 조사하는 어절의 크기는 30바이트로 제한된다. 30바이트가 넘는 크기의 어절은 별도로 참조할 수 있도록 보조 파일에 기록된다. 긴 어절을 기록하는 보조 파일 이름은 '목적파일명.LNG'가 된다.

(9). 조사 대상 파일에 TEI 헤더가 있을 경우 그 부분은 어절 조사 대상에서 제외한다.[40)]

(10). 조사 보고는 KSSM 코드의 일반 텍스트 파일로 기록된다.

---

39) ○ 유의 사항

(1). 어절 조사 결과를 기록할 때 KSSM 코드로 변환될 수 없는 HWP 코드는 다음과 같은 형식으로 표시된다.

仮 -> &H-5318;　　싹 -> &H-E462;　　≡ -> &H-250D;

이러한 결과는 H2K.EXE에 의해 <훈글> 내부코드 파일로 완벽하게 복원될 수 있다.

(2). 검사하는 파일 중 HWP 혹은 H2B 형식이 아닌 파일이 있으면 정상적인 검사가 이루어지지 않을 수 있다.

(3). H2B에 의한 <훈글> 파일 변환 중 파일에 이상이 있으면 H2B 에러에 의해 전체 검사가 중단된다.

40) 국어연구원 방식의 헤더는 향후의 표준화에서 세종계획의 TEI 방식 헤더로 바뀔 터이므로 고려하지 않았다.

○ 옵션의 활용

(1). 프로그램 이름만을 주면 다음과 같은 타이틀과 도움말 화면이 나타난다.

```
CollectWORD, v2.4A, for <21st Century Sejong Project>
-----------------------------------------------
H.Kim & B.Kang, Center for Electronic Texts, IKC, Korea Univ., 1998

Syntax: CWord <SourceMask or @ListFile> [TargetFile] [/options]
Default Target = CWORD.OUT
* Remove Default options
    S: Strip TEI Marks,          T: Trim Formal Chars
    N: Head Numbers to DGX,          H: Hangul/Hanja Words Only
    E: Skip '<entry>' Lines
* Hanja String Collection = /J
    U: Unify KS Double Hanja
- Source files may be in HWP/H2B, KSSM, KSC5601 state.
  H2B files are assumed to have '.2B?/.H2B/.HWP' extension.
```

(2). 첫번째 그룹의 다섯 가지 옵션들은 모두 기본값으로 설정되어 있다. 즉, 별도로 해제하지 않으면 이 기능이 모두 적용된다. 그 내용은 다음과 같다.

S: Strip TEI Marks

텍스트에서 ' , , <head>, </head> ...' 등의 TEI 표지를 제거한 뒤 어절을 조사한다.

T: Trim Formal Chars

어절의 앞뒤에 다음과 같은 형식 문자가 있을 경우 이들을 제거하고

어절을 조사한다.
반각 문자: ' ! " , . / : ; < = > ? \ ' |
전각 문자: 〔〕 ◇ 《》 【】

N: Head Numbers to DGX
한글 또는 '%' 앞에 숫자가 있을 경우 다음과 같은 형식으로 바꾸고 어절을 조사한다.
예) 1949년 -> dgx4년, 35시간이나 -> dgx2시간이나, 1.5%를 -> dgx3%를

H: Hangul/Hanja Words Only
한글, 한자, 전각문자가 하나라도 포함된 어절만을 조사 대상으로 한다. 즉, 영문, 숫자, 아스키 부호만으로 된 어절은 조사 대상에서 제외한다.

- 이 옵션들 중 사용하지 않고자 하는 것은 '/' 표 뒤에 해당 알파벳을 붙여 써서 표시한다. 둘 이상의 옵션도 하나의 '/' 뒤에 연속으로 쓴다. 순서나, 대소문자는 구별할 필요가 없다.
  예) /ST, /st, /St, /nhs
- 이 옵션들을 모두 다 제거하겠다고 표시하면, 어절 조사는 오직 띄어쓰기에 의해서만 구분되는 원본의 어절 형태 그대로를 대상으로 이루어진다.

(3). 두 번째 그룹의 옵션들은 사용자의 선택에 의해서만 활성화된다. 즉, 별도로 지정하지 않으면 이 기능들은 적용되지 않는다. 그 내용은 다음과 같다.

J: Hanja String Collection
한자의 연속으로 이루어진 문자열들만을 조사한다. 한자어와 한자 연구를 위한 기능으로 활용될 수 있다.

U: Unify KS Double Hanja

한자의 자형은 같은데 음의 차이 때문에 둘 이상의 코드가 부여된 표준코드 체계의 한자들(예: '力=역'과 '力=력', '龜=구'와 '龜=귀')을 本音 내지 우세한 음 쪽의 코드로 통일한 뒤 문자열을 조사한다. 이 옵션은 위의 'J' 옵션이 선택되었을 때만 유효하다. 즉 '/J'라는 단독 옵션과 '/JU'라는 결합 옵션은 가능하나, '/U' 또는 '/UJ'라는 옵션은 성립하지 않는다.

Jamo.EXE

○ 기능

이 프로그램은 한글 사용빈도를 조사하기 위한 것으로서, 단순한 음절빈도 외에 다양한 유형의 전이빈도와 자소 빈도를 산출한다.

○ 용법

(1). 이 프로그램은 도스 명령어행의 입력 방식으로 실행한다. 그 형식은 다음과 같다.

JAMO <자료 파일 이름>

(2). 자료 파일 이름에는 아무개 문자(wild card characters)들을 써서, 여러 개의 파일을 한꺼번에 분석할 수 있다.

(3). 자료 파일은 KSC5601-1992 조합형 코드(일명 KSSM 코드)로 된 텍스트 파일이어야 한다.

(4). 프로그램은 현재의 디렉토리에서 지정된 이름에 맞는 모든 파일들을 검사하여 다음과 같은 통계 결과 파일들을 생성한다.

EUMJUL.OUT	음절 빈도 통계
CHO.OUT	초성 빈도 통계
JUNG.OUT	중성 빈도 통계
ZONG.OUT	종성 빈도 통계
CHOJUNG.OUT	초-중성 전이빈도 통계
JUNGZONG.OUT	중-종성 전이빈도 통계
JUNGCHO.OUT	중-초성 전이빈도 통계
ZONGCHO.OUT	종성-초성 전이빈도 통계
JUNGENG.OUT	중성-기타문자 전이빈도 통계
ZONGENG.OUT	종성-영문자 전이빈도 통계
ENGCHO.OUT	영문자-초성 전이빈도 통계

Hsort_H.EXE

○ 기능

(1). 이 프로그램은 예전에 국어국문학 연구용으로 개발하여 전문 연구자들 사이에서 유용하게 쓰여 온 <한솔, HanSort>을 21세기 세종계획의 <훈글> 자료 처리용으로 개량한 것이다.

(2). 이 프로그램은 <훈글> 파일로 작성된 자료 목록 중에서 일정 순위의 어절들을 순차 판단 기준으로 삼아 재배열하고, 그 결과를 텍스트 파일로 기록한다.

○ 용법

(1). 이 프로그램은 메뉴와 선택화면에서 작업 조건들을 입력받는 방식으로 작동한다.

(2). 프로그램 이름을 주면 다음과 같은 초기 화면이 나타난다.

HanSortH [255 cols.], H.Kim, Korea University, 1992-1998, v2.3

Intelligent Sort for Hangul/Hanja Text (HWP/H2B -> KSSM Text

Source File = _ <Ctrl-A>: Help, <Ctrl-D>: Directory, <ESC>: Quit

(3). 원본 파일명과 목적 파일명을 묻는 단계 이후에 다음과 같은 조건 입력 화면이 나타난다. ◁표 이후의 입력 설명은 영문으로 되어 있으나, 여기에는 설명의 편의를 위해 우리말로 표시했다.

[1]	[2]	[3]	[4]	[5]	[6]	Sort Key Area
						◁ 구역 시작 어절 순번
						◁ 구역 끝 어절 순번
						◁ 작업 종류 : H, C, E
						◁ 배열 순서 (오름순=A, 내림순=D)
						◁ 한자음 판별에 두음법칙 적용
						◁ 영문 대.소문자 구별 무시
						◁ 문장부호를 제외한 순위
						◁ 띄어쓰기를 제외한 순위

(4). 위의 입력 조건 화면에서 '구역 시작 어절'과 '구역 끝 어절'이란 순차배열의 평가 대상이 되는 자료 영역의 시작과 끝을 가리킨다. 예컨대 위의 물음에 대해 각각 '3'과 '5'를 응답했다면 다음과 같이 [ ]로 둘러싸인 비교 영역을 선택한 것이 된다.

<예>

해도 저물녁 [그대 종종거름 훤돗] 다녀갈뿐

그밤 또 그대 [날과 샘과 셋이] 도른도른
무슨 그리 [향그런 이야기 날을] 세었다
샘은 애끈한 [젊은꿈 이제도 그지] 지녔으리
이밤 내 혼자 [나려가볼꺼나 나려가볼꺼나]

- 위의 자료를 순차배열한 결과는 다음과 같다.

해도 저물녁 [그대 종종거름 훤돗] 다녀갈뿐
이밤 내 혼자 [나려가볼꺼나 나려가볼꺼나]
그밤 또 그대 [날과 샘과 셋이] 도른도른
샘은 애끈한 [젊은꿈 이제도 그지] 지녔으리
무슨 그리 [향그런 이야기 날을] 세었다

(5). 순차배열 기준 구역은 6개까지 설정할 수 있으며, 각 구역마다 오름순 또는 내림순을 자유로이 선택하여 혼합할 수 있다. 순차배열 결과는 KSSM 코드의 텍스트 파일로 기록된다.

Correct.EXE

○ 기능

이 프로그램은 많은 파일들에 출현하는 여러 종류의 오류 어절들을 한꺼번에 교정하는 작업을 효율화하기 위한 것이다.

가령 수십, 수백 개의 파일에 여기 저기 들어 있는 100여 개의 오류 어절들을 수작업으로 고친다고 할 경우 막대한 시간과 인력이 소요된다. 뿐만 아니라 교정자의 부주의에 의해 고쳐지지 않은 오류 어절들이 남거나 새로운 오류가 생길 수도 있다. 이 프로그램은 면밀하게 교정된 목록에 의거하여 해당 오류

들을 한꺼번에 교정함으로써 그러한 부담을 덜어 준다.

<용법>

(1). 이 프로그램은 도스 명령어행의 입력 방식으로 실행한다. 그 형식은 다음과 같다.

Correct <교정 목록 파일 이름> <원본 파일 이름>

(2). 프로그램 이름만을 입력하면 다음과 같은 도움말 화면이 나온다.

```
Correct, v1.0, for <21st Century Sejong Project>
-----------------------------------------------
H.Kim, Center for Electronic Texts, IKC, Korea Univ., 1998

Syntax: Correct <Correction List> <Source File Mask>
        This program makes automatic correction to HWP/H2B files.
        Correction list have to be provided by user.
        Please refer manual file/document for further information.
```

(3). 교정 목록 파일과 원본 파일은 모두 <훈글> 파일이거나 <훈글>의 내부코드 파일이어야 한다.

(4). 원본 파일 이름에는 아무개 문자(wild card characters)들을 써서, 여러 개의 파일을 한꺼번에 교정할 수 있다. 교정 목록 파일은 당연히 사용자가 만들어야 한다. 교정용 목록은 아래의 예와 같이 '!'을 경계로 하여 그 앞에 교정할 원본형의 어절이, 뒤에 바른 형태가 오도록 작성한다. 원본형 어절은 그 안에 빈칸을 포함할 수 없으나, 이를 대체할 바른 형태는 여러 개의 어절로 이루어질 수 있다. 즉 프로그램은 '!' 뒤의 문자열 전체를 앞의 어절에 대체할 것으로 간주한다. 단, 맨 앞에 '!#'를 붙인 행은 교정용 어절 목록이 아닌 것으로 간주하므로,

목록에 관한 설명이나 참고 정보를 이 방식으로 기록해 둘 수 있다. 교정용 어절 목록의 행 수에는 제한이 없다. 그러나 기능상의 효율성을 위해 교정 대상 어절과 교정 목표형 문자열의 길이에는 상한선이 있다. 영문 · 부호 · 숫자를 1바이트로, 한글 · 한자 · 전각문자를 2바이트로 환산한 길이의 상한선은 다음과 같다.

△ 교정 대상 어절 = 50 바이트
△ 교정 목표 문자열 = 60 바이트

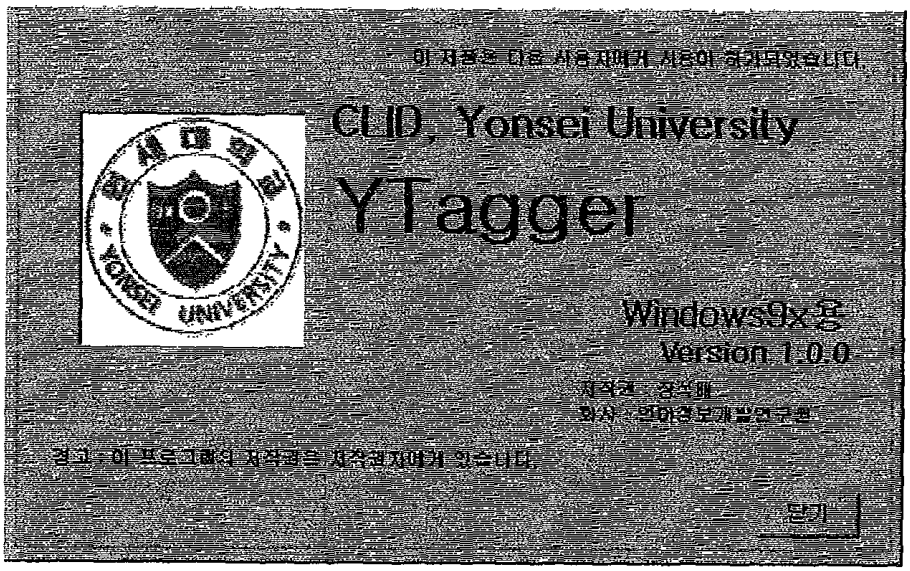
교정용 어절 목록의 본보기를 약간량 제시하면 다음과 같다.

<예>

```
!# 이것은 본보기용 교정 어절 목록임.
!# 작성: 1998년 11월 15일.
!# 작성자: 김 아무개.
마춤법 ! 맞춤법
마춤법만이 ! 맞춤법만이
大阮君 ! 大院君
大阮君이 ! 大院君이
서울시성북구 ! 서울시 성북구
부산시 ! 부산직할시
1998,10,23 ! 1998.10.23
김철수는 ! 김철수(회사원, 32세, 인천 출생)는
```

(6). 교정된 결과 파일은 원본 파일 이름에 '.CRR'이라는 꼬리가 붙은 형태로 새로 생성되어, 원본 파일과 동일한 디렉토리에 저장된다. 교정 결과 파일은 <훈글>의 내부코드 파일이므로, 이를 편집하려면 '2B' 방식으로 읽어들여야 한다.

2.4.3 말뭉치 주석 및 계량 도구 'Ytagger'



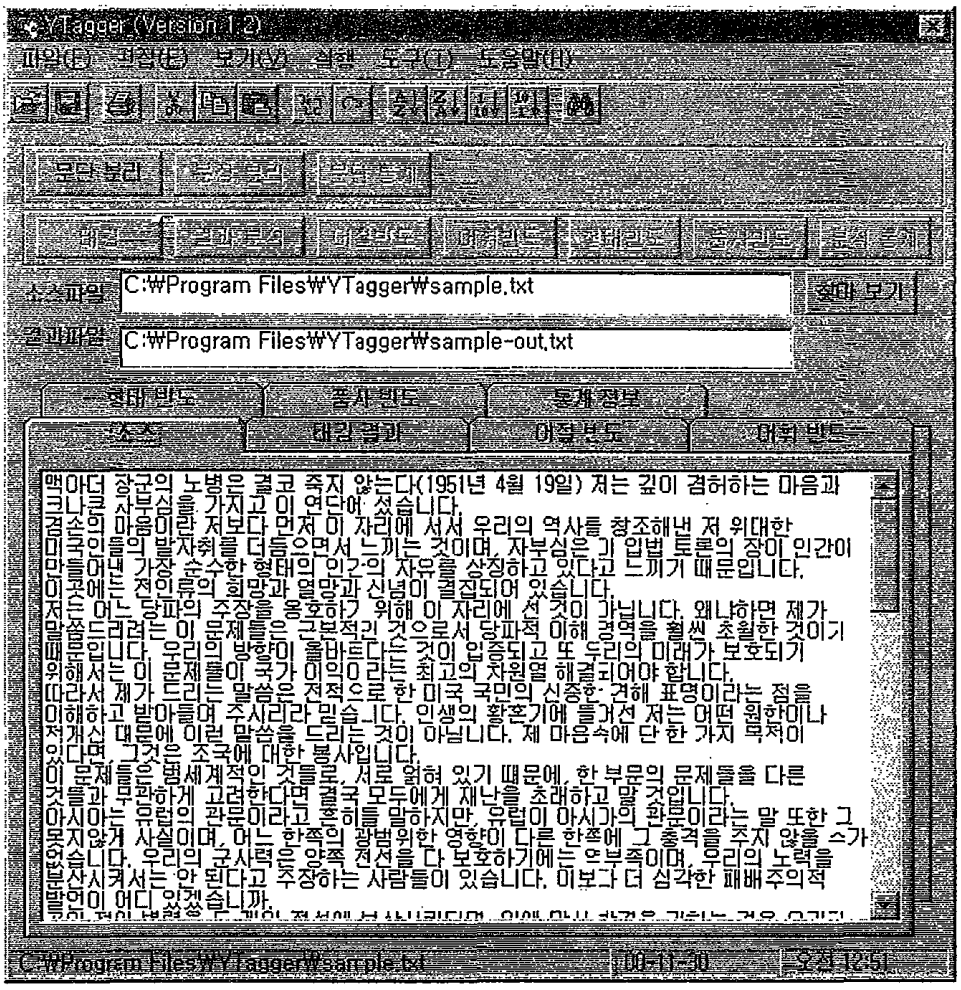
【소개】 이 프로그램은, 완성형 텍스트 파일을 대상으로 하여, 형태소 분석 결과를 주석하도록 한 뒤, 각종 언어 통계 정보를 뽑아 볼 수 있도록 개발된 것으로, 원하는 정보를 저장하여 활용할 수 있다.

【사양】 윈도우9x용으로 개발되었다.

【개발자】 연세대학교 언어정보개발연구원 장석배 연구원

【사용법】

○ 작업 대상 파일 열기 : 텍스트 파일 열기-소스 창에 텍스트 파일 표시됨.

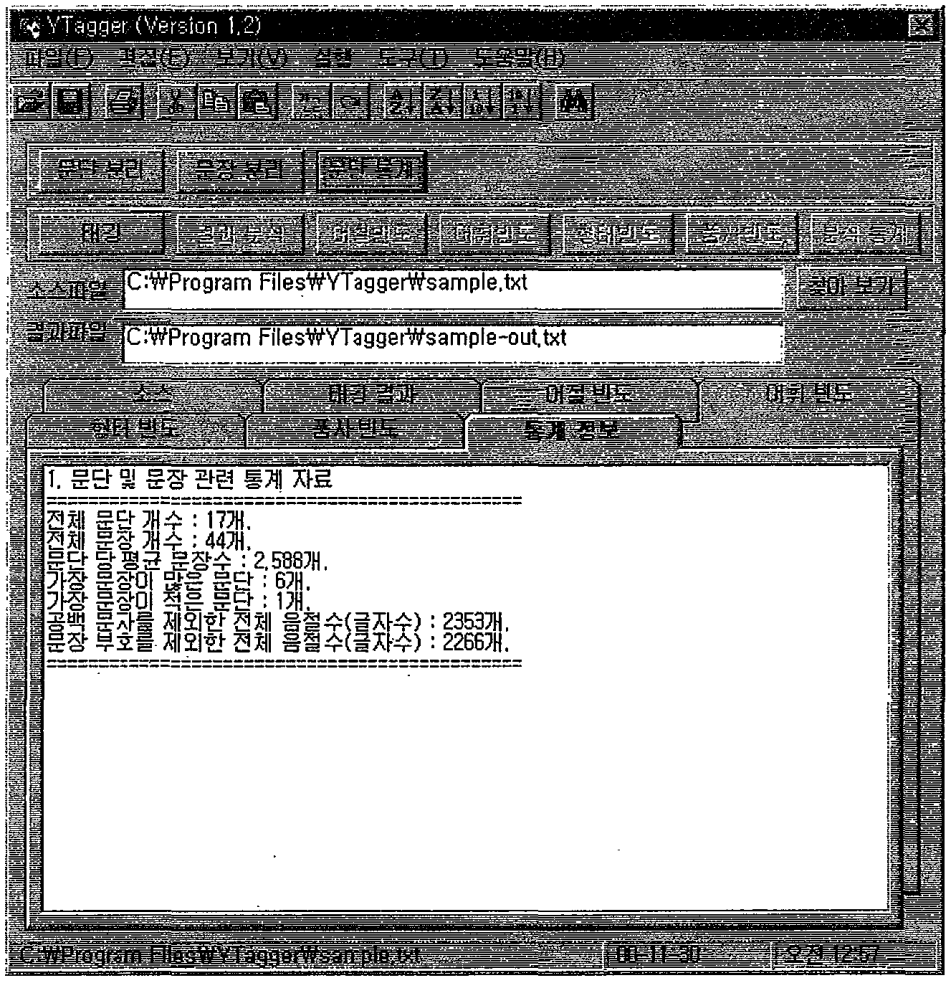


○ 문단 전처리

(1). [문단 분리] : 불러들인 텍스트 파일을 문단 단위로 분리한다.

(2). [문장 분리] : 문단 분리된 파일을 문장 단위로 분할하여, 품사 주석이 가능한 상태로 만들어 준다.

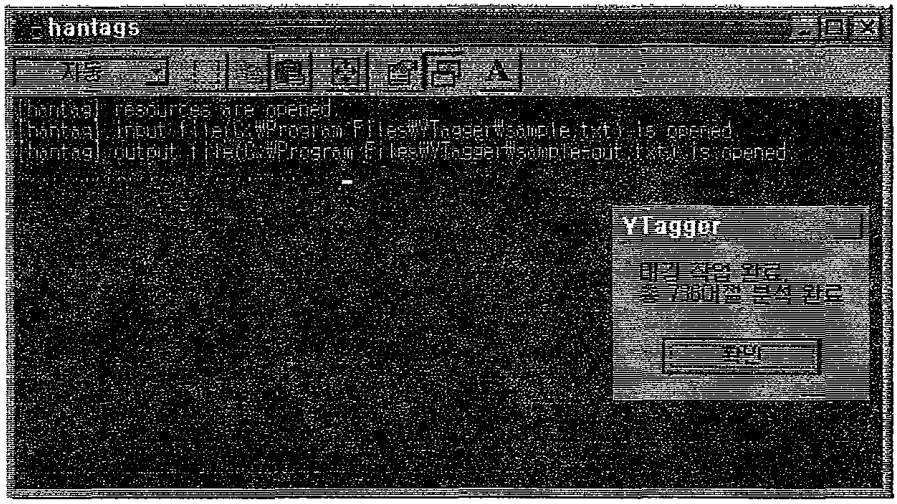
(3). [문단 통계] : 문단 및 문장 분리가 이루어지면, 아래의 그림에서와 같이, 기본적인 통계 자료를 얻을 수 있다.



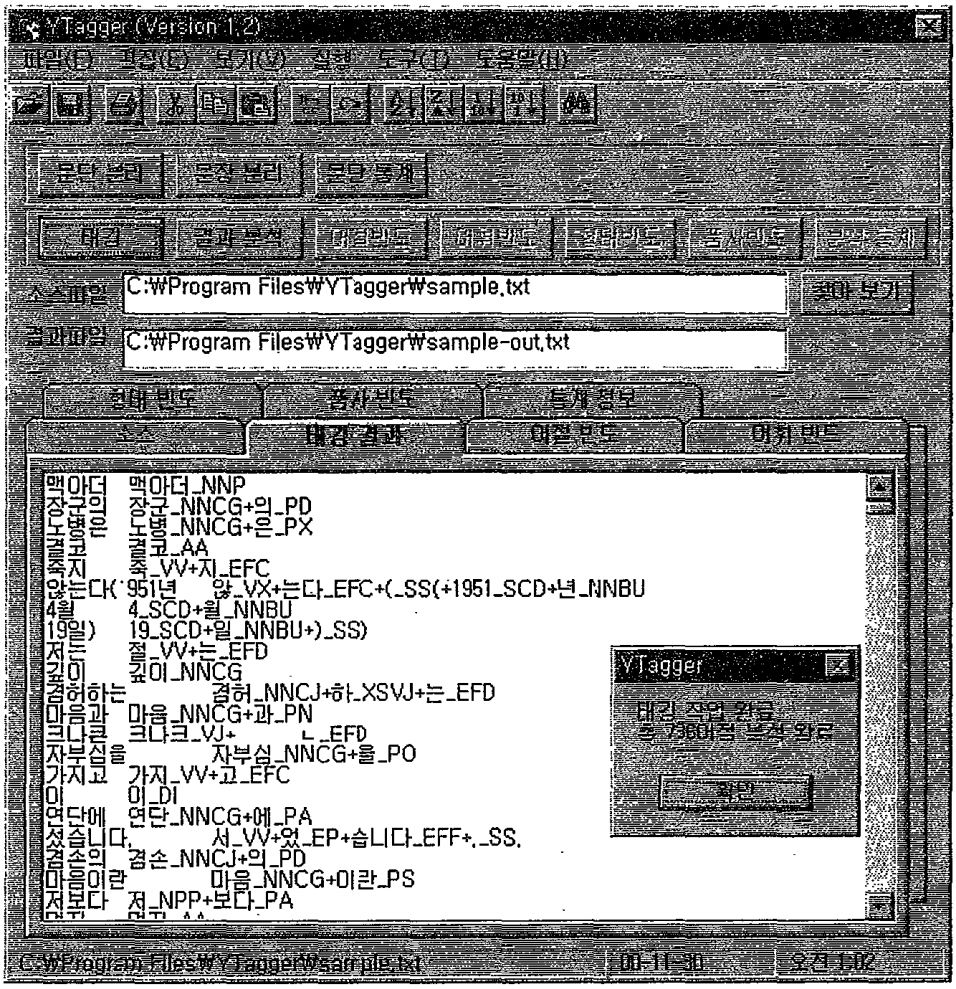
○ 자동 품사 주석

(1). 이 프로그램에서 원하는 통계 정보를 얻기 위해서는, 먼저 자동 품사 주석된 말뭉치로 만들어야 한다.

(2). 메뉴의 '실행'의 첫번째 탭을 누르거나, [태깅] 단추를 누르면 자동적으로 품사 주석을 수행한다.



(3). 작업이 완료되면, DOS창이 자동적으로 닫히고, 「태깅 결과」 탭 화면이 열림.

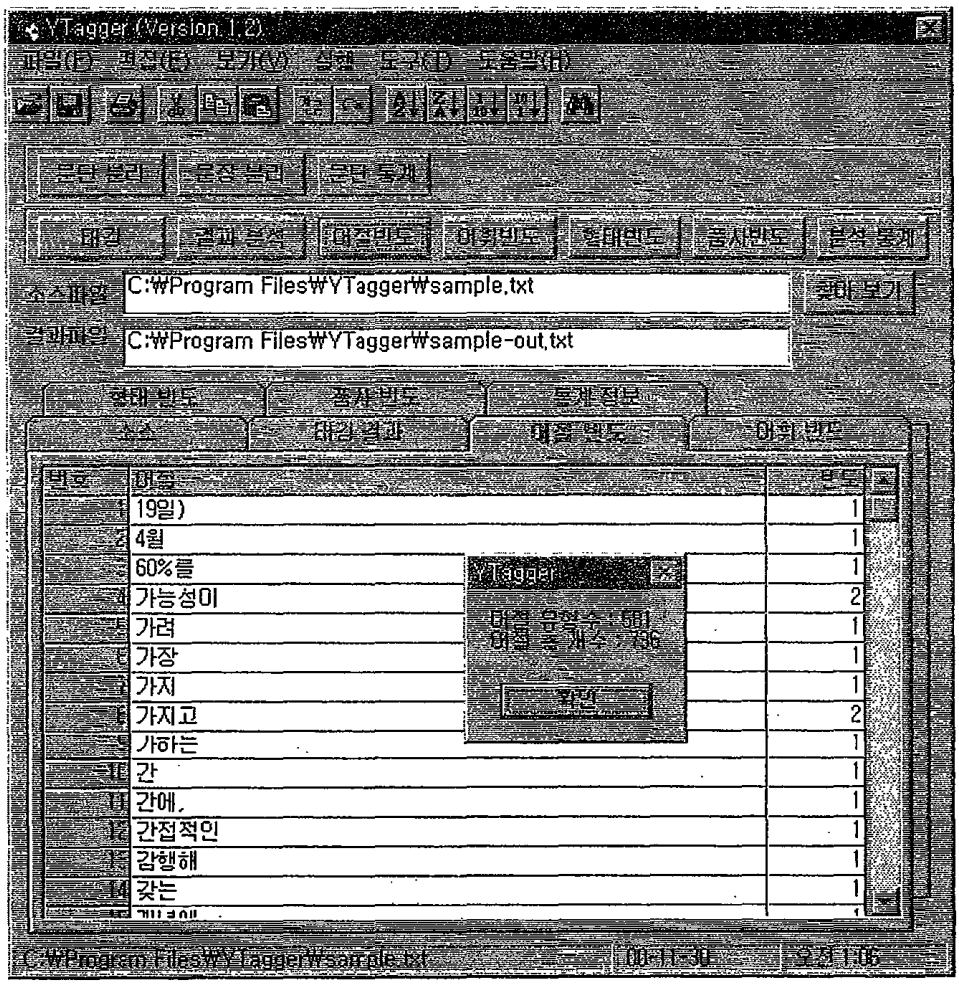


○ 결과 분석

(1). 품사 주석된 말뭉치로부터 각종 빈도 정보를 추출하기 위해서는, [결과 분석] 단추를 눌러, 실행이 완료 된 후, 원하는 빈도 추출 명령 단추(어절빈도, 어휘빈도, 형태소빈도, 품사 빈도)를 누름. 또한, 이 결과를 저장할 경우에는, 메뉴 [파

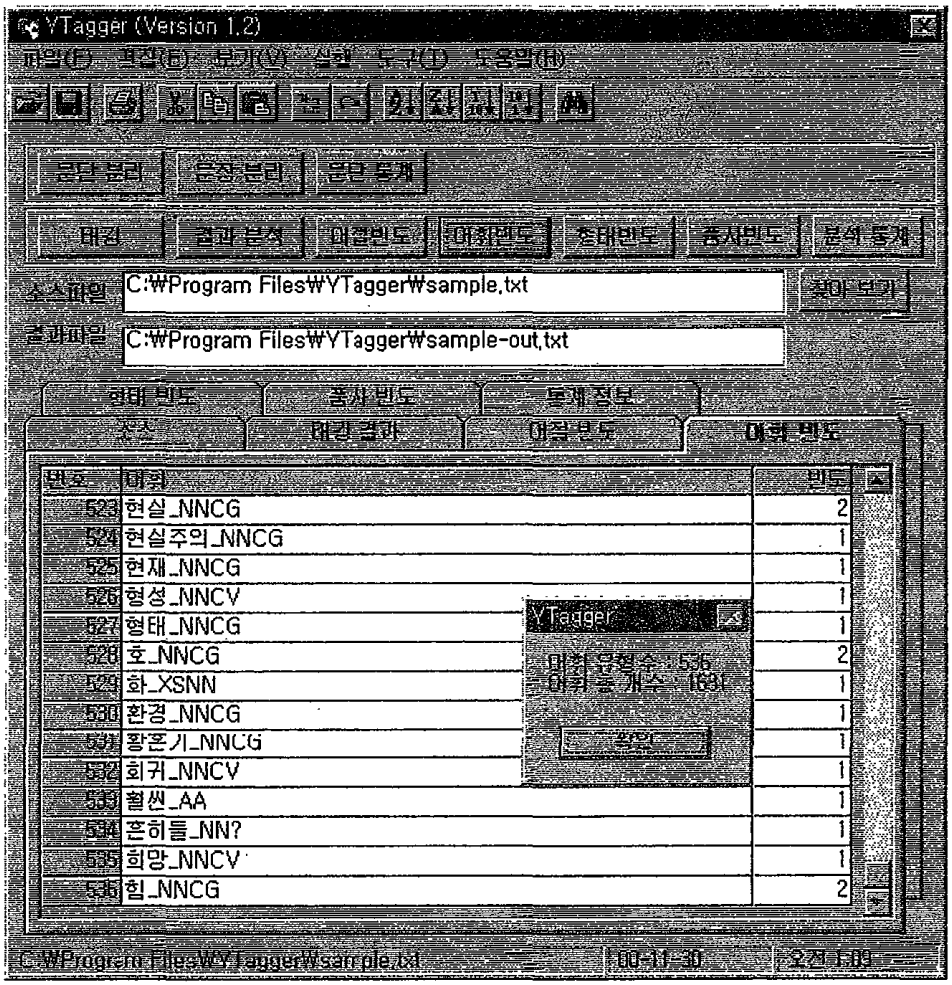
일]의 '저장' 또는 '다른 이름으로 저장'을 사용한다.

(2). [어절 빈도] :



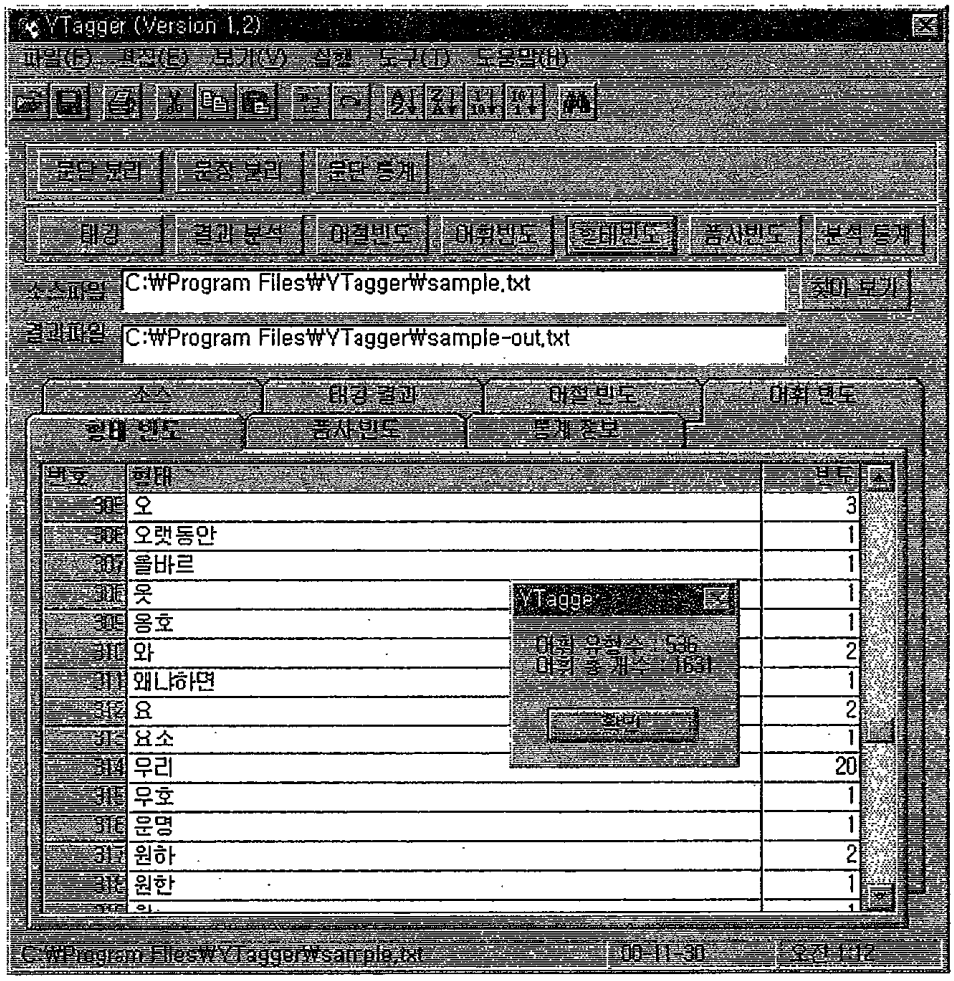
그림에서와 같이, 전체 어절 유형 수와 어절의 총 개수를 나타내는 작은 창이 뜬 후에, 결과 화면이 '어절 빈도' 탭 창에 표시되며, 오른쪽의 막대로 위아래로 이동하면서, 결과를 볼 수 있다.

(3). [어휘 빈도]



이 그림에서와 같이, 전체 어휘 유형 수와 어절의 총 개수를 나타내는 작은 창이 뜬 후에, 결과 화면이 [어휘 빈도] 탭 창에 표시되며, 오른쪽의 막대로 위아래로 이동하면서, 결과를 볼 수 있다.

(4). [형태 빈도]



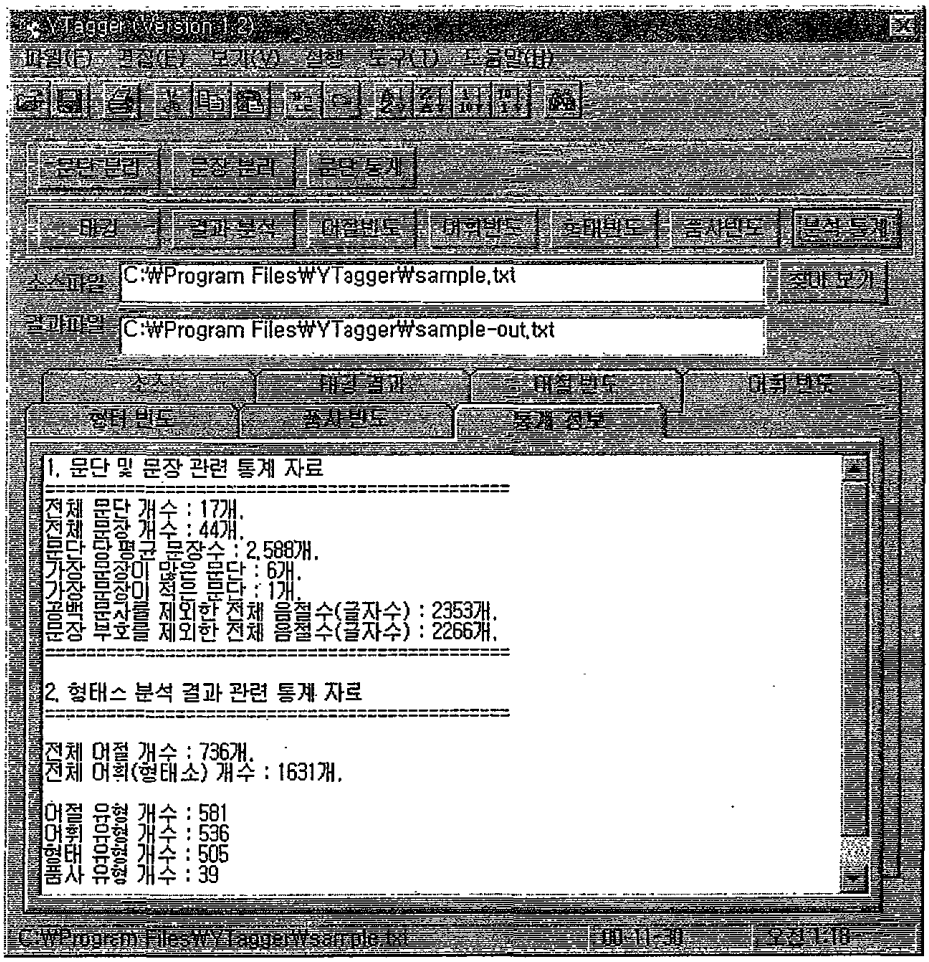
이 그림에서와 같이, 전체 형태 유형 수와 형태의 총 개수를 나타내는 작은 창이 뜬 후에, 결과 화면이 [형태 빈도] 탭 창에 표시되며, 오른쪽의 막대로 위아래로 이동하면서, 결과를 볼 수 있다.

(5). [품사 빈도]

전체 품사 유형 수와 품사의 총 개수를 나타내는 작은 창이 뜬 후에, 결과 화면이 [품사 빈도] 탭 창에 표시되며, 오른쪽의 막대로 위아래로 이동하면서, 결과를 볼 수 있다.

○ 통계 정보 보기

(1). 전체 모든 종류의 빈도를 확인할 경우에는, [분석 통계] 단추를 누른다.



2.4.4 반자동 의미 주석용 프로그램 '참뜻'

【소개】 이 프로그램은, 이미 품사 표지 주석된 말뭉치를 대상으로 하여, 어휘 의미 표지를 주석하도록 개발된 것이다. 어휘 의미의 기준이 되는 사전은 mdb 데이터베이스 형식의 데이터베이스로 만들면 되는데, 여기에서는 『연세 한국어 사전』의 데이터 샘플을 붙여서 제공한다. (10개 어휘)

【사양】 윈도우9x용으로 개발되었다.

【개발자】 연세대학교 언어정보개발연구원 장석배 연구원

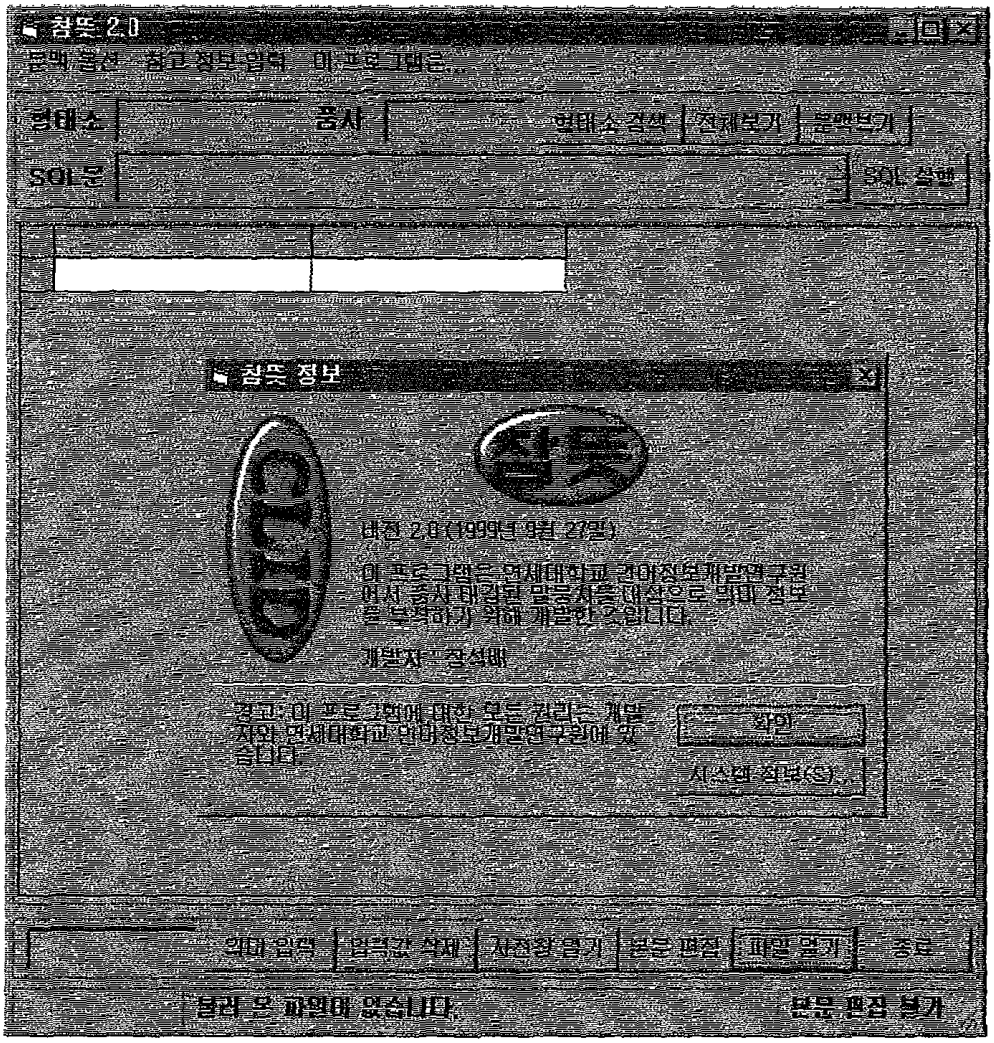
【부속 프로그램】 품사 주석된 말뭉치 파일(텍스트 파일)을 의미 표지 주석용의 파일로 변환하기 위한 '전처리개'를 사용하여, 먼저 파일을 변환하여야 한다.

【개발시기】 1999년 3월 버전 1.0, 5월 버전 1.5, 2000년 9월 버전 2.0

【사용법】

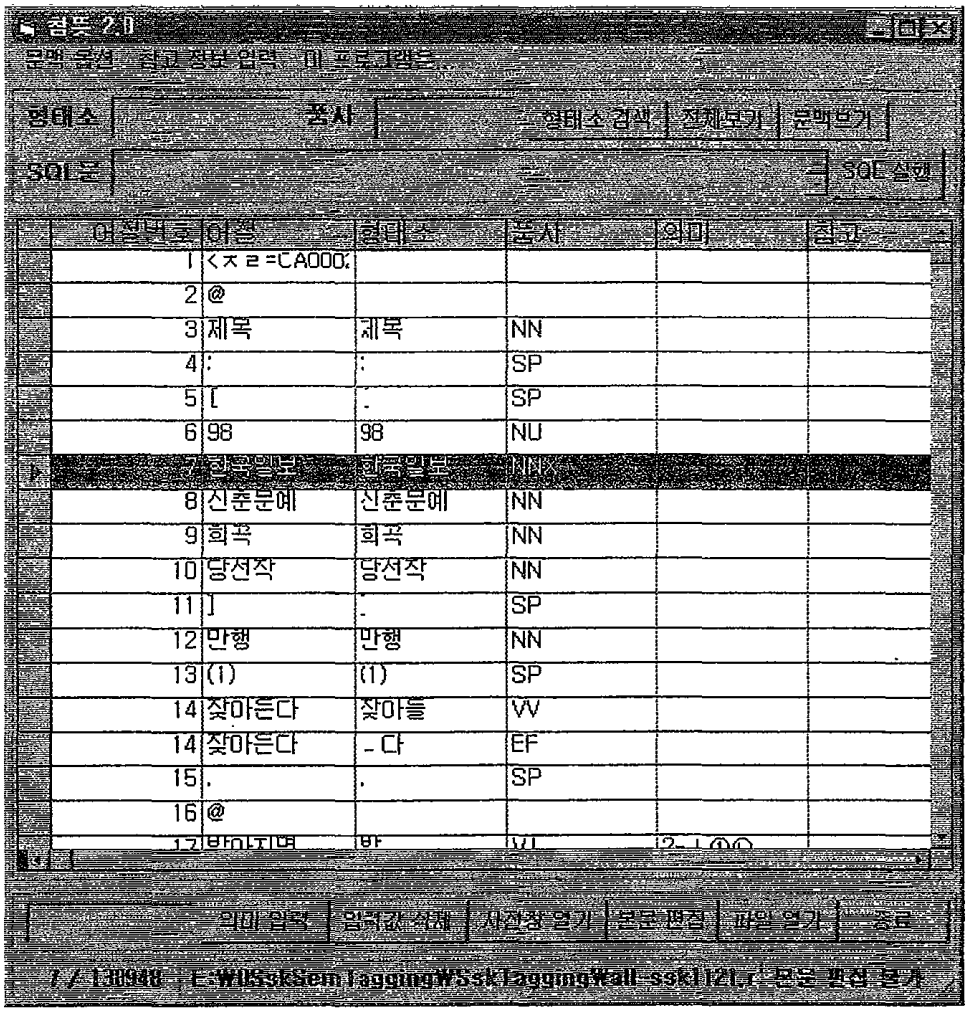
○ 초기화면

프로그램을 기동하면, 다음과 같은 초기 화면이 뜬다. 기본적인 기능은 맨 아랫부분의 5개의 메뉴 창과 맨 위의 옵션 등의 메뉴를 사용하게 된다.


참뜻 2.0
형태소
품사
형태소검색
전체보기
문맥보기
SQL문
SQL 실행
참뜻 정보
버전 2.0(1999년 9월 27일)
이 프로그램은 연세대학교 관광정보개발연구원
에서 중국어 관광 말뭉치 자료 데이터 정보
를 보완하기 위해 개발한 것입니다
개발자 : 장석배
시스템 정보(S)
엑셀 입력
입력값 삭제
자료창 열기
본문 편집
파일 열기
종료
불러 온 파일이 없습니다.
본문 편집 보기

○ 파일 열기

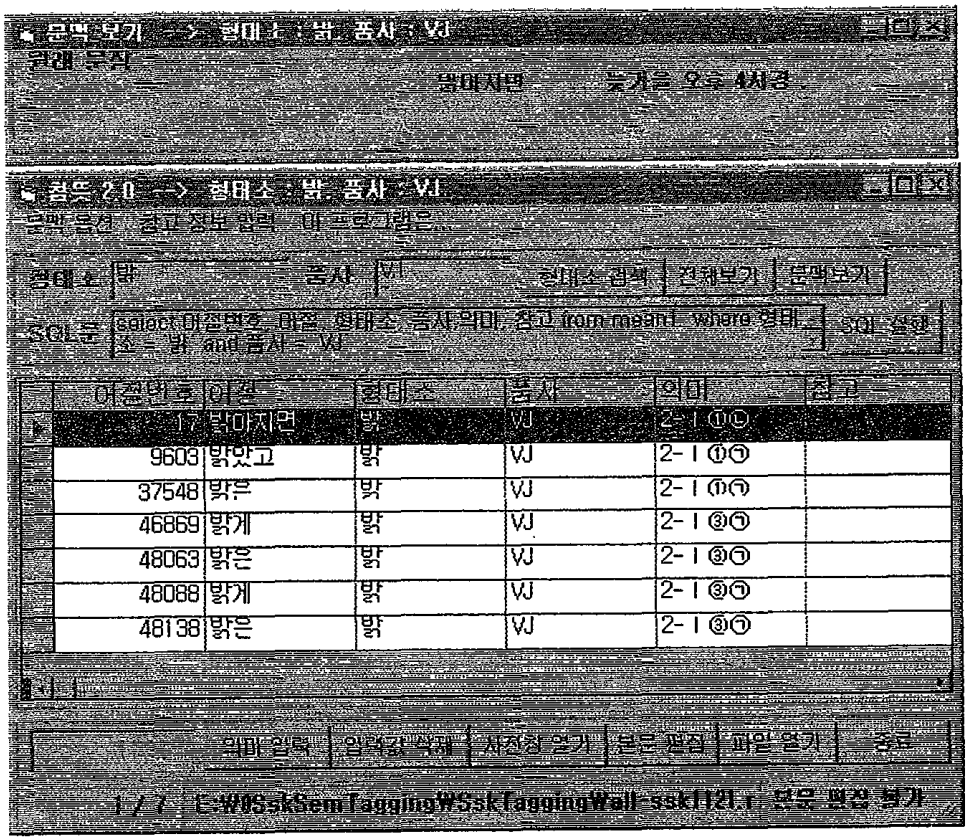
(1). [파일 열기] 단추를 눌러서, 원하는 폴더에 저장되어 있는 의미 주석 대상 말뭉치 파일을 연다. 여기에서는 'all-ssk1121.mdb' 파일을 연 화면을 볼 수 있다. 맨 아래의 숫자표시에서 보듯이, 이 파일은 모두 13만 8,948줄로 된 데이터임을 표시해 주고 있다.



의미 주석 대상 파일은, '어절번호, 어절, 형태소, 품사, 의미, 참고'의 항목으로 구성되도록 한다.

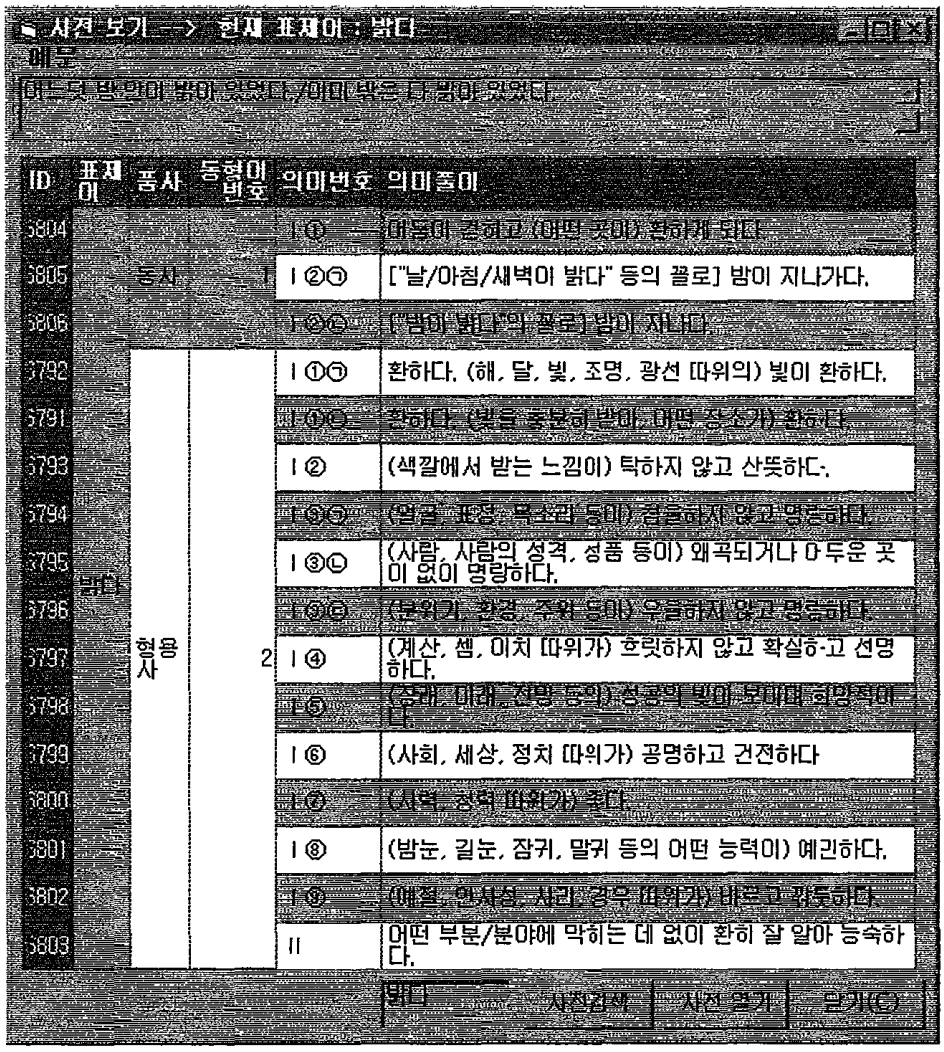
(2). 이 파일을 대상으로 하여, 의미 표지 주석을 할 대상 형태(어휘)를 검색하는데, 윗편의 '형태소'란에 원하는 어휘 형태를, 품사란에 품사 표지를 넣어서 검색할 수 있으며, 어느 한 쪽만으로도 검색이 가능하다.

(3). '밝다'(형용사 VJ)를 검색하면 다음과 같은 결과를 보여 준다.



(4). 검색어에 대해서 7개의 결과가 나왔으며, 맨 첫 줄을 선택하면, 본문 문맥을 윗창에 표시해 준다. (위의 예제에서는 이미 '의미'란에 의미 주석이 된 결과를 보여 주고 있으며, 작업이 되어 있지 않은 경우, 공란으로 표시된다.)

○ 사전창 열기



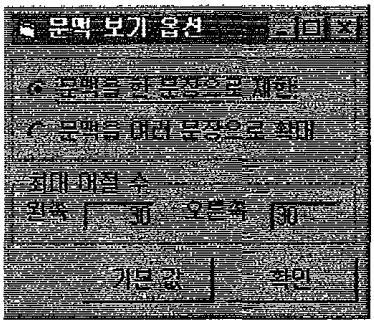
(1). 이제 작업 파일에 대해서 의미 주석을 하기 위한 사전을 열도록 한다. 아랫 메뉴의 가운데에 있는 '사전창 열기' 단추를 눌러, 사전이 있는 폴더와 파일을 선택해 주면(윈도의 탐색기 기능과 같음), 다음과 같은 사전창이 열린다.

(2). 여기에서는 사전을 열고, '밝다'를 검색한 결과를 보이기로 한다.

(3). 사전이 열리고, 사전의 의미 항목을 마우스로 누르면, 윗창(예문)에 해당되는 사전의 용례를 보여 준다.

○ 문맥 보기 옵션

(1). 데이터 창에서, 문맥을 조절할 수 있다.
(2). 윗메뉴에서 '문맥 옵션'을 누르면 아래와 같은 창이 열린다.

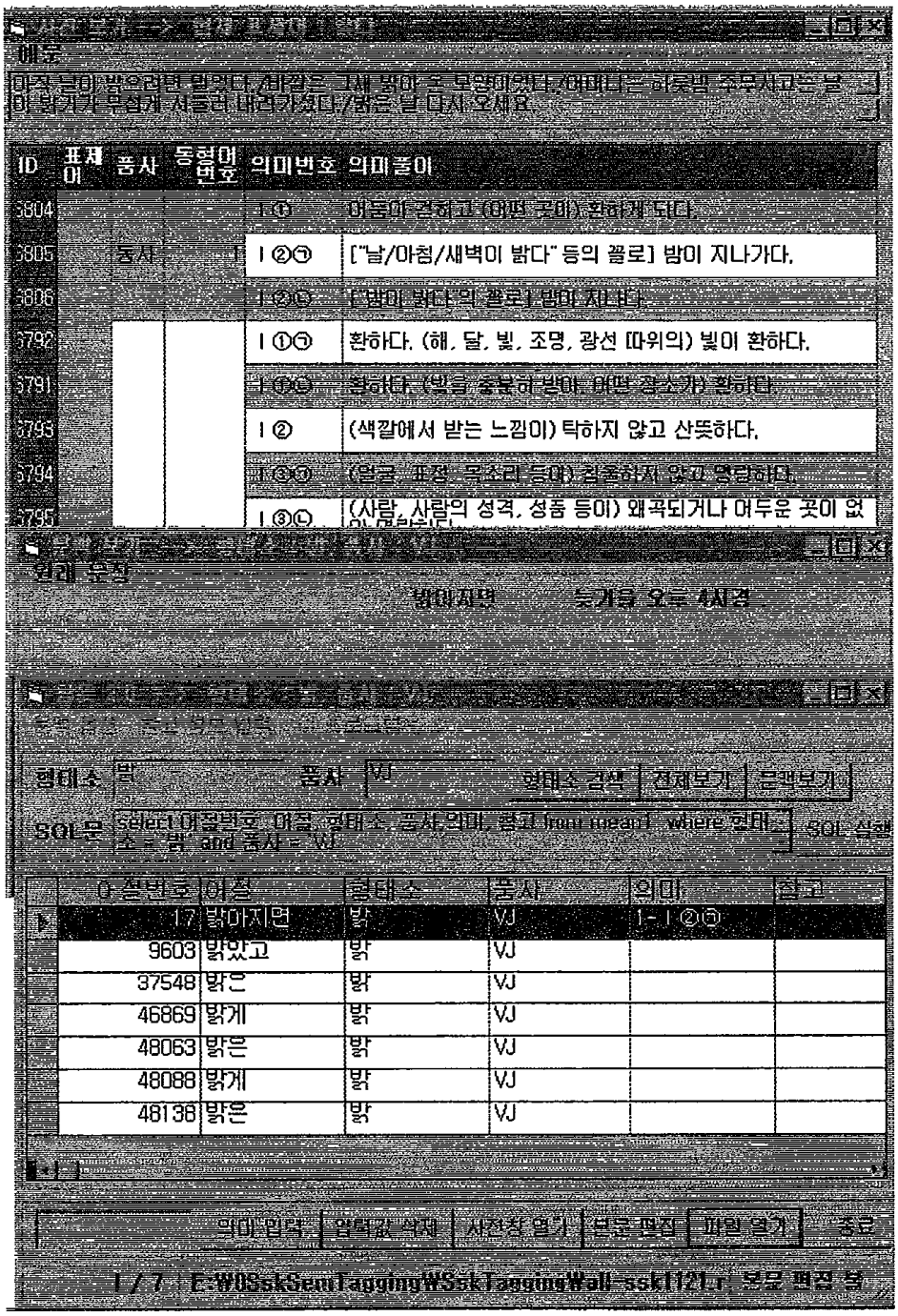


(3). 윗 부분에서는 보고 싶은 문맥을 해당 문장 하나에 국한할 것인지, 전후까지 볼 것인지를 지정할 수 있다.
(4). 한편, 보고 싶은 문맥의 길이를 어절 수로 지정할 수도 있다. 왼쪽과 오른쪽의 빈 상자에 숫자로 써 넣고, 확인을 누르면 된다.
(5). 초기값으로 돌리고 싶은 경우에는 '기본 값'을 누르면 된다. 기본 값은, 위의 그림과 같이 지정되어 있다.

○ 의미 주석

(1). 모니터 화면에 작업하기 쉽도록 적절히 데이터와 사전, 예문 창을 배열한 뒤[41], 예문의 의미에 해당하는 사전의 의미 풀이 부분을 더불클릭하면, 데이터 창에 의미 표지가 자동으로 입력되게 된다.
(2). 이 의미 풀이를 지우고 싶을 경우에는, 아랫편 메뉴의 '입력값 삭제'를 누르면 된다.

41) 대체적으로 800×600의 해상도로는 전체를 다 표시할 수 없는 경우가 많아, 최소한 1024×768 이상의 해상도가 필요하다.


해운
아직 달이 밝으려면 멀었다. 새벽은 그새 밤이 온 모양이었다. 아파트는 하룻밤 조용해졌는 달이 밝기가 무섭게 사흘러 내려가 있다./밝은 날 다시 오세요.
ID 표제어 품사 동형어 번호 의미번호 의미풀이
6804 1①
6805 동사 1 1②㉠ ["날/아침/새벽이 밝다" 등의 꼴로] 밤이 지나가다.
6806 1②㉡ ["밤이 밝다"의 꼴로] 밤이 지나다.
6792 1①㉠ 환하다. (해, 달, 빛, 조명, 광선 따위의) 빛이 환하다.
6791 1①㉡ 환하다. (방, 등, 등불 따위, 어떤 장소가) 환하다.
6793 1② (색깔에서 받는 느낌이) 탁하지 않고 산뜻하다.
6794 1③㉠ (말, 표정, 목소리 등이) 침울하지 않고 명랑하다.
6795 1③㉡ (사람, 사람의 성격, 성품 등이) 왜곡되거나 어두운 곳이 없
원래 문장
밤이지만 늦가을 오후 4시경
형태소 밝 품사 VJ 형태소 검색 전체보기 본문보기
SOL문 Select 어절번호, 어절, 형태소, 품사, 의미, 참고 from mean1 where 형태소 = '밝' and 품사 = 'VJ' SOL 실행
어절번호 어절 형태소 품사 의미 참고
17 밝아지면 밝 VJ 1-1②㉠
9603 밝았고 밝 VJ
37548 밝은 밝 VJ
46869 밝게 밝 VJ
48063 밝은 밝 VJ
48088 밝게 밝 VJ
48138 밝은 밝 VJ
의미 입력 입력값 삭제 사전창 열기 본문 편집 파일 열기 종료
1 / 7 : E:₩0SskSemTaggingWSskTaggingWall_sskt121.r, 본문 편집 중

○ 의미 주석된 말뭉치 보기

1	1	[즐거운생활2-2]		1		
353	233	@	1			
359	239	1	1	1	ASC	
360	240	.	1	.	SP	
361	241	팔	1	팔	NU	2-Ⅱ
362	242	월에도	1	월	NX	2-Ⅱ
363	242	월에도	2	에도	PA	
364	243	추석날은	1	추석날	NN	
365	243	추석날은	2	은	PA	
366	244	즐거운	1	즐겁	VJ	
367	244	즐거운	2	은	EF	
368	245	명절	1	명절	NN	
369	246	,	1	,	SP	
370	247	밤	1	밤	NN	2-
371	248	먹고	1	먹	VV	1-Ⅰ①㉠
372	249	대추	1	대추	NN	
373	250	먹고	1	먹	VV	1-Ⅰ①㉠
374	250	먹고	2	고	EF	
375	251	송편도	1	송편	NN	
376	251	송편도	2	도	PA	
377	252	먹고	1	먹	VV	1-Ⅰ①㉠
378	252	먹고	2	고	EF	
379	253	.	1	.	SP	
380	254	@	1			
381	255	2	1	2	ASC	
382	256	.	1	.	SP	
383	257	팔	1	팔	NU	2-Ⅱ
384	258	월에도	1	월	NX	2-Ⅱ
385	258	월에도	2	에도	PA	
386	259	추석날은	1	추석날	NN	
387	259	추석날은	2	은	PA	
388	260	달이	1	달	NN	1-
389	260	달이	2	이	PA	
390	261	밝은	1	밝	VJ	
391	261	밝은	2	은	EF	
392	262	밤	1	밤	NN	1-
393	263	,	1	,	SP	
394	264	손에	1	손	NN	1-①
395	264	손에	2	에	PA	
396	265	손을	1	손	NN	1-①
397	265	손을	2	을	PA	
398	266	잡고서	1	잡	VV	1-Ⅰ①

2.4.5 전처리 프로그램 "전처리개 Ver1.5"

【소개】 이 프로그램은, 이미 품사 표지 주석된 말뭉치를 대상으로 하여, 어휘 의미 주석을 위한 mdb 파일 형식으로 바꾸어 저장하도록 개발된 것이다.

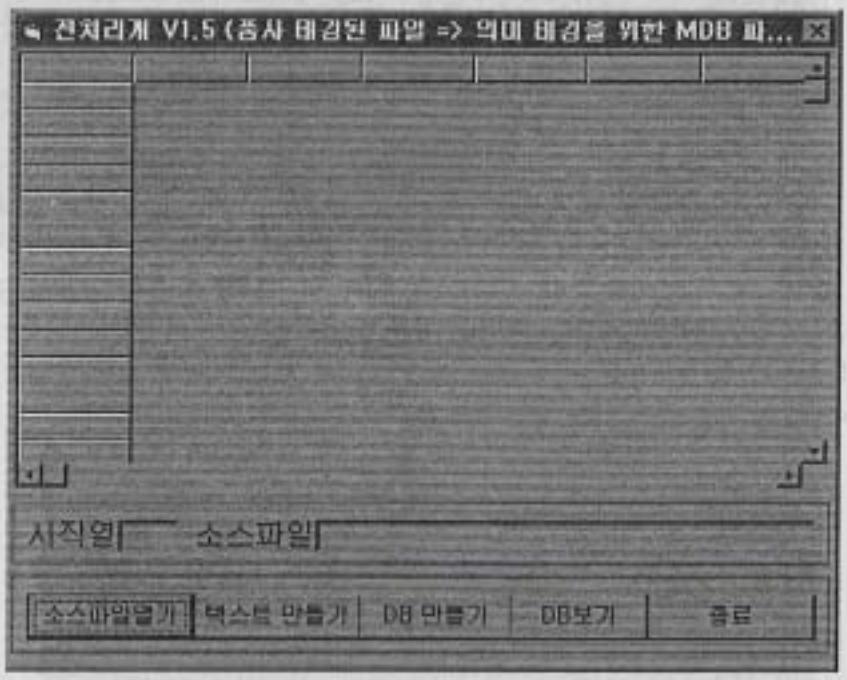
【사양】 윈도우9x용으로 개발되었다.

【개발자】 연세대학교 언어정보개발연구원 최준호 연구원

【개발시기】 1999년 9월 버전 1.0, 5월 버전 1.5, 2000년 3월 버전 1.5

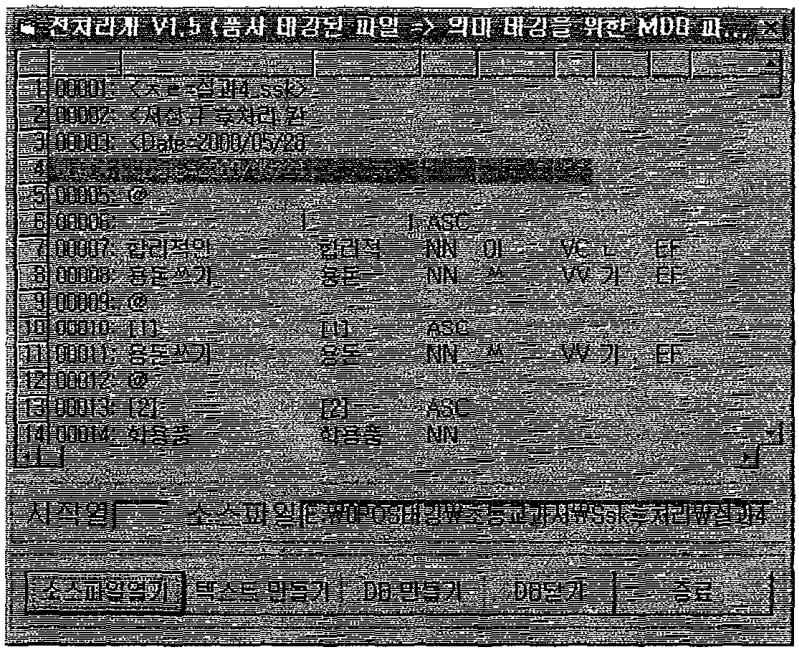
【사용법】

○ 초기화면

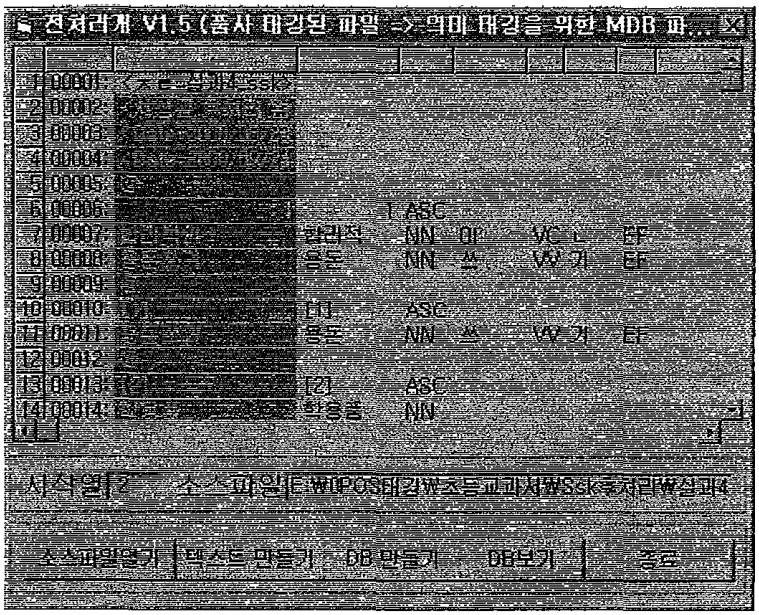


○ 소스 파일 열기

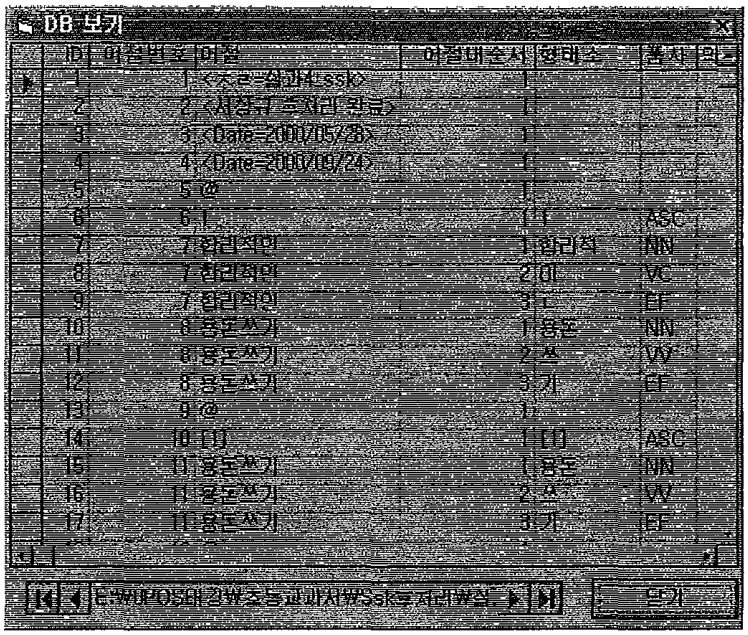
(1). 원하는 폴더에서 품사 주석된 파일을 선택한다. 그러면 다음과 같이, 항목 구분이 되어 있는 데이터의 내용이 창에 표시된다. 이 때, 품사 주석 말뭉치의 항목은 맨 왼쪽으로부터, 어절 번호, 본문 어절, 분석 결과 및 품사의 순으로 구분되어 있다.



(2). 품사 주석 말뭉치 파일을, 의미 주석용을 바꿀 때에는 시작되는 열을 설정하여야 하는데, 윗 화면에서처럼 본문 파일의 시작 부분을 선택한 뒤, 'DB 만들기'를 누른다.



(3). 변환 작업이 끝나고, 결과를 보기 위해서는 'DB 보기'를 누르면 되는데, 그 결과를 보이면 다음과 같다.



3. 실습용 말뭉치의 활용

대표적 한국어 말뭉치를 중심으로 구성된 실습용 말뭉치를 구성하며, 이 과정을 통하여 실제 목적에 맞는 말뭉치 활용 방법을 익힌다.

3.1 날말뭉치

1) 말뭉치의 구성

표본의 대표성과 균형성을 고려하기 위해서 독서 실태에 대한 설문을 중심으로 하여 구성된 연세 말뭉치 I 의 구성안[42]을 따랐다.

2) 표본의 수와 크기

약 2,000 어절 크기의 파일 100개로 20만 어절에 해당하는 날말뭉치를 포함한다.

3) 파일의 소스

21세기 세종계획의 결과로 구축된 말뭉치와 연세말뭉치로 이루어져 있다.

4) 세부 목록

날말뭉치목록.xls(첨부 시디롬)

3.2 품사 주석 말뭉치

1) 말뭉치의 구성

품사 주석 말뭉치는 4개의 종류로 이루어져 있다. 연세대학교, 한국과학기술원, 한국 전자통신 연구원, 21세기 세종계획 프로젝트에서 각각 구축한 품사 주석

42) '우리말 낱말 빈도 조사 표본의 선정 기준'(정찬섭 외), 『사전편찬학 연구 3집』,탑출판사, 1990 참조 바람. 구체적인 텍스트의 내용과 목록은, 부속 시디롬의 "날말뭉치목록.xls" 파일 참조.

말뭉치의 샘플을 보였다.

2) 표본의 수와 크기

약 2000 어절 크기의 파일을 구축기관별로 25개씩 모두 100개를 선정하여 20만 어절에 해당하는 품사주석말뭉치의 샘플을 보였다.

3) 세부 목록

품사주석말뭉치.xls(첨부 시디롬 참조)

4) 주석 지침(첨부 시디롬 참조)

① 세종품사주석말뭉치태그셋.hwp
② 연세품사주석말뭉치태그셋.hwp
③ ETRI품사주석말뭉치태그셋.doc
④ KAIST품사주석말뭉치태그셋.hwp

3.3 의미 주석 말뭉치

1) 문화관광부 주관 한국어 세계화 프로젝트의 결과물로 구축된 연세 의미 주석 말뭉치에서 약 5만 어절의 샘플을 제시했다.

2) 세부 목록

의미주석말뭉치.xls(첨부 시디롬 참조)

3.4 파일 이름

파일 이름에는 네 가지의 요소를 반영했다. 각각의 요소는 '_'로 연결된다.

1) 말뭉치의 종류

① 1 : 날말뭉치
② 2 : 품사주석말뭉치

③ 3 : 의미주석말뭉치

2) 내용의 분류

각 구축 기관의 분류를 고려하여 소설, 신문, 정보 등의 간단한 내용 분류를 표현했다.

3) 구축 기관

① 연세대학교 : ys

② 한국과학기술원 : kaist

③ 한국전자통신연구원 : etri

④ 세종계획 : sj

4) 제목

알아보기 쉽게 하기 위해서 되도록 책 혹은 텍스트의 제목을 붙였다.