

연구보고서 【2004년】

21세기 세종계획 국어 특수자료 구축

총괄연구책임자
임 용 기

세부과제연구책임자
김 향 희
이 석 재
전 형 식
소 강 춘
임 용 기
이 원 표

문 화 관 광 부
국 립 국 어 원



제 출 문

문화관광부 장관 귀하

<21세기 세종계획> 국어 특수자료 구축에 관하여 귀부와 체결한
연구 용역 계약에 의하여 연구 보고서를 작성하여 제출합니다.

2004년 12월 14일

총괄 연구 책임자: 임용기(연세대학교)

- 세부 과제명: 현대 국어 구어 전사 말뭉치 개발
책 임 자: 김향희(연세대학교)
연 구 원: 서상규(연세대학교)
연구보조원: 김수진·성연숙 외 5명
보 조 원: 정은숙 외 10명
- 세부 과제명: 한영 병렬 말뭉치 개발
책 임 자: 이석재(연세대학교)
연구보조원: 봉미경·황은하 외 10명

- 세부 과제명: 한일 병렬 말뭉치 개발
 책 임 자: 전형식(고려대학교)
 연구보조원: 마지마 도모히데 외 6명

- 세부과제명: 북한 및 해외 한국어 말뭉치 개발
 책 임 자: 소강춘(전주대학교)
 연구보조원: 김지선 외 1명
 보 조 원: 손영숙 외 9명

- 세부과제명: 역사자료 말뭉치 개발
 책 임 자: 임용기(연세대학교)
 연 구 원: 김상태(충주대학교)
 연구보조원: 권호진 외 2명
 보 조 원: 이은지 외 3명

- 세부과제명: 전문용어 말뭉치 개발
 책 임 자: 이원표(연세대학교)
 연구보조원: 최정도
 보 조 원: 김동휘 외 3명

요 약 문

1. 사업명

국어 기초자료 구축 — 국어 특수자료 구축 소분과 —

2. 연구의 목적

이 연구는 우리나라의 선진 정보 문화의 기본 바탕과 자원을 확충하는 국어정보화 중장기 발전계획의 일환으로 수립된 <21세기 세종 계획>의 10년간(1998~2007) 사업 중 제 3단계의 1차년도 작업에 해당한다.

21세기 정보화 사회에서 국민 전체의 언어·문화 생활을 향상시키고 과학 기술 및 학문을 발전시키기 위해서는 대규모 언어 자료를 축적하여 정보·지식으로 가공하는 일이 중요하다. 언어 규범과 추상적 이론 위주의 언어 정책이나 교육으로는 효과적 대응이 어려우며, 국어 생활을 과학적 토대 위에서 개선시키고 국어의 특성에 적합한 정보 문화와 새로운 정보 처리 이론 및 기술을 발전시켜야 한다. 이를 위해서는 기초적 연구로서 우수한 품질과 대용량의 규모를 갖춘 국어 자료의 지속적 축적이 필요하다. 국어 자료의 데이터베이스를 구축함으로써 언어 정책의 수립과 교육의 효율화, 정보 처리의 생산성 향상을 꾀할 수 있고 국민들 모두가 쉽고 편리하게 언어 정보를 찾고 활용할 수 있다. 이러한 기초 자료에는 문어 말뭉치뿐만 아니라 구어 말뭉치, 병렬 말뭉치 등의 특수 자료 말뭉치 또한 포함되어야 한다. 따라서 본 사업은 국가적 역량을 기반으로 국어 기초 자료 중 쉽게 접근하기 힘든 특수 자료를 대규모의 국가 말뭉치 수준으로 구축하는 것을 목적으로 한다.

1단계부터 2단계까지 구축한 자료를 효율적으로 이용하고 보완·확장해 나가기 위해서는 국어 정보화를 위한 국가적 기반의 조직적, 지속적 구축이 필요하다. 따라서 본 연구는 국어 정보 처리, 국어 연구, 국어 교육, 국어 정책, 국어 생활, 정보 산업, 번역 등 여러 부문에 활용될 수 있는 특수자료의 기초 자원과 응용 자원을 개발하고자 한다.

특수자료 구축 소분과 3단계 1차년도 사업의 목적은 다음과 같다.

첫째, 3단계 1차년도 사업에서는 그동안 추진해 온 현대국어 구어 말뭉치, 한영 병렬 말뭉치, 한일 병렬 말뭉치, 역사자료 말뭉치, 북한 및 해외 한국어 말뭉치 구축과 함께, 그동안 전문용어 정비 분과에서 추진해 온 전문용어 원시 말뭉치 구축을 진행시킨다.

둘째, 자료의 활용도를 높이기 위하여 구축된 기초 자료를 지속적으로 가공한다. 구축된 자료를 다방면으로 활용하고 일반 말뭉치와 비교 연구하기 위해서는 자료를 언어학적으로 분석하고 기본적인 형태소 분석 표지를 부여하는 작업이 필수적이다. 따라서 2단계에서 형태소 주석 말뭉치 구축을 위한 기초 연구 및 시험 구축을 토대로 형태소 분석 말뭉치를 구축해 온 바 있다. 당해연도인 3단계 1차년도에서도 2단계 작업에 이어서 형태소 분석 말뭉치 구축을 지속적으로 진행시킬 예정이다. 또한 실질적인 연구 및 활용을 위한 연구, 배포용 말뭉치 구성 및 통합·조정 작업을 추진해 나갈 예정이다. 이렇게 구축된 말뭉치는 문어와 구어의 대조 연구, 한국어와 외국어의 대조 연구, 한·영, 한·일 번역문 간의 대조 연구, 일반 어 텍스트와 전문 분야 텍스트의 대조 연구, 북한어와 남한어의 대조 연구, 현대어와 고어의 대조 연구 등에 응용될 수 있을 것이다.

셋째, 일반 현대 국어 말뭉치에 비해 자료가 희귀하고 구축의 난이도가 높은 특수자료를 말뭉치의 균형성과 사업의 궁극적인 목표에 맞도록 구축한다.

넷째, <21세기 세종 계획>의 다른 분과 사업과의 연관성을 고려하면서 사업을 추진한다. 전문용어 말뭉치는 '전문용어 정비' 분과에서 응용하기에 적합한 기초자료로 구축하고 역사자료 말뭉치는 한민족 언어정보화와 사업과의 학제적 교류 속

에서 응용가능성을 모색하며 추진한다.

3. 연구의 필요성

국어 생활을 과학적 토대 위에서 향상시키고 국어의 특성에 적합한 정보 문화와 새로운 정보 처리 이론 및 기술을 발전시키기 위해서는 그 기초적 연구로서 대규모의 국어 텍스트 말뭉치가 필요하다. 국어 자료의 데이터베이스를 구축함으로써 언어 정책의 수립과 교육의 효율화, 그리고 정보 처리의 생산성을 향상시킬 수 있고, 국민들 모두가 쉽고 편리하게 언어 정보를 찾고 활용할 수 있게 해야 하는 것이다. 그러기 위해서는 우수한 품질과 대용량의 규모를 갖춘 국어 자료의 지속적 축적이 필요하다.

구어는 말뭉치 언어학 연구의 초기부터 언어 연구와 긴밀한 관계를 맺어 왔으나, 구어 말뭉치의 필요성에 대한 문제가 제기되고 이에 대한 기초적인 연구가 시작된 것은 비교적 최근의 일이다. 구어 말뭉치의 개발은 그동안의 문어 중심 국어 연구의 문제점에 대한 해결의 실마리를 제공할 수 있다. 또한 구어의 실제 쓰임이 반영된 사전이나 문법서를 제작할 수 있는 말뭉치를 제공하여 국어 교육이나 외국어로서의 한국어 교육에 이바지할 수 있다. 뿐만 아니라 최근 음성 언어 정보 처리에 대한 각 연구 기관과 기업체의 관심과 투자가 확대되고 있음을 생각할 때, 폭넓은 영역에서 구축된 구어 자료와 이에 대한 기초 가공 작업은 이러한 기술 개발에 일정 부분 기여를 할 것이다.

2003년까지 구축된 세종 구어 전사 말뭉치는 원시 294만 어절, 형태 주석 말뭉치 50만 어절에 불과하다. 그에 비해 서구 선진국의 경우에는 국가와 수많은 정보 산업 관련 기업체, 기술 개발자들을 중심으로 대규모의 사업들이 진행되어 왔다. 이 말뭉치들은 대부분 1천만 어절을 넘었으며 실제적 이용을 위해 다양한 언어학적 층위에서의 주석 작업이 진행되고 있다. 한국어 구어 말뭉치의 경우에도 기초 가공 작업이 지속적으로 이루어져야 하며, 이를 바탕으로 3단계 이후 배포될 연구용 구어 말뭉치가 구성될 수 있을 것이다.

● 영국국가말뭉치(British National Corpus)

- 구축 기관 : 옥스퍼드대학 출판부, 롱맨 출판사, 옥스퍼드 대학, 랭카스터 대학 등 6개 기관이 영국 상공부와 과학기술연구협의체로부터 50%의 지원을 받음
- 구축 규모 : 1천만 어절(문어 포함 1억 어절)
- 구축 기간 : 1991.1 ~ 1994년
- 구축 목적 : 사전편찬학, 인공 지능, 음성 인식·합성, 문헌학 등 다방면의 연구 및 응용을 위한 기초 자료 제공. 2001년에는 구어 자료 1천만 어절 전체에 대한 어휘 빈도 작업과 문어 어휘 빈도와 비교 연구.

● 코빌드 구어 말뭉치(The COBUILD Spoken Corpus)

- 구축 기관 : 버밍햄 대학, 콜린즈 출판사
- 구축 규모 : 1천만 어절 이상(문어 포함 4억 5천만 어절, 구어 자료는 대부분 90년대 이후의 자료임)
- 구축 기간 : 1991년 ~ 현재
- 구축 목적 : 사전편찬학, 영어학, 언어 교육, 번역학 등의 연구를 위한 기초 자료 제공 및 철자 검사기, 기계 번역 시스템 등 정보처리산업 분야로의 응용.

● The International Corpus of English

- 구축 기관 : 런던 대학 SEU(the Survey of English Usage) 센터를 중심으로, 15개 국가(지역)가 프로젝트에 참가
- 구축 규모 : 각 나라에서 100만 어절씩 모두 1천 500만 어절 구축(문어·구어 포함)
- 구축 목적 : 전 세계의 영어 변이형 연구. 통사론, 형태론, 어휘론, 사회언어학, 담화 분석, 음성학, 음운론 등의 연구를 위한 자료 제공 및 언어 교육, 자연 언어 처리 등 관련 분야로의 응용.

자연 언어 처리 및 언어 공학적 요소가 필요한 분야(기계 번역, 번역 검증 시스템, 언어 교육, 대조언어학 등)에서 경쟁력을 높이기 위해서는 둘 이상의 언어를 병렬적으로 대응시켜 구성한 병렬(대역) 말뭉치가 필수적이다. 특히 국가 간의 경계를 넘나드는 정치·문화·경제 등의 교류가 활발해지고 있는 오늘날에는 기계 번역의 필요성이 급속하게 커지고 있다. 병렬 말뭉치는 언어간의 효율적인 정보 소통을 위한 통역, 번역 등의 실용적인 측면뿐만 아니라 언어간의 상이한 특징을 비교하는 언어의 대조 분석 연구라는 학문적인 측면에서도 중요한 역할을 한다.

외국에서의 병렬 말뭉치는 대부분 유럽과 미국을 중심으로 한 서구어 사이의 대응을 기반으로 활발히 구축되고 있다(예: 영어-불어, 영어-독일어, 영어-노르웨이어 등). 또한 말뭉치의 구축뿐만 아니라 이를 활용하기 위한 사용자 인터페이스 도구에 대한 개발 및 텍스트의 정렬 등이 다양하게 연구되고 있다. 더욱이 병렬 말뭉치에 대한 응용 분야의 요구가 보다 증가할 것으로 여겨지기 때문에 병렬 말뭉치의 구축 및 도구의 개발은 더욱 가속화될 전망이다

서구 선진국가들(특히 유럽의 국가들)의 대표적인 병렬 말뭉치 구축 및 관련 기술 개발 프로젝트는 다음과 같다.

◎ The English-Norwegian Parallel Corpus(영어 : 노르웨이어)

- ① 구축 기관 : Oslo Univ.
- ② 구축 규모 : 100개의 영어 원본 텍스트와 100개의 노르웨이어 번역본 텍스트, 100개의 노르웨이어 원본 텍스트와 100개의 영어 번역본 텍스트를 합하여 약 260만 어휘의 병렬 말뭉치 구축
- ③ 구축 기간 : 1994~1997
- ④ 관련 프로그램 : 번역 말뭉치 정렬 프로그램, 번역 말뭉치 검색기
- ⑤ 중심 연구 분야 : 언어의 대조 분석, 번역 연구

◎ The Finnish-English Contrastive Corpus(영어 : 핀란드어)

- ① 구축 기관 : Jyväskylä Univ.
- ② 중심 연구 분야 : 언어의 대조 분석, 언어의 구조와 사용 양상에 대한 연구, 말뭉치 기반 언어 연구를 반영한 응용 프로그램 개발

◎ Text_based contrastive studies in English(영어 : 스웨덴어)

- ① 구축 기관 : Lund Univ.
- ② 구축 규모 : 160만 어휘
- ③ 구축 기간 : 1993년 ~ 현재
- ④ 중심 연구 분야 : 인식론적 양상, 담화 표지, 문법적 초점 기제

◎ MULTEXT(Multilingual Text Tools and Corpora)

- ① 구축 기관: EC(the European Commission: LRE - the Linguistic Research and Engineering, Copernicus), 미국의 NSF(US: the Vassar/CNRS collaboration), 프랑스의 AUPELF-UREF(the Fonds Francophone pour la Recherche)와 CNRS(French)the Universite de Provence)
- ② 대상 언어: 체코어, 네덜란드어, 영어, 불어, 독일어, 스페인어, 이탈리아어, 스웨덴어, 밤바라어(Bambara), 불가리아어, 카탈란어, 에스토니아어 등

◎ MULTEXT-EAST

- ① 구축 기관: Laboratoire Parole et Langage, Centre National de la Recherche Scientifique (CNRS)
- ② 구축 목적 : MULTEXT의 기술, 방법론, 도구를 응용하여 동구언어에 적용
- ③ 대상 언어: 6개의 동구언어(불가리아어, 체코어, 에스토니아어, 헝가리어, 루마니아어, 슬로바니아어)
- ④ 중심 연구 분야: 언어표준과 부호화(CES)의 검증과 적용, 다중언어(6개국어) 말뭉치 개발(형태소 태깅, 병렬 말뭉치), MULTEXT의 말뭉치 도구의 적용, 형태소-어휘 자원 개발

◎ MLCC (Multilingual Corpora for Cooperation)

- ① 구축 기관: 제네바 대학의 ISCCO
- ② 구축 목적: EAGLES의 말뭉치 습득 프로젝트로서 6개의 대상언어로 된

comparable sets와 9개 언어의 병렬데이터 세트 구축

③ 지원 기관: EC의 EC Telematics program과 스위스 연방정부

◎ The ECI Multilingual Corpus (ECI: European Corpus Initiative)

① 구축 기관: 영국 에딘버러 대학의 HCRC와 스위스 제네바 대학의 ISSCO

② 구축 목적: 다중언어 말뭉치의 구축과 다중언어의 문자, 음성 말뭉치의 설계, 수집, 출판 지원

③ 구축 규모: 9천8백만 단어의 ECI/MCI 말뭉치 구축

④ 대상 언어: 주요 유럽어와, 터키어, 일어, 러시아어, 중국어, 말레이어 등

위와 같은 선진 외국의 개발 상황과, 기계 번역 산업계나 연구계가 자력으로 병렬 말뭉치를 구축하지 못하는 국내의 현 상황을 고려할 때, 세종 병렬 말뭉치의 개발은 필수적이다. 그러나 세종 병렬 말뭉치는 아직까지 한·영과 한·일 병렬 말뭉치만을 대상으로 개발되고 있는 실정이며, 더욱이 한·일 병렬 말뭉치의 경우 구축 시작 시기가 늦었던 만큼 한·영 병렬 말뭉치에 비해 상대적으로 구축량의 증가가 더 필요하다고 할 수 있다. 또한, 국제적 언어 정보의 교류를 위해서는 이국어 병렬 말뭉치에서 다국어 병렬 말뭉치로의 확장이 필요하다. 따라서 3단계 1차년도부터는 병렬 말뭉치의 구축량을 균형있게 증가시키면서 향후 다양한 대응어의 병렬 말뭉치 개발을 위한 시험 말뭉치 구축 연구가 이루어져야 한다.

국어 자료의 데이터베이스 구축 작업에서 역사 자료의 말뭉치 구축 작업은 특히나 중요하다. 국어의 발달 변천사나 방언론과 같은 국어 연구와 국어 교육, 그리고 어문 생활의 발전을 기하기 위해서는 역사 자료들의 정보화가 필요한 것이다. 그것은 훈민정음이라는 우리의 세계적 문화유산을 보존하고 더욱 드높이는 길이기도 하다. 고어와 방언에 대한 이해는 이해 그 자체에 그치는 것이 아니라 현대국어의 근원을 파악할 수 있게 해주고, 현대의 어문 생활의 길잡이가 된다는 점에서도 중요하다. 최근 해외의 말뭉치 구축 사업의 동향도 현대어 이외의 고어와 방언을 포괄하는 방향으로 확산되고 있음은 그 필요성을 잘 말해주고 있다.

현대 국어의 올바른 언어 분석을 위해서는 역사적인 언어 정보에 토대를 두지 않으면 안 될 것이며, 역사 자료는 국어 연구와 교육뿐 아니라, 사전편찬학, 우리 문화의 정보화·국제화의 기반을 이루게 될 것이다.

1단계에서부터 꾸준히 구축해 온 역사 자료 말뭉치를 실제 연구에 이용하기 위해서는 역사 자료를 적절한 언어단위로 분석하고 형태소 분석 작업을 수행하는 것이 필수적이다. 역사 자료의 형태소 분석 말뭉치를 개발하는 데에 있어서 가장 큰 문제점은 띄어쓰기의 구분이 없다는 자료 자체의 특성이다. 띄어쓰기의 구분이 없고 자료에 포함되어 있는 한자음, 방점 등의 요소를 동시에 고려해야 하기 때문에 역사 자료에 자동적으로 형태소 분석 표지를 부착하거나 검색하는 것은 거의 불가능하다. 따라서 형태소 분석 표지를 부착할 대상이 되는 언어 단위를 결정하고 역사 자료의 특성을 그대로 드러낼 수 있는 태그의 셋을 정하는 등의 기초 연구부터 단계별로 형태소 분석의 작업을 진행하기로 한다.

4. 연구내용 및 결과물

당해연도의 세부 과제별 연구 내용 및 결과물은 다음과 같다.

<당해년도 연구 목표와 내용>

1) 현대 국어 구어 전사 말뭉치 개발

분량 : 57만 어절

원시 말뭉치

39만 어절

형태소 분석 말뭉치

18만 어절

내용 : 말뭉치 기반 기술언어학, 사전편찬학, 담화 분석, 실험음성학 등 구어에 대한 언어학적 연구와 언어 교육, 언어 병리학, 자연언어 처

리 분야, 음성공학 분야 등 기타 관련 분야로의 응용을 위한 다양한 구어 전사 말뭉치와 형태소 분석 말뭉치를 개발한다.

2) 한·영 병렬 말뭉치 개발

분량 : 67만 어절

원시 말뭉치(한·영)

48만 어절

형태소 분석 말뭉치

19만 어절

내용 : 기계 번역, 외국어로서의 한국어 교육, 한국인을 위한 외국어 교육, 기타 비교언어학 연구와 응용을 위한 한국어-영어 병렬 말뭉치와 형태소 분석 말뭉치를 개발한다.

3) 한·일 병렬 말뭉치 개발

분량 : 34만 어절

원시 말뭉치

22만 어절

형태소 분석 말뭉치

12만 어절

내용 : 기계 번역, 한국어 교육 등의 응용 분야와 문체 연구, 비교 언어학 연구 등의 연구를 위한 한국어-일본어 병렬 말뭉치와 2003년도의 기초 연구를 토대로 한 형태소 분석 말뭉치를 개발한다. 또한, 2004년도는 전년도 형태소 분석 말뭉치 구축 시에 드러난 문제점을 개선하여, 한일 형태소 분석 말뭉치를 구축함에 있어서 유용한 정보를 제공할 수 있도록 연구한다.

4) 북한 및 해외 한국어 말뭉치 개발

분량 : 135만 어절

북한 및 해외 한국어 원시 말뭉치 100만 어절

형태소 분석 말뭉치 35만 어절

내용 : 남북한의 언어적 통합을 위한 연구와 정책 수립에 기초가 되는, 북한 지역의 현대 한국어 말뭉치와 해외 한국어 말뭉치를 구축하고, 이의 형태소 분석 말뭉치를 개발한다.

5) 역사 자료 말뭉치 개발

분량 : 원시 말뭉치 70만 어절

주석 말뭉치 5만 어절

내용 : 국어의 발달 변천사를 밝히고, 국어 교육과 어문 생활의 발전을 위한 국어사 자료 말뭉치를 개발한다.

6) 전문용어 말뭉치 개발

분량 : 원시 말뭉치 100만 어절

내용 : 전문용어의 형태와 목록을 추출할 수 있도록, 균형성과 정보성이 고려된 전문분야 도서와 자료들을 선정하여 전문분야 말뭉치를 구축한다.

<특수 자료 구축 소분과 총괄 결과물>

1. 특수 자료 구축 결과 보고서 책자 200부
2. 결과 보고서 파일(CD-ROM 형태로 10장)
3. 말뭉치 결과물(CD-ROM 형태로 10장, 총 433만 어절)

- (1) 연구 결과물은 유니코드 텍스트 파일 또는 한글 2002 이상으로 구축한다.
- (2) 표준화 지침에 따라 각종 문서 정보를 기입하고, 최종 오류율이 0.1% 이내가 되도록 한다.
- (3) 각 과제별 말뭉치 결과물의 양은 다음과 같다.
 - ① 구어 전사 자료 57만 어절 구축
(원시 말뭉치 39만 어절+형태소 분석 말뭉치 18만 어절)
 - ② 한·영 병렬 말뭉치 67만 어절 구축
(원시 말뭉치 48만 어절+형태소 분석 말뭉치 19만 어절)
 - ③ 한·일 병렬 말뭉치 34만 어절 구축
(원시 말뭉치 22만 어절+형태소 분석 말뭉치 12만 어절)
 - ④ 북한 및 해외 한국어 말뭉치 135만 어절
(원시 말뭉치 100만 어절+형태소 분석 말뭉치 35만 어절)
 - ⑤ 역사 자료 말뭉치 75만 어절
(원시 말뭉치 70만 어절+형태소 분석 말뭉치 5만 어절)
 - ⑥ 전문용어 말뭉치 100만 어절
(원시 말뭉치 100만 어절)

4. 저작권 승인을 받은 저작권 양의 각서 일체(원본)

- (1) 현대 국어 구어 전사 말뭉치
- (2) 한·영 병렬 말뭉치
- (3) 한·일 병렬 말뭉치
- (4) 전문용어 말뭉치

5. 말뭉치 원자료 사본

- (1) 현대 국어 구어 전사 말뭉치 : 녹음 자료 사본 일체(오디오·비디오 테이프 또는 CD-ROM 형태의 음성 파일)
- (2) 한·영 병렬 말뭉치 - 자체 보관
- (3) 한·일 병렬 말뭉치 - 자체 보관
- (4) 북한 및 해외 한국어 말뭉치 - 자체 보관
- (5) 역사 자료 말뭉치
- (6) 전문용어 말뭉치

5. 기대효과

- ① 우리말 사용 실태를 정확하게 분석할 수 있는 기초 자료를 제공하여 국어 생활과 국어 교육, 국어의 변천사, 국어 정보 처리 등 관련 어문 정책의 수립과 운용의 토대를 마련한다. 또한 이는 일반 언중의 어문 생활의 발전으로 연결된다.
- ② 국어 연구와 국어 정보 처리에 유용한 자료를 제공하여 정보 검색, 언어 지식 정보 기반 구축, 기계 번역 등의 발전을 촉진함으로써 국가적 생산성과 경쟁력을 향상시키는 기간 자원이 된다.
- ③ 국어의 변천사를 연구함에 있어 좀더 효율적인 연구를 가능하게 한다. 방언 연구의 경우 어형 확인이라든지 처음 사용된 시기 등을 쉽게 알 수 있으며, 어원 연구의 경우 어떤 어휘의 어원 및 의미 용법의 변화 등을 정밀하게 밝힐 수 있도록 한다. 또한 사라져가는 순수 국어 어휘 등을 발굴할 수 있고 이것을 바탕으로 현대 국어의 조어법과 관련시켜 연구할 수 있는 토대를 제공한다.
- ④ 장기간의 남북 분단으로 인한 언어적 이질화의 실상을 구체적으로 파악할 수 있게 하고, 이질화를 해소하기 위한 자료적 토대를 제공한다.
- ⑤ 인간 언어의 자연스러운 모습을 드러내 주는 음성적 실체에 보다 용이하게 접근할 수 있게 함으로써 음성 언어 연구 및 응용을 위한 기반을 마련한다.

6. 활용방안

본 연구의 성과는 이후의 기초 언어 자료 구축 연구에 계속 반영될 것이며, 적절한 과정을 거쳐 전문연구기관, 학자, 정보 산업체 및 국어 연구 교육 분야에 보급됨으로써 국어 정보화의 전사회적 역량을 증진하는 데 기여하도록 한다.

결과물의 보급은 자료의 특성과 저작권 등의 사항을 고려하여 다음의 두 가지 경로를 통해 이루어지도록 한다.

1) 세종 계획 홈페이지와 국어 정보화 종합 처리실

- 현재 운영되고 있는 세종 계획 홈페이지는 국어 정보화 사업인 '21세기 세종 계획'의 홍보와 함께 저작권 등의 일반 보급에 문제가 없는 자료들을 일반에 공개하고 있다. 기초자료뿐 아니라 세종 계획의 다른 분과의 결과물도 일부 제공되고 있다.
- 결과물 센터에서는 2002년도에 '국어 정보화 종합 처리실'을 개장해 '21세기 세종 계획'의 모든 결과물(공개·비공개)을 검색할 수 있도록 하고 있다.

2) 전문기관 연구용 배포

- 매년 세종 계획을 통해 구축되는 자료들은 주관기관(국립 국어연구원)이 판단하여 전문기관에 한정적으로 배포하고 있다.
- 이 자료들은 사업의 평가나 세종 계획 내 분과간 협력 등의 목적으로 이용되고 있다.

차 례

<사업의 개요>	1
1. 1·2단계 연구 추진 실적 및 2004년도 연구 목표	1
2. 2004년도 사업 목표	3
3. 연구 추진 방법	5
 제1부 원시 말뭉치의 구축	15
 1장. 현대 국어 구어 전사 말뭉치 개발	19
1. 연구의 필요성 및 목표	19
2. 구어 전사 원시 말뭉치 개발	21
3. 말뭉치 구축 지침	35
4. 현대 국어 구어 전사 말뭉치 구축 실례	36
 2장. 한영 병렬 말뭉치 개발	45
1. 연구의 필요성 및 목표	47
2. 한영 병렬 원시 말뭉치 개발	48
3. 말뭉치 목록 및 구성	55
4. 한영 병렬 말뭉치 구축 실례	65
 3장. 한일 병렬 말뭉치 개발	91
1. 개요	93
2. 설계 원칙	94
3. 작업 지침	94
4. 작업 파일 목록	103
5. 말뭉치 지도	108
6. 한일 병렬 말뭉치 구축 실례	110
 4장. 북한 및 해외 한국어 말뭉치 개발	127
1. 연구의 필요성, 목표, 대상	129
2. 북한 말뭉치 구축 결과	131

3. 연구 추진 방법	134
4. 결론 및 제언	147
5장. 역사자료 말뭉치 개발	149
1. 연구 목적	151
2. 원시 말뭉치 개발	151
3. 역사자료 구축 지침	152
4. 기존 말뭉치와의 차이점	162
5. 실제 말뭉치 작업의 예	163
6. 구축된 말뭉치 현황	170
6장. 전문용어 말뭉치 개발	173
1. 연구의 필요성, 목적 및 방법	175
2. 말뭉치 구축 결과	177
3. 연구 추진 방법	184
4. KORTEM과 세종 말뭉치의 차이점	198
5. 문제점 및 향후 계획	206
제2부 형태소 분석 말뭉치의 구축	209
1장. 현대 국어 구어 전사 말뭉치 개발	211
1. 연구의 필요성 및 목표	213
2. 구어 전사 형태소 분석 말뭉치 개발	214
3. 현대 국어 구어 형태소 분석 말뭉치 구축 사례	222
2장. 한영 병렬 말뭉치 개발	229
1. 연구의 필요성 및 목표	231
2. 한영 형태소 분석 말뭉치 구축	232
3. 형태소 분석 지침 보완 및 논의 사항	235
4. 한영 형태소 분석 말뭉치의 예	241
3장. 한일 병렬 말뭉치 개발	253

1. 연구 개요	255
2. 한일 형태소 분석 말뭉치 구축	256
3. 한일 형태소 분석 말뭉치 구축 도구 및 구축 방법	261
4. 논의 사항에 대한 견해	263
5. 한일 병렬 형태소 분석 말뭉치 구축의 실례	264
<부록> 한일 병렬 일본어 문서 형태소 분석 말뭉치 구축 지침	286
 4장. 북한 및 해외 한국어 말뭉치 개발	305
1. 주식 말뭉치 개발	307
2. 북한어 말뭉치 주식 지침(2003-수정본)	309
 5장. 역사자료 말뭉치 개발	329
1. 품사 주식 말뭉치	331
2. 역사자료 말뭉치 주식 체계	333
 제3부 전문가 초청 특강 발표 요지	351
 1장. 전문가 초청 특강 일지	353
1. 제1차 전문가 초청 특강	355
2. 제2차 전문가 초청 특강	379
3. 제3차 전문가 초청 특강	397
4. 제4차 전문가 초청 특강	407

〈사업의 개요〉

〈사업의 개요〉

1. 1·2단계 연구 추진 실적 및 2004년도 연구목표

본 연구는 세계적 수준의 한국어 언어 자료 데이터베이스를 구축하고, 구축되는 자료가 국어정보화의 광범위한 영역에서 지속적으로 활용되기 위한 표준적 규격과 안을 마련하는 것을 그 목표로 한다. 본 연구의 제1 단계는 총 10년간의 전체 사업 기간 중 1998~2000년까지 이루어졌고, 본 연구의 제 2단계는 2001~2003년까지의 3년 동안 진행되었다. 제 3단계는 2004~2007년까지의 4년 동안에 이루어지며 올해는 3단계 시작 연도로서 그간의 연구 실적과 3단계 1차년도의 최종 목표를 정리하면 다음의 표와 같다.

① 1단계(1998~2000) 추진 실적

구분 \ 연도		1단계			합계 (단위:어절)
		1998년	1999년	2000년	
현대국어 구어 전사 말뭉치		50만	50만	50만	150만
한영 병렬 말뭉치		기초 연구	25만	75만	100만
북한 해외	북한	35만	105만	130만	270만
	해외	100만	25만		125만
역사 자료 말뭉치		70만	75만	100만	245만
구비문학 말뭉치		100만	130만		230만
합계		355만	410만	355만	1,120만

② 2단계(2001-2003) 추진 실적 및 목표

구분 \ 연도			2001	2002	2003	합계		주안점
현대국어 구어 전사 말뭉치	원시 말뭉치		55만	40만	49만	144만	194만	▪균형 말뭉치 구성을 위한 다양한 말뭉치 확보에 노력
	형태소 분석 말뭉치		5만	25만	20만	50만		
한·영 병렬 말뭉치 [다국어 병렬 말뭉치]	원시 말뭉치	한·영	80만	90만	85만	255만	322만	▪획득 자료의 특수성 극복, 균형성 확보에 노력 ▪다국어 병렬 말뭉치 개발을 위해 한·중/한·러/한·불 병렬 말뭉치를 시험 구축
		한·중 한·러 한·불			15만	15만		
	한·영 형태소 분석 말뭉치		5만	30만	17만	52만		
한·일 병렬 말뭉치	원시 말뭉치		13만	18만	24만	55만	60만	▪다양한 장르의 말뭉치 확보에 노력
	형태소 분석 말뭉치				5만	5만		
북한 및 해외 한국어 말뭉치	북한 및 해외 한국어	원시 말뭉치	140만	100만	80만	320만	394만	▪해외 한국어 자료의 균형성 고려(러시아의 한국어 추가)
		형태소 분석 말뭉치	10만	25만	39만	74만		
역사 자료 말뭉치	원시 말뭉치		62만	20만	72만	154만	206만	▪사이버 한글 박물관 디지털화 자료 중 미입력분 개발 ▪개화기 미입력자료/기독교 및 필사본 자료
	형태소 분석 말뭉치		20만	32만		52만		
말뭉치 가공 및 활용 시스템 개발			▪북한 문화어 형태소 분석 시스템 개발 ▪한·영 병렬 말뭉치 용례 검색 시스템 개발	▪한·영 병렬 말뭉치 응용 시스템 개발		2단계 말뭉치총합		▪말뭉치의 효율적인 활용을 극대화
						원시	943만	
						주석	233만	

③ 3단계(2004-2007) 추진 목표

<사업의 개요>

구분 \ 연도		2004	2005	2006	2007	합계		주안점
현대 국어 구어 전사 말뭉치	원시 말뭉치	39만	40만	42만	40만	161만	241만	• 다양한 말뭉치 확보 • 연구·배포 용 말뭉치 구성을 위한 통합 재정비 작업
	형태소 분석 말뭉치	18만	21만	20만	21만	80만		
한·영 병렬 말뭉치	원시 말뭉치	48만	47만	48만	47만	190만	288만	• 획득 자료의 특수성 극복 • 연구·배포 용 말뭉치 구성을 위한 통합 재정비 작업
	형태소 분석 말뭉치	19만	26만	27만	26만	98만		
한·일 병렬 말뭉치	원시 말뭉치	22만	23만	21만	23만	89만	145만	• 다양한 장르의 말뭉치 확보에 노력
	형태소 분석 말뭉치	12만	14만	16만	14만	56만		
북한 및 해외 한국어 말뭉치	원시 말뭉치	100만	100만	90만	100만	390만	546만	• 해외 한국어 자료의 균형성 고려
	형태소 분석 말뭉치	35만	39만	43만	39만	156만		
역사 자료 말뭉치	원시 말뭉치	70만	60만	60만	50만	240만	275만	• 15-19세기 자료와 개화기 자료를 적절한 비율로 구축
	형태소 분석 말뭉치	5만	10만	10만	10만	35만		
전문용어 말뭉치	원시 말뭉치	100만	100만	100만	100만	400만	400만	• 전문 분야의 언어 공학적 처리를 위한 기초 자료 구축
총합		468만	480만	477만	470만	1,895만	1,895만	
말뭉치의 가공 및 활용을 위한 연구 과제		음성-텍스트 통합 구어 말뭉치 구축을 위한 기초 연구				3단계 말뭉치총합		• 말뭉치의 활용을 극대화하기 위한 기초연구
						원시	1,470만	
						주석	425만	

3단계 1차년도에서는 그간 이루어진 구축 작업을 종합적으로 분석하고, 균형 말뭉치 구성을 목표로 작업을 진행할 것이다.

구어 전사 말뭉치는 3단계 이후 구성될 100~150만 어절 규모의 연구·배포용 통합 말뭉치가 유기적으로 구성될 수 있도록 원시 말뭉치의 구축 작업을 지속적으로 수행할 예정이다. 또한, 2002년에 마련된 구어 균형 말뭉치 구성 시안을 검토하여 수정·보완한 후, 이를 토대로 포괄적이고 다양한 유형의 자료를 개발·수집하고자 한다. 양적인 측면 못지않게 중요한 것이 배포용 말뭉치의 질적인 재정비 문제이다. 따라서 3단계에도 내적인 통일성 및 일관성 확보를 위한 수정, 보완 작업을 지속해 나갈 것이다. 또한 구어 말뭉치는 디지털화된 원자료와 텍스트 자료를 함께 제공하는 것이 중요하다. 음성 자료와 텍스트 자료를 매핑하는 방식으로 연구 배포용 말뭉치를 구축하고, 이를 검색 시스템 상에서 병렬적으로 검색 가능하도록 한다면 그 활용 범위를 극대화할 수 있을 것이다. 따라서 3단계의 이른 시기에 이에 대한 기초 연구가 진행되어, 이를 바탕으로 실제적인 작업이 수행될 필요가 있다.

병렬 말뭉치의 경우, 2단계까지 이루어진 연구 성과와 결과물을 토대로 구축하고 저작권 해결이 가능한 범위 내에서 말뭉치의 범용성 확보를 위해 노력한다. 특히 자연언어처리나 언어공학 분야에서의 병렬 말뭉치 활용도를 고려하여 말뭉치 구성을 설계한다. 그리고 지금까지 구축된 병렬 말뭉치의 활용도를 높이기 위한 여러 가지 방안을 모색한다.

역사 말뭉치는 3단계 사업에서 원시 말뭉치 240만 어절, 주석 말뭉치 35만 어절을 구축할 것이다. 원시 말뭉치는 15세기부터 시작하여 미입력 문헌을 입력하고 말뭉치 지도상 입력이 취약한 개화기 자료 입력에 중점을 둔다. 개화기 시대의 교과서, 기독교 자료, 신문, 사전 등을 중심으로 원시 말뭉치를 구축할 것이다. 또한 주석 말뭉치는 15세기 문헌의 주석과 개화기 문헌의 주석을 반 정도씩 할 것이다. 15세기 문헌의 주석이 끝나면 16세기 문헌 순으로 내려 올 것이며 개화기 문헌은 주로 신소설 쪽으로 초점을 맞춰 주석 말뭉치를 구축할 것이다.

전문용어 말뭉치는 2단계까지 '전문용어 정비' 분과에서 이루어 낸 결과물을 기초로 하여 추후 '전문용어 정비' 분과에서 용어 추출, 용어의 사용빈도 파악, 해당 분야 언어모델링 등 다양한 언어공학적 처리를 할 수 있도록 말뭉치를 구축한다. 또한 해당 분야의 용어학적 대표성을 가지면서 균형성 있는 말뭉치가 될 수 있도록 말뭉치 구성을 설계한다.

2. 2004년도 사업 목표

국어 특수자료 분야를 6개 영역으로 구분하고, 표준화된 대규모 말뭉치를 구축한다. 이 가운데 2004년도의 목표량은 다음과 같다.

2.1. 현대 국어 구어 전사 말뭉치 개발

분량 : 57만 어절

원시 말뭉치

39만 어절

형태소 분석 말뭉치

18만 어절

내용 : 말뭉치 기반 기술언어학, 사전편찬학, 담화 분석, 실험음성학 등 구어에 대한 언어학적 연구와 언어 교육, 언어 병리학, 자연언어 처리 분야, 음성공학 분야 등 기타 관련 분야로의 응용을 위한 다양한 구어 전사 말뭉치와 형태소 분석 말뭉치를 개발한다.

2.2. 한·영 병렬 말뭉치 개발

분량 : 67만 어절

원시 말뭉치(한·영)

48만 어절

형태소 분석 말뭉치

19만 어절

내용 : 기계 번역, 외국어로서의 한국어 교육, 한국인을 위한 외국어 교육, 기타 비교언어학 연구와 응용을 위한 한국어-영어 병렬 말뭉치와 형태소 분석 말뭉치를 개발한다.

2.3. 한·일 병렬 말뭉치 개발

분량 : 34만 어절

원시 말뭉치 22만 어절

형태소 분석 말뭉치 12만 어절

내용 : 기계 번역, 한국어 교육 등의 응용 분야와 문체 연구, 비교 언어학 연구 등의 연구를 위한 한국어-일본어 병렬 말뭉치와 2003년도의 기초 연구를 토대로 한 형태소 분석 말뭉치를 개발한다. 또한 2004년도는 전년도 형태소 말뭉치를 구축함에 있어서 유용한 정보를 제공할 수 있도록 연구한다.

2.4. 북한 및 해외 한국어 말뭉치 개발

분량 : 135만 어절

북한 및 해외 한국어 원시 말뭉치 100만 어절

형태소 분석 말뭉치 35만 어절

내용 : 남북한의 언어적 통합을 위한 연구와 정책 수립에 기초가 되는, 북한 지역의 현대 한국어 말뭉치와 해외 한국어 말뭉치를 구축하고, 이의 형태소 분석 말뭉치를 개발한다.

2.5. 역사 자료 말뭉치 개발

분량 : 원시 말뭉치 75만 어절

원시 말뭉치 70만 어절

형태소 분석 말뭉치 5만 어절

내용 : 국어의 발달 변천사를 밝히고, 국어 교육과 어문 생활의 발전을 위한 국어사 자료 말뭉치를 개발한다.

2.6. 전문용어 말뭉치 개발

분량 : 100만 어절

원시 말뭉치

100만 어절

내용 : 전문용어의 형태와 목록을 추출할 수 있도록, 균형성과 정보성이 고려된 전문분야 도서와 자료들을 선정하여 전문분야 말뭉치를 구축한다.

3. 연구 추진 방법

3.1. 현대 국어 구어 전사 말뭉치 개발

1) 구어 말뭉치의 설계와 구성

- 현대 국어 구어 전사 말뭉치의 2단계까지의 구축량은 원시 말뭉치 194만 어절, 형태 분석 말뭉치 50만 어절이다. 3단계에서의 구축 목표량은 원시 말뭉치 161만 어절, 형태 분석 말뭉치 80만 어절로, 이 중 3단계 1차년도인 2004년에는 원시 39만, 형태 분석 18만 어절이 구축되었다. 계획대로 구어 말뭉치 구축이 진행된다면, 원시 말뭉치는 총 355만 어절, 형태 분석 말뭉치는 130만 어절이 구축 완료될 것이다.
- 3단계 이후의 100만 어절의 연구·배포용 통합 말뭉치가 양적·질적인 면에서 대표성과 균형을 가질 수 있도록 포괄적이고 다양한 원시 말뭉치의 지속적인 구축을 위해 노력한다. 2004년 수정한 구어 균형 말뭉치 시안과 현재까지의 구축량은 다음과 같다.

구분			2003	2002	2001	2000	1999	1998	합계
화자수	텍스트 유형								
1 인 담 화	공직 (강의, 강연, 발표, 자기소개, 업무보 고, 업무 프리젠테 이션, 연설, 설교 등)	일 반	126,623	37,958	44,496	94,216	121,984	15,560	440,837
		방 송	0	10,840	3,713	0	84,993	155,377	254,923
	사적 (경험담 이야기하 기, 영화, 소설 등 의 줄거리 이야기 하기, 전화 음성 메 시지 등)		80,158	45,759	11,916	0	0	0	137,833
	합계(비율)		206,781 (42.02%)	94,557 (23.58%)	60,125 (10.93%)	94,216 (18.84%)	206,977 (41.77%)	170,937 (67.90%)	833,593 (30.98%)
2 인 이 상 의 담 화	공직 (토론, 상담, 회의, 구직 인터뷰, 면접, 수업대화, 구매대 화, 진료대화, 공공 기관에서의 대화 등)	일 반	34,879	0	33,489	39,502	125,556	0	233,426
		방 송	0	28,557	92,807	157,420	163,037	76,612	518,433
	사적 (일상대화, 식사대 화, 전화대화 등)		250,491	277,923	363,689	208,989	0	4,211	1,105,303
	합계(비율)		285,370 (57.98%)	306,480 (76.42%)	489,985 (89.07%)	405,911 (81.16%)	288,593 (58.23%)	80,823 (32.10%)	1,857,162 (69.02%)
기타			0	0	0	0	7552	0	7,552
합계			492,151	401,037	550,110	500,127	495,570	251,760	2,690,755

- 또한 구어 말뭉치의 활용 및 배포를 위한 기본적인 제약 조건이라 할 수 있는 저작권 승인 문제를 해결하고자, 2001년부터 발화자들로부터 자료 구축과 활용에 대한 '국가 언어 자원 구축 및 활용 승낙서'를 받고 있으며, 2003년 이후부터 저작권 해결 전담자를 두어 원시 말뭉치 신규 구축분에 대하여 70% 이상의 저작권 양의 각서 확보율을 유지할 수 있도록 최대한 노력하고

있다.

- 자료 사용 및 보관의 용이성과 정밀한 언어학적 연구 및 자연 언어 처리 분야에서의 직·간접적 이용을 위해 녹음 작업을 디지털화하여 텍스트 형태의 구어 전사 자료와 음성 파일을 함께 확보하고 있다.
- 지금까지 구축된 형태 분석 말뭉치의 장르별 구성비와 내용 및 화자 정보를 재검토하여 연구·배포용 구어 통합 말뭉치를 구성할 수 있는 방향으로 진행하고자 한다.

2) 말뭉치 규격

- 표준화 지침에 의거하여 부호화 방법과 문서 형식을 통일한다.
- 말뭉치의 효율적 관리와 이용에 필요한 헤더 정보와 마크업을 첨부한다. 헤더 정보와 마크업, 형태소 태그 모두 표준안을 설정하며, 국제 표준 규격을 최대한 반영하도록 한다.
- 품질 수준의 유지에 유의하여, 최종 오류율이 0.1% 이내가 되도록 한다.

3) 추진 방법

- 구어 말뭉치 구축을 위한 각종 영상·음성 자료의 녹취 및 수집
- 구어 말뭉치의 전사 및 주석 방법론 보완
- 지침에 따른 1, 2차 전사와 수작업 마크업
- 프로그램을 이용한 1차 형태 분석 태깅과 수작업을 통한 보완
- TEI 헤더와 기계적인 내용 태그 부착
- 원자료 편집
- 최종 점검과 오류 검증

3.2. 한·영 병렬 말뭉치 개발

1) 설계

- 한·영 병렬 말뭉치의 3단계 구축 목표량은 한·영 원시 말뭉치 190만 어절

한·영 형태소 분석 말뭉치 98만 어절이다. 2단계 3차년도까지 한·영 원시 355만 어절, 형태소 분석 52만 어절이 구축되어 있으며, 3단계 1차년도인 2004년에는 한·영 원시 48만, 형태소 분석 19만 어절을 구축한다.

- 말뭉치의 저작권 사용 승인을 받는 것을 전제로 하여 기계번역, 한국어와 영어의 대조 연구 및 언어 교육 등에 널리 응용할 수 있도록 말뭉치의 범용성과 응용상의 다양성을 기본적으로 고려하여 설계한다.
- 다음과 같은 병렬 말뭉치의 주된 이용 분야를 고려하여 각 분야에 맞는 말뭉치를 고루 구비할 수 있도록 설계한다.
 - 언어 교육 분야: 번역 및 해석 과정에서의 이해와 생산을 위한 번역 등가성, 표현과 용법상의 적합성이 고려된 텍스트가 우선되어야 한다. 따라서 이해하기 쉽고, 문법적 오류가 상대적으로 적은 텍스트, 실생활과 관련된 언어 표현, 교육용 어휘 사용과 교육용 문장 구조가 고려된 텍스트가 선정되어야 한다. 영어 교재의 경우, 번역의 방향성을 상징할 수 없는 대응 자료가 있을 수 있다는 점에서 다른 분야의 텍스트와 구별된다.
 - 공공 부문 분야: 실생활에 직접적으로 도움이 되는 텍스트의 선정이 요구된다. 공공 부문에서의 병렬 말뭉치 요구는 이해와 생산을 위한 번역 예시의 기능이 우선된다. 의료, 은행, 행정 분야 등의 신청 양식, 안내문, 공고문, 공문서 등 특정 유형의 텍스트가 선정될 수 있다.
 - 언어 공학 분야: 공학적 처리 우선성(편리성)이 고려된다. 기계 번역용 어휘부 구축을 위한 등가어 추출 시스템을 개발하고자 한다면 문장 구조가 복잡한 텍스트보다는 간결한 문장 구조의 텍스트가 선호된다. 정렬 프로그램 및 기계번역에 응용될 실험용 또는 정보 추출용 말뭉치로 사용하고자 한다면 앵커 포인트나 앵커 단어가 확실한, 직역 위주의 텍스트가 선택되어야 한다. 또한 선호되는 번역 분야가 우선적으로 고려된다면 기술적 텍스트, 매뉴얼 등을 선정할 수 있다.
 - 학술 연구 분야: 특정 텍스트 유형이나 장르의 선정 문제보다는 말뭉치의 전체 구조 설계에 영향을 받는다. 번역 연구나 대조 연구용 말뭉치는 대조되는 두 언어의 언어 현상과 구조의 비교를 위해 텍스트 선정과 말뭉치 규모에 대응성을 유지해야 한다.

<사업의 개요>

- 1998년 이후 2003년까지 구축된 한·영 병렬 말뭉치의 매체·장르별 분포를 보면, 주로 상상적 텍스트나 인문 사회 관련의 텍스트에 치중한 경향을 보인다. 3단계까지의 전체적인 병렬 말뭉치의 구성 및 향후 병렬말뭉치의 활용도 제고를 위해 기술적 텍스트나 제품 매뉴얼, 교재 등 언어교육과 언어공학 분야에서 널리 사용될 수 있는 영역의 텍스트를 중심으로 선정한다.
- 병렬 말뭉치의 경우 저작권 문제 해결이 가장 시급한 과제인 만큼 저작권 해결을 위한 다각도의 모색과 실험을 통해 원시 말뭉치 신규 구축분에 대하여 저작권 양의 각서 확보율을 높이기 위해 최대한 노력한다.
- 국가 간의 정치·문화·경제 등의 교류가 활발해지고 있는 오늘날 언어 간의 효율적인 정보 소통과 상호 문화 이해를 위해 통역·번역 등의 실용적인 측면이나 기계 번역, 언어의 대조 분석 연구라는 학문적인 측면에서 병렬 말뭉치는 매우 중요한 역할을 한다.

2) 추진 방법

- 설계 원칙에 따라 자료를 구성하고 저작권 확보를 위한 노력을 시도한다.
- 자료의 입력과 문장 단위의 정렬 문서를 생성한다.
- 형태소 분석 말뭉치에 대한 중의성 해결 작업과 형태소 분석 결과 검토 작업을 실시한다.
- 품질 수준의 유지에 유의하여, 최종 오류율이 0.1% 이내가 되도록 한다.

3.3. 한·일 병렬 말뭉치 개발

1) 설계

- 한·일어의 대조 연구와 기계 번역, 언어 교육 등에 널리 활용할 수 있도록 범용성을 갖도록 한다.
- 한·일/일·한 병렬 텍스트로 구성하며, 국제적 기준인 CES(Corpus Encoding Standard)의 규범과 기준을 바탕으로 한다.

- 현대어의 각 부분이 잘 나타나 있는 여러 장르와 종류의 자료를 대상으로 구축한다. 구축 대상 자료는 신문 잡지 기사와 실용문, 소설, 교과서, 사전 자료, 상품 매뉴얼, 외교문서(조약문), 교양 서적, 텔레비전(라디오) 대담, 시나리오 등이다.
- 한·일 대역과 일·한 대역문을 50%씩 구축한다.
- 자료의 선정은 연구진 내부의 자료 평가 심의를 거쳐 추천된 것을 대상으로 한다.
- 저작권 해결을 위한 여러 가지 방안을 검토하여 원시 말뭉치 신규 구축분에 대하여 저작권 양의 각서 확보율을 70% 이상으로 높이기 위해 노력한다.

2) 추진 방법

- 설계 원칙에 따라 자료를 구성하고 저작권 확보를 위한 노력을 시도한다.
- 자료의 입력과 정렬 문서를 생성한다.
- 최종 오류율이 0.1% 이내가 되도록 한다.

3.4. 북한 및 해외 한국어 말뭉치 개발

1) 설계

- 남북 분단 이후 현재까지 북한에서 쓰인 국어 텍스트와 중국 연변 지역을 중심으로 한 해외 한국어 텍스트를 말뭉치로 구축한다.
- 대상 자료는 교과서, 신문, 잡지, 문예지, 기타 전공 서적류 등으로 하되, 북한 사회의 언어적 실상이 충실하게 포착될 수 있도록 한다.
- 원시 말뭉치 구축 목록은 국립국어원에서 추진하고 있는 북한 자료 구축 목록과 중복되지 않도록 사전 협의하여 선정한다
- 북한어의 이질화 현실을 폭넓게 파악하고, 통일에 대비한 언어통합 추진에 활용할 수 있도록 구성한다.
- 3단계 이후 100만 어절의 (연구·배포용의) 북한 및 해외 한국어 균형 말뭉

치 구성을 목표로 다양한 말뭉치 확보를 위해 노력한다.

2) 추진 방법

- 표준화 지침에 의거하여 부호와 문서 형식을 통일한다.
- 말뭉치의 효율적 관리와 이용에 필요한 헤더 정보와 마크업을 첨부한다.
- 형태소 분석 지침에 따라 형태소 분석 말뭉치를 구축한다.
- 품질 수준의 유지에 유의하여, 최종 오류율이 0.1% 이내가 되도록 한다.

3.5. 역사 자료 말뭉치 개발

1) 설계

- 사이버 한글 박물관 디지털화 자료 중 미입력분을 개발한다.
- 개화기 미입력 자료와 기독교 및 필사본 자료를 구축한다.
- 대상 자료를 구축의 난이도별로 적절하게 구분하여 효율적인 구축 과정을 설계한다.

2) 추진 방법

- 데이터베이스 구축 대상 자료 및 문헌을 선정한다.
- 역사 자료 구축에 필요한 말뭉치 표준화 및 입력상의 지침을 개발한다.
- 원전의 서지적 특징과 표기에 충실하게 전자화하고, 표준적 헤더 정보와 마크업을 첨부한다.

3.6. 전문용어 말뭉치 개발

1) 설계

- KORTERM의 전문용어 정비 분과와 협조 하에 전공서적 및 학술지 등을 수집하여 '전문용어 표준화'를 위한 기초 자료인 전문분야 말뭉치 100만 어절을 구축한다.
- 대상 자료는 전문용어의 정비 분과에서 추진해오던 구축 목록을 참조하여 해당 학술학회와 협의하여 최대한 활용할 수 있도록 한다.
- 국내 각 전문분야의 학회와 협의하여 전문용어 목록 데이터를 작성한다.
- 대상 원시 말뭉치를 구축하여 KORTERM의 '전문용어 정비'에서 활용할 수 있도록 한다. 말뭉치 입력에 대한 것들은 KORTERM과 협의하여 사업을 원활히 수행할 수 있도록 한다.

2) 추진 방법

- KORTERM과의 협력관계가 있기 때문에 기존의 '전문용어정비' 분과에서 말뭉치를 활용할 수 있도록 KORTERM의 입력방식을 따른다.
- 말뭉치의 효율적 관리와 이용에 필요한 헤더 정보와 마크업을 첨부한다.
- 단 마크업은 KORTERM 입력방식을 따르고, 헤더 정보와 문서 형식은 세종 말뭉치 전체에 대한 일관성 때문에 '세종 말뭉치 표준화 지침'(1998)에 의거하여 통일한다.
- 품질 수준의 유지에 유의하여, 최종 오류율이 0.1% 이내가 되도록 한다.

제 1 부

원시 말뭉치의 구축

여 백

제 1 장

현대 국어 구어 전사 말뭉치 개발

연구책임자 : 김 향 희

여 백

제 1 장

현대 국어 구어 전사 말뭉치 개발

1. 연구의 필요성 및 목표

1.1. 연구의 필요성

구어 말뭉치는 문어 말뭉치와 함께, 인간 의사소통행위 연구를 위한 기본적인 언어 자료로서 중요성을 지닌다. 구어 자료의 구축은 문어와 균형을 이루는 언어 자료로서의 가치 이외에도, 자연스러운 상황에서의 실제적인 언어 사용 양태를 파악할 수 있다는 점과, 시대적·사회적인 흐름에 따라 변화하는 언어의 역동성을 사실적이고 구체적인 자료를 통해 파악할 수 있다는 점 등에서 그 의의를 찾을 수 있다. 이 때문에 국어학적 연구 외에도 한국어 교육 및 사전 편찬학, 그리고 전산 언어학, 음성 공학 분야에 이르기까지 다양한 분야에서 적극적인 관심의 대상이 되고 있는 것이다.

더욱이 한국인의 언어생활의 역동적 변화를 그대로 담고 있는 현대 한국어 구어 말뭉치는, 한국말이 음성 언어적 측면에서 발현되는 모습을 연구하고, 정보 처리 및 IT 산업 분야에서 자연 언어 검색 및 처리를 효율적으로 응용할 수 있도록 하는 데에 문어 말뭉치보다 훨씬 더 큰 가치를 지닌다. 특히 최근의 음성 언어 정보 처리를 위한 근간 자료로서 구어 말뭉치에 대한 각 연구 기관 및 기업체의 관

심과 문의를 생각한다면, 이러한 IT 산업 분야에서의 기술 개발에서 발화 상황이나 화자의 성별, 연령대 등의 정보가 고려되어 구축된 구어 말뭉치의 필요성과 역할은 그 중요성을 강조해도 지나치지 않겠다 하겠다.

이와 같이, 구어 말뭉치의 필요성에 대한 인식은 많이 진전되었다 하겠으나, 사실상 아무런 제약이나 통제가 없는 상황에서 발화되는 담화를 녹취하여 전사한 진정한 의미의 구어 자료를 구축하는 데에는 많은 노력과 시간이 필요하다. 서구 선진국에서는 1990년대 이후 국가와 수많은 정보 산업 관련 기업체 및 기술 개발자들을 중심으로 문어 말뭉치 및 대규모의 구어 말뭉치가 포함된 말뭉치 구축 사업들이 진행되어 왔으며, 그 규모도 대부분 1천만 어절 이상에 달한다. 이에 비하여 국내에서 현재까지 구축되고 있는 구어 말뭉치의 경우, 2004년까지 구축된 구어 원시 말뭉치가 333만 어절, 형태 주석 말뭉치가 68만 어절에 불과하다. 그러므로, 유의미한 대규모 국가 언어 자료의 개발을 위해서는 한국어를 대상으로 한 구어 말뭉치 구축이 지속적으로 이루어져야 할 필요가 있다.

또한 구어 말뭉치는 디지털화된 원 자료와 텍스트 자료를 함께 제공하는 것이 중요하다는 점에 대해 많은 연구자들이 인식을 같이 하고 있다. 음성 자료와 텍스트 자료를 매핑하는 방식으로 연구 배포용 말뭉치를 구축하고, 이를 검색 시스템 상에서 병렬적으로 검색 가능하도록 하여 그 활용 범위를 극대화하는 방안에 대해서는 좀더 구체적이고 지속적인 논의가 필요하다.

1.2. 연구 목표

현대국어 구어 전사 말뭉치 개발은 한국어의 실제 사용 양상에 대한 정보를 자료로 축적하여 사실적인 언어 사용에 기반한 한국어 연구의 기본 자료를 확보하기 위한 국가적 규모의 사업이다. 지속적으로 구축되어 온 구어 원시 말뭉치는, 이후 가공 작업을 통하여 자연언어처리 분야에 문맥 및 담화 처리를 위한 연구 자료로서 사용될 수 있다.

세종 계획 제 3단계 1차년도 2004년도 구어 전사 말뭉치는 3단계 이후 구성될 100만 어절 규모의 연구·배포용 통합 말뭉치가 유기적으로 구성될 수 있도록 원시 말뭉치의 구축 작업을 지속적으로 수행하는 것이다. 양적인 측면 못지 않게 중요한

것이 배포용 말뭉치의 질적인 재정비 문제이므로, 3단계에서도 내적인 통일성 및 일관성 확보를 위한 기 구축분 말뭉치의 수정 및 보완 작업을 지속해 나간다.

2. 구어 전사 원시 말뭉치 개발

2.1. 말뭉치 목록

파일명	표기 언어	제목	장르	생산 연도	어절수	저작권
7CM00001	한글	강의_공대수업	기타 녹음>전사(공학)	2004	5,090	○
7CM00002	한글	강의_방송음향	기타 녹음>전사(총류)	2004	2,979	○
7CM00003	한글	강의_윈도우	기타 녹음>전사(공학)	2004	3,493	○
7CM00004	한글	강의_인터넷	기타 녹음>전사(공학)	2004	5,434	○
7CM00005	한글	독백_군대	기타 녹음>전사(총류)	2004	4,177	○
7CM00006	한글	방송_스포츠 중계(축구)	기타 녹음>전사(총류)	2004	13,895	○
7CM00007	한글	수업대화_과외 수업#1	기타 녹음>전사(총류)	2004	7,042	○
7CM00008	한글	수업대화_과외 수업#2	기타 녹음>전사(총류)	2004	5,669	○
7CM00009	한글	수업대화_과외 수업#3	기타 녹음>전사(총류)	2004	13,124	○
7CM00010	한글	수업대화_오디오EQ	기타 녹음>전사(예술)	2004	7,670	○
7CM00011	한글	수업대화_콘솔#1	기타 녹음>전사(예술)	2004	6,629	○
7CM00012	한글	수업대화_콘솔#2	기타 녹음>전사(예술)	2004	11,216	○
7CM00013	한글	일상대화_가족사	기타 녹음>전사(총류)	2004	4,234	○
7CM00014	한글	일상대화_게임	기타 녹음>전사(총류)	2004	4,900	○
7CM00015	한글	일상대화_결혼#1	기타 녹음>전사(총류)	2004	5,283	○
7CM00016	한글	일상대화_결혼#2	기타 녹음>전사(총류)	2004	14,713	○
7CM00017	한글	일상대화_고등학교동창	기타 녹음>전사(총류)	2004	8,277	○
7CM00018	한글	일상대화_교보문고커피숍	기타 녹음>전사(총류)	2004	7,300	○

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

7CM00019	한글	일상대화_민속주점#2	기타 녹음>전사(총류)	2004	12,560	○
7CM00020	한글	일상대화_생활	기타 녹음>전사(총류)	2004	5,507	△
7CM00021	한글	일상대화_여고동창모임	기타 녹음>전사(총류)	2004	5,882	○
7CM00022	한글	일상대화_연애이야기	기타 녹음>전사(총류)	2004	5,845	○
7CM00023	한글	일상대화_일상#1	기타 녹음>전사(총류)	2004	6,836	○
7CM00024	한글	일상대화_일상#4	기타 녹음>전사(총류)	2004	8,341	○
7CM00025	한글	일상대화_저녁 식사#1	기타 녹음>전사(총류)	2004	4,735	○
7CM00026	한글	일상대화_저녁 식사#2	기타 녹음>전사(총류)	2004	8,253	○
7CM00027	한글	일상대화_저녁식사#3	기타 녹음>전사(총류)	2004	2,365	○
7CM00028	한글	일상대화_점심 식사	기타 녹음>전사(총류)	2004	4,674	○
7CM00029	한글	일상대화_주점 대화#1	기타 녹음>전사(총류)	2004	3,878	○
7CM00030	한글	일상대화_초등학교동창	기타 녹음>전사(총류)	2004	10,276	○
7CM00031	한글	일상대화_친구#1	기타 녹음>전사(총류)	2004	3,152	○
7CM00032	한글	일상대화_친구와의 대화#1	기타 녹음>전사(총류)	2004	962	○
7CM00033	한글	일상대화_친구와의 대화#3	기타 녹음>전사(총류)	2004	8,359	○
7CM00034	한글	일상대화_친구와의 대화#4	기타 녹음>전사(총류)	2004	3,083	○
7CM00035	한글	일상대화_친구와의 대화#5	기타 녹음>전사(총류)	2004	9,776	○
7CM00036	한글	일상대화_친척과의 대화	기타 녹음>전사(총류)	2004	3,344	○
7CM00037	한글	일상대화_커피숍	기타 녹음>전사(총류)	2004	12,322	○
7CM00038	한글	일상대화_휴식시간	기타 녹음>전사(총류)	2004	2,090	○
7CM00039	한글	전화대화_전화 통화	기타 녹음>전사(총류)	2004	368	○
7CM00040	한글	주제대화_녹음기 사용법	기타 녹음>전사(총류)	2004	196	○
7CM00041	한글	주제대화_농구	기타 녹음>전사(총류)	2004	581	○
7CM00042	한글	주제대화_대학진학	기타 녹음>전사(총류)	2004	9,323	○
7CM00043	한글	주제대화_보고서	기타 녹음>전사(총류)	2004	3,392	○
7CM00044	한글	주제대화_취업	기타 녹음>전사(총류)	2004	3,936	○
7CM00045	한글	진료 대화_식이요법	기타 녹음>전사(의학)	2004	2,527	○
7CM00046	한글	토론_고등학생 토론#1	기타 녹음>전사(사회)	2004	10,192	△
7CM00047	한글	토론_고등학생 토론#2	기타 녹음>전사(사회)	2004	10,379	△

제1장 현대국어 구어 전사 말뭉치 개발

7CM00048	한글	토론_고등학생 토론#3	기타 녹음>전사(사회)	2004	10,443	△
7CM00049	한글	토론_고등학생 토론#4	기타 녹음>전사(사회)	2004	11,480	△
7CM00050	한글	토론_고등학생 토론#5-1	기타 녹음>전사(사회)	2004	4,379	△
7CM00051	한글	토론_고등학생 토론#5-2	기타 녹음>전사(사회)	2004	5,805	△
7CM00052	한글	토론_고등학생 토론#6	기타 녹음>전사(사회)	2004	8,844	△
7CM00053	한글	토론_사회학이론 세미나#1,2	기타 녹음>전사(사회)	2004	17,711	△
7CM00054	한글	토론_세계화세미나#1,2	기타 녹음>전사(사회)	2004	14,021	○
7CM00055	한글	회의_총학생회전체회의	기타 녹음>전사(총류)	2004	9,216	○
7CM00056	한글	회의_학원강사	기타 녹음>전사(총류)	2004	17,726	△
TOTAL					393,884	

<<2003년 원시 말뭉치 목록>>

2.2. 저작권 해결 현황

구어 전사 말뭉치의 원활한 보급과 활용을 위해서는 기본적으로 저작권 문제가 해결되어야 한다. 이를 위해, 2001년부터 현재까지 발화자들로부터 자료 구축과 활용에 대한 ‘국가 언어 자원 구축 및 활용 승낙서’를 받고 있으며, 2003년부터는 저작권 해결 전담자를 두어 원시 말뭉치 구축분에 대하여 저작권 양의 각서를 확보하고자 노력해 오고 있다. 올해 구축분의 경우 전체 발화자 중 5.12%를 제외한 94.88%의 발화자들로부터 저작권 양의 각서를 확보한 상태이다.

위의 원시 말뭉치 목록에서 저작권이 ○로 표시된 것은 저작권이 해결되었다는 의미이고, △으로 되어 있는 것은 각 파일의 담화 참여자 중 단 한 명이라도 저작권 양의 각서를 확보하지 못했다는 의미이다. 식당 같은 장소에서 종업원이 한 마디 정도 발화한 경우처럼 예상치 못한 화자가 발화한 경우나, 담화 참여자 본인이 저작권 양의 각서에 서명하지 않아 이를 확보하지 못한 경우이다. 그러나 이때 이들이 담화에 참여한 비율은 전체적인 담화 상황 중 그 비중이 낮은 편인 경우가

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

대부분이기 때문에, 위의 파일들이 연구·배포용 통합 말뭉치에 포함된다고 해서 저작권이 문제가 될 수준은 아니라 하겠다.

2.3. 말뭉치 구축에서의 고려 사항

2.3.1. 균형 말뭉치 설계 및 구성

우선, 2003년까지 구축된 세종 구어 전사 말뭉치는 원시 294만 어절, 형태 주석 말뭉치 50만 어절로 이를 도표로 보이면 다음과 같다.

제1장 현대국어 구어 전사 말뭉치 개발

구분			2003	2002	2001	2000	1999	1998	합계
화자수	텍스트 유형								
1 인 담 화	공적 (강의, 강연, 발표, 자기소개, 업무보 고, 업무 프리젠테 이션, 연설, 설교 등)	일 반	126,623	37,958	44,496	94,216	121,984	15,560	440,837
		방 송	0	10,840	3,713	0	84,993	155,377	254,923
	사적 (경험담 이야기하 기, 영화, 소설 등 의 줄거리 이야기 하기, 전화 음성 메 시지 등)		80,158	45,759	11,916	0	0	0	137,833
	합계(비율)		206,781 (42.02%)	94,557 (23.58%)	60,125 (10.93%)	94,216 (18.84%)	206,977 (41.77%)	170,937 (67.90%)	833,593 (30.98%)
2 인 이 상 의 담 화	공적 (토론, 상담, 회의, 구직 인터뷰, 면접, 수업대화, 구매대 화, 진료대화, 공공 기관에서의 대화 등)	일 반	34,879	0	33,489	39,502	125,556	0	233,426
		방 송	0	28,557	92,807	157,420	163,037	76,612	518,433
	사적 (일상대화, 식사대 화, 전화대화 등)		250,491	277,923	363,689	208,989	0	4,211	1,105,303
	합계(비율)		285,370 (57.98%)	306,480 (76.42%)	489,985 (89.07%)	405,911 (81.16%)	288,593 (58.23%)	80,823 (32.10%)	1,857,162 (69.02%)
기타			0	0	0	0	7552	0	7,552
합계			492,151	401,037	550,110	500,127	495,570	251,760	2,690,755

<<2003년도까지의 원시 말뭉치 기 구축분 구성>>

현대 구어 전사 말뭉치의 3단계 전체 구축 목표량은 원시 말뭉치 161만 어절, 형태 주석 말뭉치 80만 어절로, 이중 올해 2004년도 원시 말뭉치 구축량은 39만 어절, 형태 주석 말뭉치는 18만 어절이다.

2004년 구어 전사 원시 말뭉치 구축에서는 2002년도에 마련된 구어 균형 말뭉치 시안을 보완하고 이에 근거해 다양한 유형의 구어 말뭉치를 개발·수집하는 것을 목표로, 일상생활의 주제에 초점을 맞춘 자연 발화를 보강한 39만 어절의 구어 전사 말뭉치를 구축하였다. 2004년 구축된 원시 말뭉치 구성은 다음과 같다.

<<2004년 원시 말뭉치 구성>>

화자수	구분	장르	어절수	비율
1인 담화	공적	강의	16,996	4.31%
	사적	사적 독백	4,177	1.06%
2인 이상의 담화	공적	토론	93,254	23.68%
		수업 대화	51,350	13.04%
		진료 대화	2,527	0.64%
		방송	13,895	3.53%
		회의	26,942	6.84%
	사적	일상 대화	166,947	42.38%
		전화 대화	368	0.09%
		주제 대화	17,428	4.42%
		합계	393,884	100.00%

2004년 구축한 원시 말뭉치에서는 기 구축분에 대한 정보를 데이터베이스화하여, 3단계 이후 배포될 연구·배포용 말뭉치의 구성을 위해, 미리 준비된 스크립트가 없는 다양한 발화 상황 및 다양한 화자, 그리고 진료 대화와 같이 극히 녹취하기 힘든 자료 등을 포함시키고자 노력하였다. 위의 표는 2004년에 국한된 자료로서, 다만, 1인 담화 중 사적 독백의 비율 및 2인 이상의 담화 중 공적 담화의 진료 대화나, 사적 담화의 전화 대화의 비율이 상대적으로 낮은 편인데, 이는 전체적인 말뭉치 지도상에서 균형을 맞추어 가면 되는 것이며, 앞으로 남은 사업 기간 동안 구축될 전사 말뭉치를 통해 보강해 나가야 할 부분이다.

구어 말뭉치의 경우는 문어 자료와는 다르게 화자의 수·연령·성별 및 담화 장면, 담화 목적 등에 따라서 문어 말뭉치에서는 고려되지 않거나 파악되지 않는 다양한 변이형을 포함하게 된다. 그러므로, 구어 전사 말뭉치 구축 시에는 다양한 담화 상황과 화자별 특성이 고려된 말뭉치 구축이 이루어져야 함은 주지의 사실이다. 특히 말뭉치의 균형성 문제에 대해서는 많은 논의가 이루어져 왔는데, 이상적인 구어 말뭉치는 담화 상황의 특성 및 대화 참여자가 1인이나 혹은 2인 이상이나에 따른 구분, 화자의 성별·연령·출신 지역 혹은 장기 거주 지역·직업·대화 참여자들 사이의 관계 등을 세밀히 고려한 구축이 이루어지는 것이 바람직하다.

- 참여자수- 독백
 - 대화 : 2인 대화/ 3인 대화/ 4인 이상의 대화 등
- 연령별 - 10대/ 20대/ 30대/ 40대/ 50대/ 60대 이후의 대화
- 성별 - 동성끼리의 대화/ 이성끼리의 대화
- 지역별 - 서울·경기도/ 강원도/ 충청도/ 경상도/ 전라도/ 제주도
- 직업별 - 학생/ 직장인/ 주부 등

위와 같은 특성을 모두 제대로 반영하는 것은 말뭉치를 구성하거나 수집할 때 쉽지 않은 일이다. 그러나, 담화의 유형별 특징이나 화자의 연령별·성별·직업별·출신 지역별·계층별 언어 사용의 차이 등 특정 변인에 초점을 맞춰 연구하기 위해서는 위의 요소들을 골고루 충족하는 자료가 일정 비율이 포함되어야 할 필요가 있다.

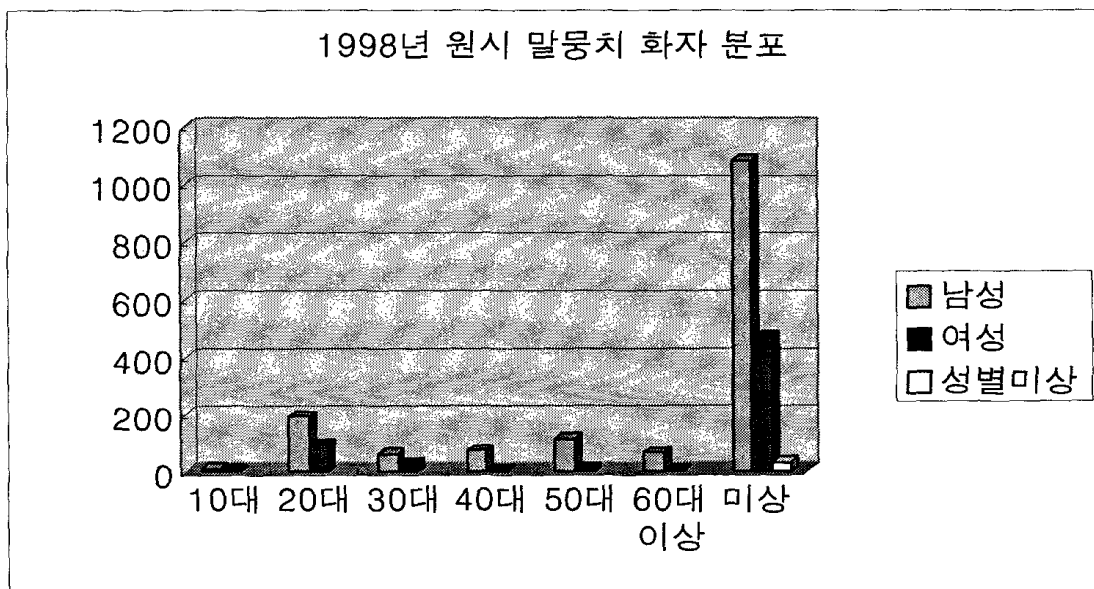
1인 담화 및 2인 이상의 담화 구분은 구어 자료 구축 시 지속적인 고려의 대상이 되어 왔으며, 이미 2002년 구축 시안에서 제시된 바와 같이 화자의 성별 및 연령별 비율에 따른 말뭉치 구축은 담화 자료 수집의 용이성을 고려하여 20대와 30대 자료를 중심으로 구축하되, 10대 화자 및 40대·50대·60대 이후 화자의 자료를 도외시하지 않는 방향에서 진행하고자 노력하고 있다.

구어 말뭉치 구축 시 이러한 다양한 변인들이 고려된 자료가 균형 있게 구축되고 활용되도록 하고자, 그동안 구축해 온 파일들에 대한 통계 자료를 지속적으로 축적해 오고 있다. 즉, 구어 말뭉치 목록과 발화 유형 분류 및 연령별·성별 화자

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

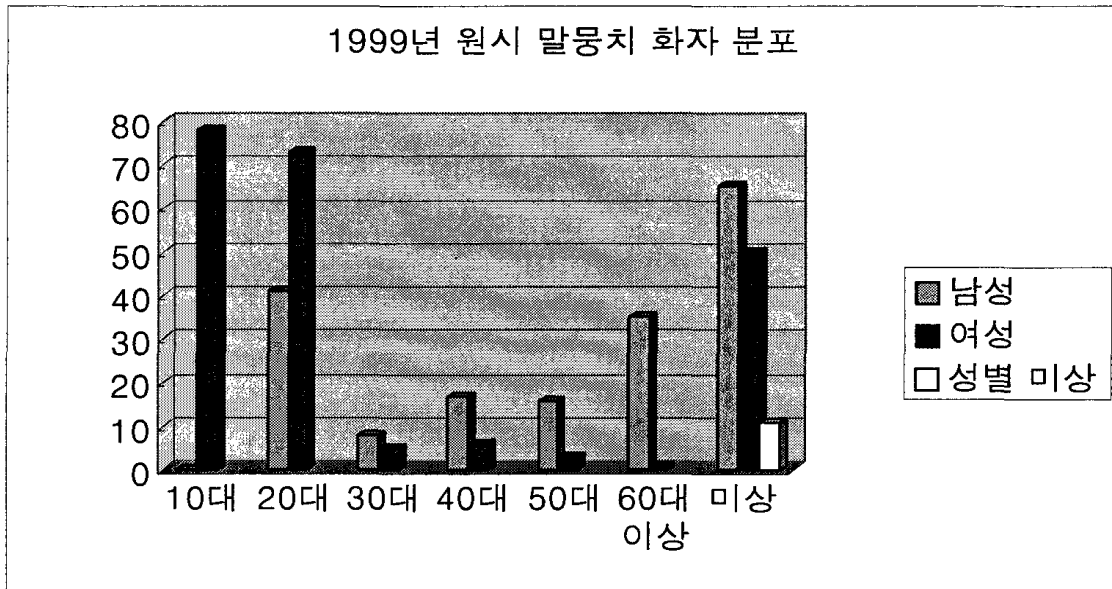
정보에 대한 누적 통계 등이 그것인데, 이는 100만 어절 규모의 연구·배포용 통합 말뭉치 구성을 위한 근간 자료로서 활용될 것이다.

여기서는 화자별 특성에 초점을 맞추어 지금까지 구축되어 온 구어 원시 말뭉치의 특성을 통계적으로 살펴보기로 한다. 지난 1998년부터 구축되어 온 구어 전사 원시 말뭉치를 올해 2004년도 구축분까지를 연도별로 살펴본 후, 98년부터 올해 2004년까지 구축된 전체 구어 원시 말뭉치의 누적 정보를 화자의 성별 및 연령대별 분포도로 다음과 같이 나타내었다.

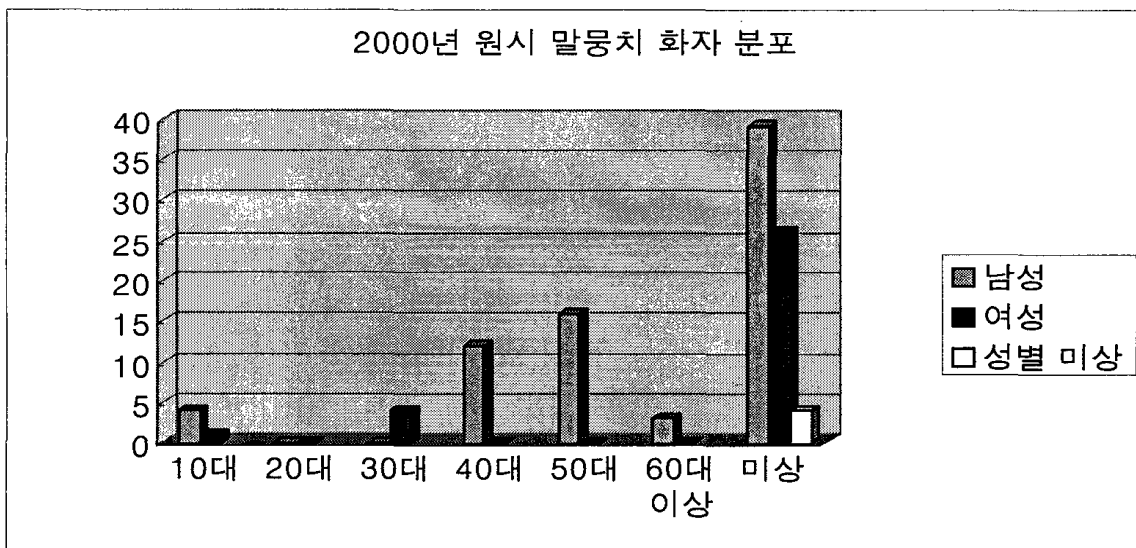


<<1998년 원시 말뭉치의 화자 분포>>

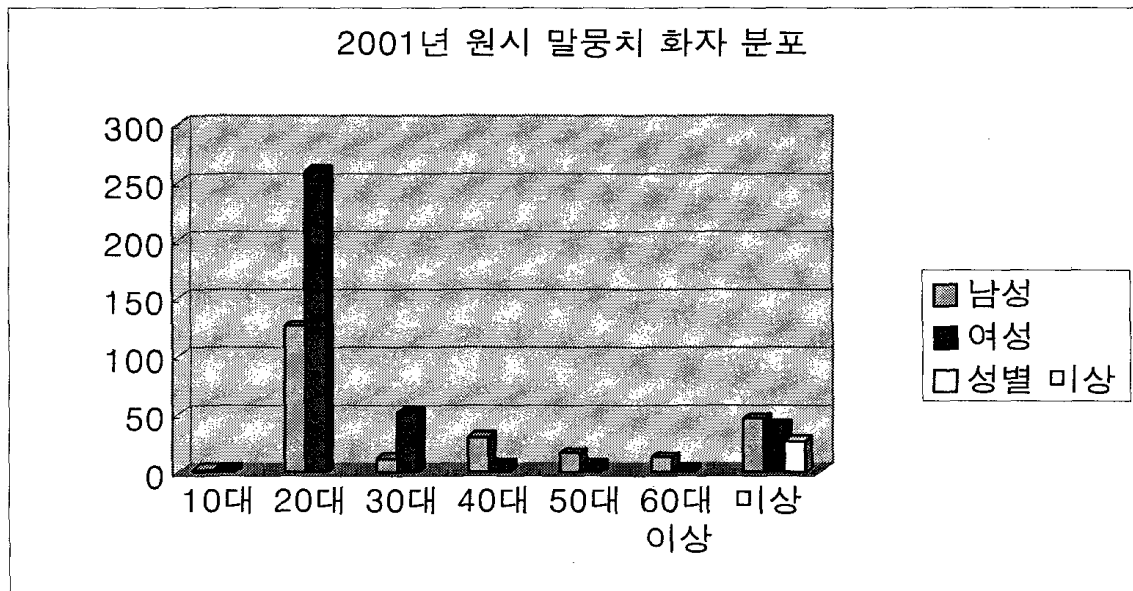
제1장 현대국어 구어 전사 말뭉치 개발



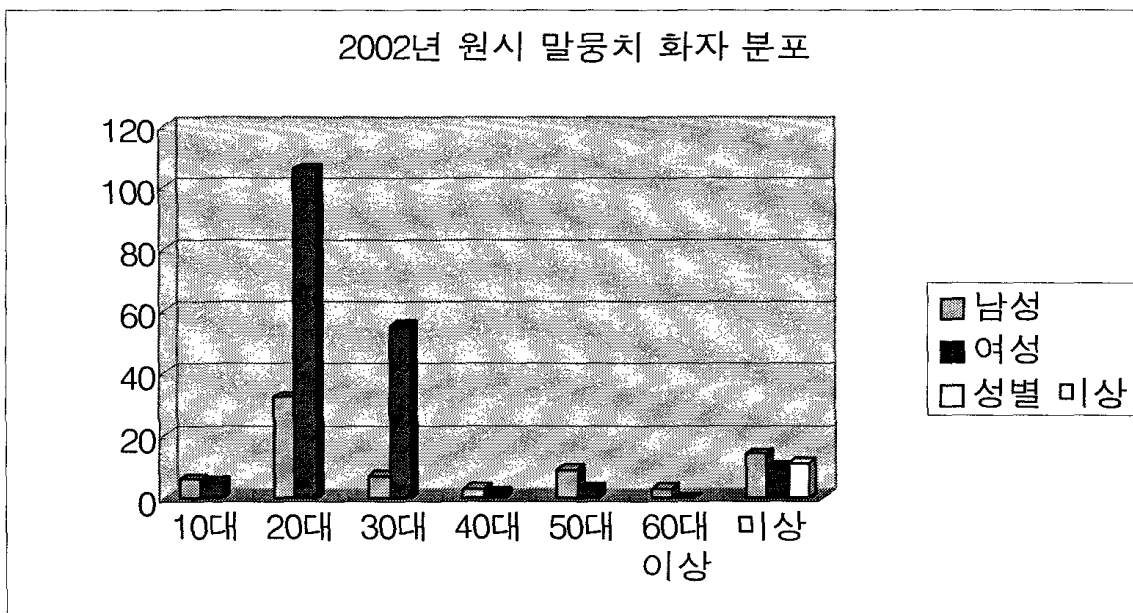
<<1999년 원시 말뭉치의 화자 분포>>



<<2000년 원시 말뭉치의 화자 분포>>

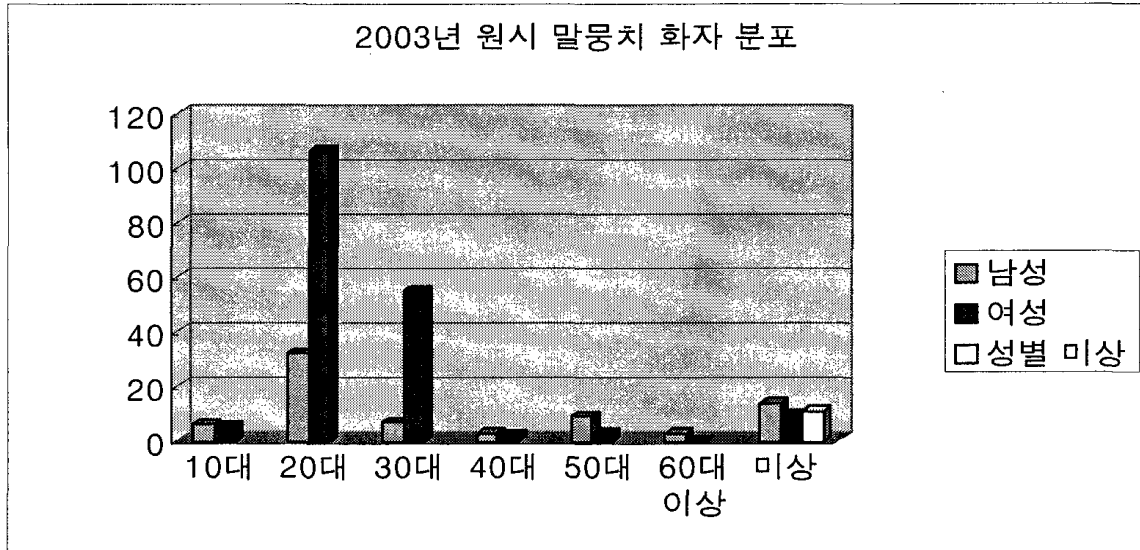


<<2001년 원시 말뭉치의 화자 분포>>

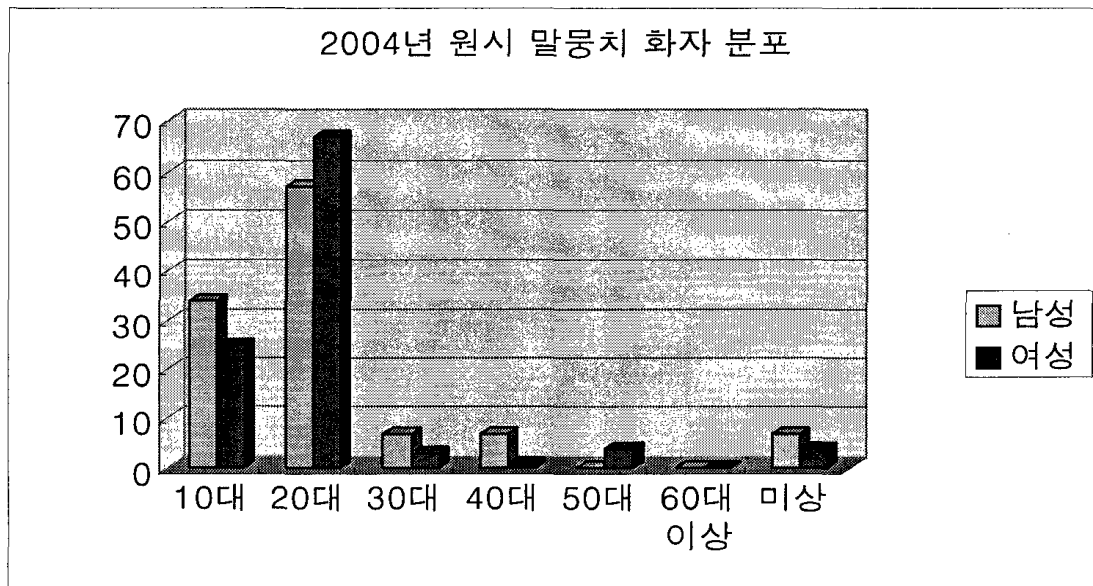


<<2002년 원시 말뭉치의 화자 분포>>

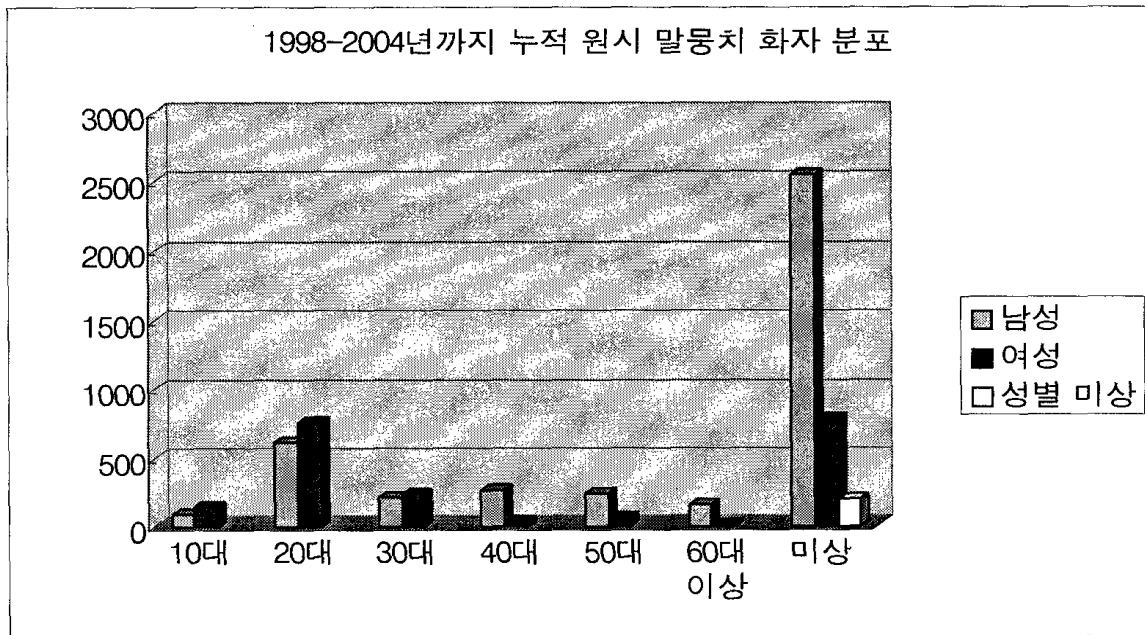
제1장 현대국어 구어 전사 말뭉치 개발



<<2003년 원시 말뭉치의 화자 분포>>



<<2004년 원시 말뭉치의 화자 분포>>



<<1998년부터 2004년 구축분까지의 원시 말뭉치 화자 분포>>

1998년부터 2004년 현재까지 구축된 전체 원시 말뭉치의 화자는 총 6337명으로 이 중 남성은 4104명, 여성은 2030명, 그리고 나머지 성별을 알 수 없는 경우는 202명으로, 전체 대비 남성 화자의 비율은 64.77%, 여성 화자는 32.04%, 성별을 알 수 없는 경우는 3.19%의 분포를 보인다.

제1장 현대국어 구어 전사 말뭉치 개발

1998-2004년	10대	20대	30대	40대	50대	60대 이상	연령대 모름	전체 합계(명)
남성	88	612	208	262	231	157	2547	4105
여성	148	752	234	42	59	13	782	2030
성별 미상							202	202
합계	236	1364	442	304	290	170	3,531	6337

<<1998년부터 2004년 구축분까지의 성별·연령별 누적 통계>>

연령별로는 10대가 3.72%, 20대가 21.52%, 30대가 6.97%, 40대가 4.80%, 50대가 4.58%, 그리고 60대 이상이 2.68%, 기타 연령을 알 수 없는 경우가 52.53%로 나타났다.

성별로는 남성 화자가 6337명 중 4105명으로 전체 화자의 64.77%에 해당되며, 여성 화자가 2030명으로 전체 화자의 32.04%, 그리고 성별이 파악되지 않은 경우가 202명으로 전체 화자 중 3.19%를 차지하여, 남성 화자의 수가 좀 많은 편으로 조사되었다.

연도별로 그래프를 보면, 1998년에 구축된 원시 말뭉치는 화자의 성별 및 연령별 분포가 비교적 고르게 나타나고 있다.

1999년 구축 말뭉치의 경우 10대·20대 여성 화자와 60대 남성 화자의 녹음분이 상대적으로 많은데, 전체 화자수가 409명이므로 이중 10대 여성 화자와 20대 여성 화자 및 60대 남성 화자가 차지하는 비율은 각각 19.07%, 17.84%, 8.56%로 특정 연령층에 크게 편중되어 있지 않다 하겠다.

2000년에 구축된 원시 말뭉치의 경우는 40대·50대 남성 화자의 비율이 각각

11.01%, 14.68%로 높은 편이다. 그러나 2000년 원시 말뭉치에는 남성 화자 연령 미상이 35.8%, 여성 화자 연령 미상이 23.9%, 성별·나이 미상이 3.7%로, 화자나 성별 미상인 경우가 전체 화자 구성의 63.3%로 높은 편이었다.

자연 발화를 중점적으로 구축하기 시작한 2001년 이후 자료 중, 2003년까지 구축된 원시 말뭉치는 20대 화자의 비율이 높고, 2004년 올해 구축분의 경우는 10대와 20대 화자 비율이 높은 편이다.

즉, 2001년 이후 자료에서는 20-30대 화자의 비중이 상대적으로 높게 나타나고 있음을 알 수 있는데, 이것은 자료 수집의 용이성 측면도 있다 하겠으나, 40대 이후의 연령층에 비해 신조어 및 유행어의 수용과 사용 등, 언어적 역동성과 생산성에 비교적 민감하게 반응하는 점도 고려되었다 할 수 있다. 20-30대 화자는 또한 10대 화자와 비교해 보면 사회에서 언어적 영향력을 행사할 수 있는 가능성이 더 높기 때문에 이들의 발화 수집이 최근 2-3년간 많이 이루어졌던 것이 사실이다. 그러나 말뭉치가 균형성을 갖추도록 하기 위해서는 지금까지는 다소 미흡했던 40대 이후 화자의 언어 자료가 좀더 확충되어야 할 것이다.

기존에 구축된 방송 자료에는 이러한 연령대별 불균형은 적게 나타나지만, 이들 초기 구축분에는 성별이나 나이 미상의 화자 비율이 98년에 69.8%, 99년에는 30.81%, 2000년에는 63.30%로 좀 높은 편이다. 이는 준 구어 자료에 해당하는 방송 자료가 다수 포함되어 있어 저작권 양의 각서를 확보하지 못했기 때문으로 보이는데, 이러한 성별 및 나이 미상 화자의 수치는 이후 점점 감소하여 2003년에는 13.15%, 2004년 올해 구축분의 경우는 5.12%로 조사되었다. 즉, 말뭉치를 초기 구축분에서는 다수의 방송 자료를 포함하고 있었기 때문에, 성별 혹은 연령 미상의 화자 수가 불가피하게 많을 수 밖에 없었으나, 자연 발화를 중점적으로 구축하기 시작한 2001년 이후에는 화자 정보가 불분명한 경우는 점차 감소하고 있음을 알 수 있다.

그래프를 통해 살펴볼 수 있는 특징 중에 하나는 자연 발화를 많이 다루기 시작한 2001년 자료부터는 20-30대 화자의 비율이 상대적으로 높다는 것인데, 남녀의

성비 불균형과, 발화자의 나이대가 20대에 다소 편중되어 분포된 점은, 차기 사업 추진에서 보완해 나가야 할 부분으로 보인다. 즉, 연령 측면에서 40대·50대·60대 화자의 구어 말뭉치 확충에 노력을 기울여야 할 것이며, 남녀 성비가 일정 정도 균등할 수 있도록 하는 자료의 확보가 이루어져야 함을 알 수 있다. 이렇듯 기 구축된 자료에 대한 지속적인 정보 관리는 균형성을 갖춘 말뭉치를 구성하는 요건을 갖춰 나가는 데에 중요한 역할을 할 것이다.

2.3.2. 녹음 작업의 디지털화

자료의 보관 및 재생의 용이성을 위해 2003년도부터 아날로그식 녹음 작업을 디지털화하여 텍스트 형태의 구어 전사 자료와 음성 파일을 함께 확보해 오고 있는데, 이것은 이후 정밀한 언어학적 연구와 자연 언어 처리 분야에서의 직·간접적 이용에 도움이 될 것이다.

녹음 과정 및 도구는 작년과 동일하게 디지털 방식으로 진행되었으며, SONY사의 휴대용 디지털 녹음기 “TCD-100”을 사용하였다. 그후 이를 음향 편집 프로그램인 “Cool Edit Pro”를 사용하여 Windows 표준 웨이브 파일(Windows PCM)로 변환을 하였고, 웨이브 파일의 형식은 “44,100Hz, 16Bit, Mono”으로 하였다.

3. 말뭉치 구축 지침

원시 말뭉치 구축은 2003년과 동일한 지침에 따라 진행하여 기 구축된 자료와의 일관성을 유지하였다. 2003년도 지침은 2002년도 지침을 보완 수정한 것인데, 기존의 전사 및 마크업 지침에 대한 세부적인 보완이 이루어졌고, 기존 전사 지침과는 별도로 축약형 등의 구어 이형태, 방언형, 약어, 숫자, 외래어/외국어, 유행어/신조어, 간투사 표기 등에 대한 세부 지침을 작년 보고서에 별첨한 바 있는 “구어 원시 말뭉치 표기 지침 2003. 2(Ver.01)”로 확정하였기 때문에, 이를 따라 원시 말뭉치 구축을 진행하였다.

4. 현대 국어 구어 전사 말뭉치 구축 실례

```
<!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>일상대화_초등학교동창, 전자전사자료</title>  
                <author>김창환, 윤재연, 박정은</author>  
                <sponsor>대한민국 문화관광부</sponsor>  
                <respStmt><resp>전자/표준화</resp>  
                    <name>연세대학교 언어정보연구원</name>  
                </respStmt>  
            </titleStmt>  
            <extent>10276 어절</extent>  
            <publicationStmt>  
                <distributor>국립국어원</distributor>  
                <idno>7CM00030.hwp</idno>  
                <availability><p>배포 불가</p></availability>  
            </publicationStmt>  
            <sourceDesc>  
                <p>원전 없음, 일상 대화를 녹음하여 전사함</p>  
            </sourceDesc>  
        </fileDesc>  
        <encodingDesc>  
            <projectDesc><p>21세기 세종계획 3단계 1차년도 말뭉치  
구축</p></projectDesc>
```

```

<samplingDecl><p>녹음하여 전사</p></samplingDecl>
<editorialDecl><p>21세기 세종계획 구어 말뭉치 전사 및 마크업 지침에
따름</p></editorialDecl>
</encodingDesc>
<profileDesc>
  <creation><date>2004</date></creation>
  <langUsage>
    <language id=k0 usage=99>한국어, 표준어</language>
  </langUsage>
  <particDesc>
    <person id=P1 sex=M age=20s><p>김창환, 의사, 경상</p></person>
    <person id=P2 sex=M age=20s><p>윤재연, 학생, 강원</p></person>
    <person id=P3 sex=F age=20s><p>박정은, 학생, 강원</p></person>
    <person id=P4 sex=M age=?><p>식당종업원</p></person>
    <person id=P5 sex=F age=?><p>식당종업원</p></person>
  </particDesc>
  <settingDesc>
    <p>식당에서 녹음함</p>
  </settingDesc>
  <textClass>
    <catRef scheme='SJ21' target='M2801'>구어 녹음 전사</catRef>
  </textClass>
</profileDesc>
<revisionDesc>
  <change><date>2004</date>
    <respStmt>
      <resp>프로젝트책임자</resp><name>김향희</name>
    </respStmt>
    <item>음성 발화를 녹음(녹화)하여 PC로 입력</item>
  </change>
  <change><date>2004-5</date>
    <respStmt><resp>보조원</resp><name>윤재연</name></respStmt>
    <item>자료 녹음</item>
  </change>

```

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

```

    <change><date>2004-6</date>
      <respStmt><resp>보조원</resp><name>허윤경</name></respStmt>
      <item>1차 전사</item>
    </change>
    <change><date>2004-9</date>
      <respStmt><resp>연구보조원</resp><name>양영하</name></respStmt>
      <item>2차 전사 후 띄어쓰기 오류 교정, 철자 오류 교정</item>
    </change>
    <change><date>2004-10</date>
      <respStmt><resp>연구보조원</resp><name>양영하</name></respStmt>
      <item>마크업, 태그 오류 수정, 세종 21 프로젝트 헤더 붙임,
마킹</item>
    </change>
  </revisionDesc>
</teiHeader>
<text>
  <u who=P1><s n=00001>야::,</s>
  <s n=00002>어떻게<phon>어떠게</phon> XX냐::,</s></u>
  <u who=P3><s n=00003><vocal desc='노래'></s></u>
  <u who=P2><s n=00004><unclear>야</unclear> 지랄을 해라,</s>
  <timeLine><when id=T1><when id=T2></timeLine>
  <s n=00005><anchor synch=T1>지랄을,<anchor synch=T2></s></u>
  <u who=P3><s n=00006><anchor synch=T1>암튼<anchor synch=T2> 멋지구,</s>
  <s n=00007>멋지구,</s></u>
  <u who=P1><s n=00008>야 이 노래,</s>
  <timeLine><when id=T1><when id=T2></timeLine>
  <s n=00009><anchor synch=T1>이 노래,<anchor synch=T2></s></u>
  <u who=P3><s n=00010><anchor synch=T1><vocal desc='웃음'><anchor
synch=T2></s></u>
  <u who=P1><s n=00011>이 노래 얼마나 좋은하는데 임마::,</s>
  <s n=00012><unclear>꼭이 마음에</unclear> 든다,</s></u>
  <u who=P2><s n=00013>시끄러,</s></u>
  <u who=P1><s n=00014><singing>뽀뽀뽀뽀,</singing></s>
  <s n=00015>형이 <unclear> XX그든::,</s>

```

<s n=00016>음::,</s>
 <s n=00017><pause dur=2.3></s></u>
 <u who=P3><s n=00018>암튼::,</s>
 <s n=00019>그래서 좋아.</s></u>
 <u who=P1><s n=00020>겨울에 꼭 놀러오기 약속,</s></u>
 <u who=P3><s n=00021>아까 했잖아.</s></u>
 <u who=P1><s n=00022>약속,</s>
 <s n=00023>진짜.</s>
 <timeLine><when id=T1><when id=T2></timeLine>
 <s n=00024><anchor synch=T1>나,</s>
 <s n=00025>빈말로 하는 사람<anchor synch=T2> 제일 싫어해.</s></u>
 <u who=P3><s n=00026><anchor synch=T1>나 나 가면::<trunc-iu><anchor
 synch=T2></s>
 <s n=00027>나 보드 가르쳐 줘?</s></u>
 <u who=P1><s n=00028>당근 뽕스지.</s></u>
 <u who=P3><timeLine><when id=T1><when id=T2></timeLine>
 <s n=00029>아 <anchor synch=T1>열나게 나 진짜<trunc-iu><anchor
 synch=T2></s></u>
 <u who=P2><s n=00030><anchor synch=T1>카빙 카빙,</s>
 <s n=00031>카빙<anchor synch=T2> 배운다 지금.</s></u>
 <u who=P3><s n=00032><trunc-iu>나 나 스키는 상급에서두 내려와,</s></u>
 <u who=P2><s n=00033>야,</s></u>
 <u who=P3><s n=00034>열라 잘 잘 타,</s>
 <timeLine><when id=T1><when id=T2></timeLine>
 <s n=00035><anchor synch=T1>진짜 잘 타.<anchor synch=T2></s></u>
 <u who=P2><s n=00036><anchor synch=T1>내 친구,<anchor synch=T2></s>
 <s n=00037>내 내 친구두,</s>
 <s n=00038>보드에서두<trunc-iu></s></u>
 <u who=P1><s n=00039>너두 내려와.</s>
 <s n=00040>약속,</s></u>
 <u who=P2><s n=00041><trunc-iu>보드 보드 타거던?</s></u>
 <timeLine><when id=T1><when id=T2></timeLine>
 <u who=P1><s n=00042><anchor synch=T1>같이 날짜 맞춰 내려와.<anchor
 synch=T2></s></u>

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

<u who=P3><s n=00043><anchor synch=T1>어우 난 못 타,<anchor synch=T2></s>
<s n=00044><unclear></s></u>
<u who=P1><s n=00045>아 진짜 약속해 니 이 새끼 니 어기면
어떡할래,</s></u>
<u who=P2><s n=00046>아 <unclear> <unclear>탈게,</unclear></s>
<timeLine><when id=T1><when id=T2><when id=T3><when id=T4></timeLine>
<s n=00047><anchor synch=T1>근데<trunc-iu><anchor synch=T2></s></u>
<u who=P1><s n=00048><anchor synch=T1>나도 나도 <unclear><anchor
synch=T2></s></u>
<u who=P3><s n=00049><anchor synch=T1>아 보드 타면::,<anchor
synch=T2></s>
<s n=00050>맨날 초급에서 <trunc>옆</trunc> <anchor synch=T3>옆어져.</s>
<s n=00051>뒤집어져.</s>
<s n=00052>나 스키 탈래.<anchor synch=T4></s></u>
<u who=P2><s n=00053><anchor synch=T3>아니 근데,</s>
<s n=00054>그게 또 있어.<anchor synch=T4></s></u>
<u who=P3><s n=00055>그냥.</s>
<timeLine><when id=T1><when id=T2></timeLine>
<s n=00056>난 <unclear>그렇게 <anchor synch=T1>할래.</s>
<s n=00057>가르쳐 줘야 돼.</unclear><anchor synch=T2></s></u>
<u who=P2><s n=00058><anchor synch=T1>야 야 내 친구 내 친구,<anchor
synch=T2></s>
<s n=00059>너 잘 모를걸?</s>
<s n=00060>보드 되게 잘 타거던?</s></u>
<u who=P1><s n=00061>어.</s></u>
<u who=P2><s n=00062>뭐 하는지 알아?</s></u>
<u who=P1><s n=00063>뭐.</s></u>
<u who=P2><s n=00064>그 새끼 차가 있는데?</s></u>
<u who=P1><s n=00065>어,</s></u>
<u who=P2><s n=00066>내 친구 새끼가 보드 잘 타는데,</s>
<s n=00067>눈 쫓::나 많이 내렸을 때::,</s></u>
<u who=P1><s n=00068>음.</s></u>
<u who=P2><s n=00069>내가 <laughing><unclear> 뭐 했는지
알아,</laughing></s>

<s n=00070>그 새끼가 좀 잘 살아 그래 갖구,</s>
 <s n=00071>차 위에다가,</s></u>
 <u who=P1><s n=00072>어,</s></u>
 <u who=P2><s n=00073>캠코더를 딱 천장에 달았어.</s></u>
 <u who=P1><s n=00074>어,</s></u>
 <u who=P2><s n=00075>위에다가,</s></u>
 <u who=P1><s n=00076>어,</s></u>
 <u who=P2><s n=00077>캠코더 딱 단 다음에,</s>
 <s n=00078>끈 준비해 갖구,</s>
 <s n=00079>친구 새끼가 운전하구,</s></u>
 <u who=P3><s n=00080>하::!</s></u>
 <timeLine><when id=T1><when id=T2></timeLine>
 <u who=P1><s n=00081><anchor synch=T1><unclear><anchor synch=T2></s></u>
 <u who=P2><s n=00082><anchor synch=T1>로드 보딩 했어.<anchor
 synch=T2></s></u>
 <u who=P3><s n=00083><unclear>그 사람 무슨 일 있었대?</unclear></s></u>
 <u who=P1><s n=00084>야,</s>
 <timeLine><when id=T1><when id=T2></timeLine>
 <s n=00085><anchor synch=T1>보드는,<anchor synch=T2></s></u>
 <u who=P3><s n=00086><anchor synch=T1>진짜 <unclear><anchor
 synch=T2></s></u>
 <u who=P2><s n=00087><anchor synch=T1>저거 저거<anchor synch=T2> 저거
 야마카시 있잖아 야마카시,</s>
 <s n=00088>뭐야 그거.</s></u>
 <u who=P1><timeLine><when id=T1><when id=T2><when id=T3><when
 id=T4></timeLine>
 <s n=00089>하이 새끼 <anchor synch=T1>니가 보드<trunc-iu><anchor
 synch=T2></s></u>
 <u who=P2><s n=00090><anchor synch=T1>완전<anchor synch=T2>
 <trunc>익스</trunc> 익스트림스 <unclear></s></u>
 <u who=P1><s n=00091><anchor synch=T3>니가 보드를 안 타 봤구나,<anchor
 synch=T4></s></u>
 <u who=P3><s n=00092><anchor synch=T3><unclear><anchor synch=T4></s></u>
 <u who=P1><s n=00093><vocal desc='혀차는소리, 껌'> 나는,</s>

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

<timeLine><when id=T1><when id=T2></timeLine>
 <s n=00094>내가 <anchor synch=T1><unclear> XX이라고<anchor synch=T2> 보드
 잘 탄 놈이 있어.</s></u>
 <u who=P3><s n=00095><anchor synch=T1><unclear><anchor synch=T2></s></u>
 <u who=P1><s n=00096><unclear> 왜<trunc-iu></s></u>
 <timeLine><when id=T1><when id=T2></timeLine>
 <u who=P3><s n=00097><unclear> 우리 오빠 보드 <anchor
 synch=T1><unclear>잘 타잖아.</unclear><anchor synch=T2></s></u>
 <u who=P2><s n=00098><anchor synch=T1><vocal desc='웃음'><anchor
 synch=T2></s></u>
 <u who=P1><s n=00099><anchor synch=T1>보드를,<anchor synch=T2></s>
 <s n=00100>뒤로 타거든,</s>
 <s n=00101>뒤로도 탈 수 있어,</s></u>
 <timeLine><when id=T1><when id=T2></timeLine>
 <u who=P3><s n=00102><anchor synch=T1><vocal desc='웃음, 혀'><anchor
 synch=T2></s></u>
 <u who=P2><s n=00103><anchor synch=T1>웅,<anchor synch=T2></s></u>
 <u who=P1><s n=00104>뒤로 타잖아,</s>
 <s n=00105>뒤로 타면서 내려오면서 다 찍었어,</s></u>
 <u who=P2><s n=00106>어 그래!</s>
 <timeLine><when id=T1><when id=T2><when id=T3><when id=T4><when
 id=T5><when id=T6></timeLine>
 <s n=00107>내 친구도 <anchor synch=T1>그래,<anchor synch=T2></s></u>
 <u who=P1><s n=00108><anchor synch=T1>어,<anchor synch=T2></s>
 <s n=00109><anchor synch=T3>내려오며 다 찍어.</s>
 <s n=00110>나 내려오면서 찍혔어,<anchor synch=T4></s></u>
 <u who=P2><s n=00111><anchor synch=T3>내 친구,</s>
 <s n=00112><unclear></s>
 <s n=00113>어 어,<anchor synch=T4></s></u>
 <u who=P1><s n=00114><unclear> 캠코더로 <anchor synch=T5>집에
 <unclear><anchor synch=T6></s></u>
 <u who=P3><s n=00115><anchor synch=T5>재수없어 재수없어.<anchor
 synch=T6></s></u>
 <u who=P2><s n=00116><anchor synch=T5><unclear><anchor synch=T6></s></u>

<u who=P1><s n=00117>짜::,</s>
 <s n=00118>XX 이만한 거 있어,</s></u>
 <u who=P2><s n=00119>어 어 나두 있어.</s></u>
 <u who=P3><s n=00120>아니::<trunc-iu></s></u>
 <u who=P1><s n=00121>형님 얼굴 날리면 멋있잖아,</s>
 <s n=00122><latching>형님,</s></u>
 <u who=P3><s n=00123>의사들이 의사들이 할 일이 없으니까,</s>
 <timeLine><when id=T1><when id=T2><when id=T3><when id=T4></timeLine>
 <s n=00124>이렇게 한 군데에 치중해서 <anchor synch=T1>뭔가
 한다니까?<anchor synch=T2></s></u>
 <u who=P2><s n=00125><anchor synch=T1>아니 그래<anchor synch=T2> 그제
 자기 취미야,</s>
 <s n=00126><anchor synch=T3><unclear><anchor synch=T4></s></u>
 <u who=P1><s n=00127><anchor synch=T3><unclear> 기분 나빠,<anchor
 synch=T4></s></u>
 <u who=P3><s n=00128><anchor synch=T3><unclear>취미야?</unclear></s>
 <s n=00129>취미야?<anchor synch=T4></s></u>
 <u who=P1><s n=00130>나는 솔직히 하동 온 게,</s>
 <s n=00131>보드 탈라고 왔어.</s>
 <s n=00132>진짜로.</s></u>
 <u who=P2><s n=00133>아니 근데,</s></u>
 (중략)
 ...

여 백

제 2 장

한영 병렬 말뭉치 개발

연구책임자 : 이 석 재

여 백

제 2 장

한영 병렬 말뭉치 개발

1. 연구의 필요성 및 목표

1.1 연구의 필요성

국제적인 문화와 정보의 교류가 증가하면서 자동 번역 및 통역을 비롯한 자연 언어 처리 및 언어 공학 분야의 필요성이 대두되고 있다. 이러한 현실을 반영하기 위해서는 국가 간에 언어 정보를 공유하고 자원화 하는 일이 필요하며, 기계번역, 번역 검증 시스템, 언어 교육, 비교 언어학 등의 연구와 이를 뒷받침 할 수 있는 소프트웨어의 개발을 위해서는 둘 이상의 언어를 대응시켜 구성한 병렬 말뭉치가 필수적이다. 이와 같은 병렬 말뭉치 개발의 필요성 및 선진 외국에서의 구축 현황을 고려해 볼 때, 이를 활용한 국제적인 언어 정보의 교류가 이루어지기 위해서는 한·영 병렬 말뭉치의 지속적인 확장 구축이 필요하다.

이 장에서는 한·영 병렬 원시 말뭉치 개발을 1) 연구 필요성 및 목표, 2) 한·영 병렬 말뭉치의 개발, 3) 한·영 병렬 말뭉치의 목록과 구성, 4) 한·영 병렬 말뭉치 구축 실례로 나누어 기술한다.

1.2. 연구 목표

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

2004년도 한영 병렬 말뭉치 개발에서는 48만 어절의 한영 병렬 원시 말뭉치를 구축하는 것을 목표로 하고 있다. 또한 새로운 유형의 텍스트에 대한 작업으로 인한 작업 지침에 대한 수정·보완 작업을 동시에 진행한다.

2. 한영 병렬 원시 말뭉치 개발

2.1 구축 설계 및 원칙

2.1.1 말뭉치의 설계 및 구성 원칙

1) 한·영 병렬 말뭉치 과제는 21세기 세종계획 2단계 3차년도까지(1999년~2003년)의 사업을 통해 355만 어절의 원시 말뭉치를 구축했다. 3단계(2004년~2007년)의 구축 목표량은 190만 어절이며, 3단계 1차년도인 2004년에는 48만 어절을 구축한다.

2) 한·영 병렬 원시 말뭉치의 개발은 현대 한국어와 현대 영어의 양질의 대역 텍스트를 말뭉치를 구축 대상으로 한다.

3) 말뭉치의 저작권 사용 승인을 받을 수 있는 자료를 우선 고려하되, 기계번역, 한국어와 영어의 대조 연구 및 언어 교육, 이개어 사전 편찬 등에 널리 응용할 수 있도록 말뭉치의 범용성과 응용상의 다양성을 기본적으로 고려하여 설계한다.

4) 지금까지 구축된 한영 병렬 원시 말뭉치의 매체·내용별 구성비와 텍스트 원본별 구성비에 대한 검토 결과를 근거로, 매체·내용별 말뭉치의 다양성과 원본 언어별 균형성을 갖춘 말뭉치를 구성할 수 있는 방향으로 진행하고자 한다.

2.1.2 장르별 다양성과 원본별 균형성을 고려한 말뭉치 설계

1) 장르별 다양성에 대한 고려

제2장 한영 병렬 말뭉치 개발

한·영 병렬 말뭉치의 구축은 텍스트 및 말뭉치의 저작권 해결을 전제로 해야 하기 때문에 전체적인 장르별 다양성이 고려될 수 없었다. 한편, 말뭉치의 주요한 특성으로 거론되는 균형성 또한 한·영 병렬 말뭉치에 적용할 만한 선행연구가 진행된 바가 거의 없다. 따라서 3단계 사업부터 한·영 병렬 말뭉치 전체의 장르별 다양성과 포괄성을 확보하기 위하여 지난 5년 동안의 장르별 구축량을 토대로 부족한 장르의 텍스트를 우선 고려하여 구축할 예정이다.

한·영 병렬 원시 말뭉치의 1단계(1999년~ 2000년)와 2단계(2001년~ 2003년) 구축량을 도표로 살펴보면 장르별 분포가 다음과 같다.

매체 분류	내용 분류	세부 분류	1단계(1999-2000)				2단계(2001-2003)				장르별 어절수	비율
			원텍스트 (ST)		번역텍스트 (TT)		원텍스트 (ST)		번역텍스트 (TT)			
			영어	한국어	영어	한국어	영어	한국어	영어	한국어		
신문												0.0%
잡지	잡지	총류	54,948			38,888	16,737			13,542	124,115	3.4%
책	교과서		88,672			66,237	14,520	44,496	71,329	10,790	296,044	8.2%
	상상적 텍스트	기타	47,496	50,988	73,046	32,608		102,688	144,828		451,654	12.5%
		성경						556,933			483,005	1,039,938
	기타 비상상 적 텍스트			301,388			203,865	416,894	39,408	55,804	384,809	1,402,168
기타 출판물		안내 문, 소책 자, 정부 문서					40,459			26,185	66,644	1.9%
기타 비출 판물		연설 문	34,978	9,448	16,490	27,285		49,055	82,005		219,261	6.1%
	언어별 어절수		527,482	60,436	89,536	368,883	1,045,543	235,647	353,966	918,331	3,599,824	100%
합계			1,046,337			2,553,487		3,599,824				

<표1> 세종계획의 텍스트 분류안에 따른 한·영 병렬 말뭉치의 매체·장르별 분포

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

- 기구축된 한영 병렬 말뭉치의 매체·내용별 분포가 일정하지 않은데, 이는 말뭉치 구축 대상 텍스트의 저작권 해결을 전제로 해야 하고, 또한 수집 가능한 자료의 제한 등 문제 때문에 전체적인 장르별 다양성이 고려될 수 없었던 것으로 풀이된다.
- 신문의 경우 번역시 의역이 많고 원본과 비교할 때 누락 또는 첨가된 내용이 많아 한영 병렬 말뭉치의 구축에 적합하지 않다고 판단되어 말뭉치 자료로 선정하지 않는다.
- ‘책·기타 비상상적 텍스트’ 장르의 구축량이 다른 장르의 텍스트에 비해 많은 비율을 차지하고 있다. 따라서 올해는 한영 텍스트의 1:1 대응 비율이 높으며 구축 비율이 상대적으로 낮은 연설문과 안내문, 소책자, 정부문서 등을 포함한 기타 출판물, 잡지 중심으로 자료 수집을 진행한다.
- 1999년 이후 2003년까지 구축된 한·영 병렬 말뭉치의 매체·장르별 분포를 보면, 주로 상상적 텍스트나 인문 사회 관련의 텍스트에 치중한 경향을 보인다. 3단계까지의 전체적인 병렬 말뭉치의 구성 및 향후 병렬말뭉치의 활용도 제고를 위해 기술적 텍스트나 제품 매뉴얼, 교재 등 언어교육과 언어공학 분야에서 널리 사용될 수 있는 영역의 텍스트를 중심으로 선정한다.
- 병렬 말뭉치의 경우 저작권 문제 해결이 가장 시급한 과제인 만큼 저작권 해결을 위한 다각도의 모색과 실험을 통해 원시 말뭉치 신규 구축분에 대하여 저작권 양의 각서 확보율을 높이기 위해 최대한 노력한다.
- 성경의 비율이 다소 높다는 지적이 있으나 성경의 경우 번역의 질이 높다는 장점과 더불어 가장 많은 언어로 번역된 텍스트인 만큼, 다국어 말뭉치로의 확장 가능성을 고려하여 완성된 성경 병렬 말뭉치의 구축을 목표로 지속적으로 구축해나갈 예정이다. 3단계 사업이 끝날 때까지 남은 성경 말뭉치의 양을 고려하여 해마다 일정량씩 계획적으로 구축해 나갈 예정이다.

2) 원본별 균형성에 대한 고려

한·영 병렬 말뭉치는 원본 언어별로 영한 대역 말뭉치, 한영 대역 말뭉치와 성경 말뭉치 세 가지 종류로 구분한다. 세종계획 2단계 사업부터 시작하여 구축한 성

경 말뭉치는 원본이 히브리어 텍스트이기 때문에 대부분의 성경은 히브리어를 원본으로 번역되며, 성경 말뭉치는 특별히 하나의 유형으로 분류하였다.

한·영 병렬 원시 말뭉치의 1단계(1999년 ~ 2000년)와 2단계(2001년 ~ 2003년) 구축량의 원본별 어절수는 다음과 같다.

	1단계(1999-2000)	2단계(2001-2003)	합계	비율
영한 대역 말뭉치	896,365	923,936	1,820,301	50.6%
한영 대역 말뭉치	149,972	589,613	739,585	20.5%
성경 말뭉치		1,039,938	1,039,938	28.9%
합계	1,046,337	2,553,487	3,599,824	100.0%

<표2> 텍스트 원본 언어별 한·영 병렬 원시 말뭉치의 분포

위의 도표에서 확인할 수 있듯이, 한영 대역 말뭉치의 비율이 영한 대역 말뭉치보다 훨씬 적다. 따라서 2004년도 사업에서는 한영 대역 말뭉치 자료의 수집에 특히 노력하였다. 또한, 성경의 경우 번역의 질이 높다는 장점과 더불어 가장 많은 언어로 번역된 텍스트인 만큼, 다국어 말뭉치로의 확장 가능성을 고려하고 기구축된 성경 말뭉치 분량을 감안할 때 3단계 사업이 끝날 때까지 완성된 성경 한영 병렬 말뭉치의 구축을 목표로 해마다 계획 있게 구축해 나갈 예정이다.

2.2 말뭉치 구축 방법

한·영 병렬 말뭉치의 구축은 다음과 같은 작업 과정을 거쳐 진행된다.

2.2.1 말뭉치의 설계와 자료 수집

말뭉치의 구축은 말뭉치의 설계가 자료 수집에 선행되어야 하나, 저작권 선해결의 문제와 양질의 번역 텍스트를 구하기 어렵다는 점을 감안하여 말뭉치의 설계와 동시에, 수집 가능한 자료에 대한 조사를 함께 진행한다. 따라서 기구축된 말뭉치의 장르를 검토하여 부족한 장르가 무엇인지를 확인한다. 세부적인 장르를 고려한

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

결과, 잡지 등 몇몇 장르의 텍스트가 부족한 것으로 확인되었다.

또한 저작권 사용 동의를 받을 수 있는 동시에 수집이 가능한 잡지, 연설문 및 상상적 텍스트인 소설 등의 구축을 중심으로 말뭉치를 설계하였다.

2.2.2 저작권 사용 동의 협조 요청

저작권자로부터 텍스트 사용 승인을 받기 위해서는 출판사 또는 저작권을 소유한 해당 기관의 협조가 중요하다. 저작권 소유자에게 협조 요청문을 발송하여 세종계획의 사업 내용을 설명하며 저작권 활용 동의서 서명을 부탁한다. 올해는 다음과 같은 저작권 해결 사례가 있다.

- 가. 정부 기관과 협력하여 해결한 경우: 소식지, 주한미국대사의 연설문 텍스트
- 나. 출판사와 협력하여 해결한 경우: 영어 소설의 한글 대역 텍스트
- 다. 저작권이 자동 말소된 경우: 영어 소설 텍스트

2.2.3 텍스트의 입력과 검수

전자 파일을 직접 입수하거나 또는 전자 파일이 없는 경우, 직접 원문을 입력하였다. 이러한 텍스트는 오·탈자가 포함될 수 있으므로 이들을 원본에 근거하여 검수하고 교정하여 사용하였다.

전자 파일의 경우, 세종계획 결과물 요구에 부합되는 형태로 전처리 작업을 한다.

2.2.4 문장 분할 마크업

전처리가 끝난 텍스트는 문장 분할을 진행한다.

2.2.5 문장 정렬 및 ID 문서 출력

문장 정렬 소도구 hep.exe를 이용하여 문장을 정렬하며, 문장 정렬이 끝나면 한글 텍스트와 영어 텍스트 및 정렬 텍스트의 각각의 ID문서를 출력한다.

2.3 구축 지침 보완 사항

한영 병렬 원시 말뭉치 구축은 기존의 지침에 따라 진행하여 기 구축된 자료와의 일고나성을 유지하였다. 다만, 2004년도 한영 병렬 원시 말뭉치는 구성상 소설의 비율이 증가하면서 입력과 문장 분할에서 발견된 문제들을 해결하기 위하여 다음과 같이 구축의 세부 지침 내용을 수정·보완하였다. 구축 지침은 향후 계속적으로 수정, 보완될 것이다.

2.3.1 말줄임표의 입력

- 1) 말줄임표는 "..."는 모두 특수문자의 "..."로 바꿔준다.
- 2) 원본에 "..."또는 ".."처럼 마침표 3개 이하로 말줄임표를 표기했을 경우 "..."로 수정한다. 마침표를 3개 이상 사용했을 경우 "....."로 수정한다.

2.3.2 말줄임표의 문장 분할

- 1) 말줄임표 앞에 종결어미가 올 경우, 문장을 분할한다.
 예1) </s><s>이것저것 일거리도 많구요...</s><s>저도 부지런히 돈벌이를 했
 죠.</s><s>
 예2) <s>그건 그래...</s><s>정말 언제나 젖은 것 같은 손톱이지...</s>
- 2) 말줄임표 앞에 연결어미가 올 경우, 아래의 문장 분할 기준을 적용하여 달리 처리한다.
 - 분할하는 경우
 - 문장 내용상 말줄임표가 명백한 생략의 의미를 지니고, 생략된 부분을 복원했을 경우 말줄임표 뒤의 문장과는 의미 통사적으로 독립적일 경우 문장

을 분할한다.

예1) <s>그래서 돈을 좀 얻으려고...</s><s>여기 온 것도 실은 그것 때문입니
다.</s>

- 말줄임표 전후 문맥이 다른 주어로 시작할 경우, 2003년까지의 문장분할 지침과의 일관성을 고려하여 분할한다.

예) </s><s>우리 둘은 기어오르다가 다시 내려가고...</s><s>개년은 눈사태라도 생길까 봐 드라보트에게 너무 노래를 크게 부르거나 휘파람을 불지 말라고 부탁했지.</s><s>

○ 분할하지 않는 경우

- 말줄임표 전후 문맥이 동일한 주어일 경우 분할하지 않는다.

예) <s>이 말은 원래 이렇게 사용하는 것이 아니고...</s><s>실제로 그렇게 쓰이지 않는다.</s><s>

- 예외 사항: 말줄임표 전후 문맥이 동일한 주어일지라도, 말줄임표다음에 접속사(그러나, 그리고 등 주로 문두에 나타남)가 있을 경우에는 분할한다.

예) <s>미국인은 대개 당신보다 돈이 많지만...</s><s>그러나 인식합니다.</s>

○ 문장 내용상 말줄임표가 명백한 생략을 의미할 경우

- 생략된 부분을 복원하였을 경우 뒤에 오는 문장과 별개의 문장으로 읽힐 경우 경우, 문장을 분할한다.

예) <p><s>"편지 말인가?</s><s>그래!</s><s>편지...</s><s>제발 내 두 눈 사이만 똑바로 쳐다 보라니깐!</s><s>그건 매듭 편지야.</s><s>우린 편잡지방의 장님 거지에게서 그 실 매듭 편지를 배웠지."</s></p>

○ 대화 속에 나타난 말줄임표가 한 문장 안에서 명백한 휴지의 의미로 사용되었을 경우 분할하지 않는다.

예) <s>그리고 그 다음 그 아래에...이렇게, 또 지그가 읊니다. </s>

○ 감탄사 또는 간투사 뒤에 말줄임표가 올 경우

- 웃음소리, 울음소리 등 감탄사 또는 간투사 뒤에 말줄임표가 올 경우, 뒤의 문장과 분할하여 독립된 문장으로 인정하여 분할한다.

예) <s>헛헛...</s><s>맞다. </s>

예) <s>똑바로 직선 말입니다.</s><s>그리고...</s><s>알겠습니까, 그 아래 끝에서 이번에는 경사지게,...

- 웃음소리, 울음소리 등 감탄사 또는 간투사가 말줄임표를 사이에 두고 중복 출현할 경우에는 한 문장으로 인정한다.

예) <s>"음...음...</s>아니, 당신처럼 되어보고 싶군요."

- "예/아니" 뒤에 말줄임표가 올 경우 문장을 분할하지 않는다.

예) <s>Yes... they do always have wet finger-nails...</s>

○ 말줄임표가 단어들의 나열에서 생략을 나타낼 경우, 문장 분할을 하지 않는다.

예) <s>누구의 발이나 다 피와 뼈와 근육...모두 이런 쓸모없는 것으로 되어 있어요.</s>

3. 말뭉치 목록 및 구성

3.1 말뭉치 목록

구분	파일명	제목	언어	장르	어절수	어절수 합계	저작권
영한	re0405ee.hwp	The Desert Islander	영어	책 - 소설	7,075		○
	re0405ek.hwp	무인도에 사는 사람	한글		6,043	2004	○
	re0405al.hwp	The Desert Islander	정렬				-
	re0406ee.hwp	The Door in the Wall	영어		6,727	12,948	○
	re0406ek.hwp	담장의 문	한글		6,221		○
	re0406al.hwp	The Door in the Wall	정렬				-
	re0407ee.hwp	The Garden-Party	영어		4,954	9,644	○

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

re0407ek.hwp	가든 파티	한글		4,690		O
re0407al.hwp	The Garden-Party	정렬				-
re0408ee.hwp	The Imaginative Woman	영어		9,489	17,229	O
re0408ek.hwp	환상을 좇는 여인	한글		7,740		O
re0408al.hwp	The Imaginative Woman	정렬				-
re0409ee.hwp	The Man Who Would Be King	영어		14,276	24,726	O
re0409ek.hwp	왕이 되고 싶은 사나이	한글		10,450		O
re0409al.hwp	The Man Who Would Be King	정렬				-
re0410ee.hwp	To Please His Wife	영어		6,637	12,606	O
re0410ek.hwp	아내를 위하여	한글		5,969		O
re0410al.hwp	To Please His Wife	정렬				-
re0411ee.hwp	The Whirligig of Life	영어		2,175	3,981	O
re0411ek.hwp	돌고도는 인생	한글		1806		O
re0411al.hwp	The Whirligig of Life	정렬				-
rj0412ee.hwp	Keynote Address to Hwarangdae International Symposium on "The Threat of WMD and the Security of the Korean Peninsula"	영어	기 타 비 출 판 물 - 연 설 문	2,204	3,802	O
rj0412ek.hwp	화랑대 국제 심포지엄 기조연설	한글		1,598		O
rj0412al.hwp	Keynote Address to Hwarangdae International Symposium on "The Threat of WMD and the Security of the Korean Peninsula"	정렬				-
rj0413ee.hwp	Remarks by the Ambassador at the AFTAK Breakfast Mee.hwpiting	영어		1,878	3,084	O
rj0413ek.hwp	한국무역대리점협회(AFTAK) 조찬연설	한글		1,206		O
rj0413al.hwp	Remarks by the Ambassador at the AFTAK Breakfast Mee.hwpiting	정렬				-
rj0414ee.hwp	Remarks by The Ambassador Thomas C. Hubbard to the Citizens United for a Better Society CEO Forum	영어		1,409	2,299	O
rj0414ek.hwp	바른사회를 위한 시민회의 CEO 포럼 연설	한글		890		O
rj0414al.hwp	Remarks by The Ambassador Thomas C. Hubbard to the Citizens United for a Better Society CEO Forum	정렬				-
rj0415ee.hwp	Spee.hwpch to the Chungcheong Forum	영어		1,638	2,639	O

제2장 한영 병렬 말뭉치 개발

	rj0415ek.hwp	토마스 C. 허바드 주한미국대사 총칭포럼 연설문	한글		1,001		O
	rj0415al.hwp	Spee.hwpch to the Chungcheong Forum	정렬				-
	rj0416ee.hwp	The United States and Korea Face the Challenges of 2002	영어		2,342	3,798	O
	rj0416ek.hwp	한국과 미국이 직면한 2002년의 도전	한글		1,456		O
	rj0416al.hwp	The United States and Korea Face the Challenges of 2002	정렬				-
	rj0417ee.hwp	U.S. Policy Approach to the Korean Peninsula	영어		1,684	2,728	O
	rj0417ek.hwp	미국의 대한반도 정책 접근방법	한글		1,044		O
	rj0417al.hwp	U.S. Policy Approach to the Korean Peninsula	정렬				-
	rj0418ee.hwp	U.S.-Korea Relations in Transition	영어		2,216	3,613	O
	rj0418ek.hwp	전환기의 한미 관계	한글		1,397		O
	rj0418al.hwp	U.S.-Korea Relations in Transition	정렬				-
	rj0419ee.hwp	U.S.-ROK Relations a Month After the Summit	영어		2,084	3,505	O
	rj0419ek.hwp	한미 정상회담 한달 후의 한미관계	한글		1,421		O
	rj0419al.hwp	U.S.-ROK Relations a Month After the Summit	정렬				-
합계	영한 대역 말뭉치				66,788 한글	119,720	
한영	rb0420ek.hwp	국제교류재단 소식지 2002년 2월호	한글	잡 지 - 총 류	4,187	11,297	O
	rb0420ee.hwp	Korea Foundation News Letter, Feb. 2002	영어		7,110		O
	rb0420al.hwp	국제교류재단 소식지 2002년 2월호	정렬				-
	rb0421ek.hwp	국제교류재단 소식지 2002년 5월호	한글		5,912	14,754	O
	rb0421ee.hwp	Korea Foundation News Letter, May. 2002	영어		8,842		O
	rb0421al.hwp	국제교류재단 소식지 2002년 5월호	정렬				-
	rb0422ek.hwp	국제교류재단 소식지 2002년 8월호	한글		4,411	11,750	O
	rb0422ee.hwp	Korea Foundation News Letter, Aug. 2002	영어		7,339		O
	rb0422al.hwp	국제교류재단 소식지 2002년 8월호	정렬				-
	rb0423ek.hwp	국제교류재단 소식지 2002년 11월호	한글		4,736	13,104	O
	rb0423ee.hwp	Korea Foundation News Letter, Nov. 2002	영어		8,368		O
	rb0423al.hwp	국제교류재단 소식지 2002년 11월호	정렬				-
	rb0424ek.hwp	국제교류재단 소식지 2003년 2월호	한글		5,879	14,988	O
	rb0424ee.hwp	Korea Foundation News Letter, Feb. 2003	영어		9,109		O

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

rb0424al.hwp	국제교류재단 소식지 2003년 2월호	정렬				-
rb0425ek.hwp	국제교류재단 소식지 2003년 5월호	한글		5,165	13,583	O
rb0425ee.hwp	Korea Foundation News Letter, May. 2003	영어		8,418		O
rb0425al.hwp	국제교류재단 소식지 2003년 5월호	정렬				-
rb0426ek.hwp	국제교류재단 소식지 2003년 8월호	한글		4,463	11,810	O
rb0426ee.hwp	Korea Foundation News Letter, Aug. 2003	영어		7,347		O
rb0426al.hwp	국제교류재단 소식지 2003년 8월호	정렬				-
rb0427ek.hwp	국제교류재단 소식지 2003년 11월호	한글		5,869	14,671	O
rb0427ee.hwp	Korea Foundation News Letter, Nov. 2003	영어		8,802		O
rb0427al.hwp	국제교류재단 소식지 2003년 11월호	정렬				-
rb0428ek.hwp	국제교류재단 소식지 2004년 2월호	한글		4,879	12,414	O
rb0428ee.hwp	Korea Foundation News Letter, Feb. 2004	영어		7,535		O
rb0428al.hwp	국제교류재단 소식지 2004년 2월호	정렬				-
rb0429ek.hwp	국제교류재단 소식지 2004년 5월호	한글		6,314	16,180	O
rb0429ee.hwp	Korea Foundation News Letter, May. 2004	영어		9,866		O
rb0429al.hwp	국제교류재단 소식지 2004년 5월호	정렬				-
rh0430ek.hwp	외국인 근로자를 위한 안전작업 길잡이	한글	책 - 생 활 / 설 명 서	5,971	16,975	O
rh0430ee.hwp	Safety Guidline for Foreign Workers	영어		11,004		O
rh0430al.hwp	외국인 근로자를 위한 안전작업 길잡이	정렬				-
rh0431ek.hwp	외국인 노동자를 위한 생활 매뉴얼	한글		20,215	52,524	O
rh0431ee.hwp	Living in Korea-A Manual for Migrant Workers	영어		32,309		O
rh0431al.hwp	외국인 노동자를 위한 생활 매뉴얼	정렬				-
ri0432ek.hwp	참여정부의 국방정책 2003	한글	기 타 출 판 물 - 국 방	20,759	58,114	O
ri0432ee.hwp	participatory government defense policy 2003	영어		37,355		O
ri0432al.hwp	참여정부의 국방정책 2003	정렬				-
합계	한영 대역 말뭉치			98,760 영어	262,164	

제2장 한영 병렬 말뭉치 개발

성경	rg0401ee.hwp	Proverbs	영어	책 - 기타 상상적 텍스트	14,154		X
	rg0401ek.hwp	잠언	한글		10,892		O
	rg0401al.hwp	Proverbs	정렬				-
	rg0402ee.hwp	Ecclesiastes	영어		5,013	9,039	X
	rg0402ek.hwp	전도서	한글		4,026		O
	rg0402al.hwp	Ecclesiastes	정렬				-
	rg0403ee.hwp	Song of Songs	영어		2,528	4,403	X
	rg0403ek.hwp	아가	한글		1,875		O
	rg0403al.hwp	Song of Songs	정렬				-
	rg0404ee.hwp	Isaiah	영어		34,302	60,290	X
	rg0404ek.hwp	이사야	한글		25,988		O
	rg0404al.hwp	Isaiah	정렬				-
합계	성경 말뭉치				55,997	98,778	
					한글		
총합	한영 병렬 원시 말뭉치				194,473	480,662	424,665
					영어		
	목표량 대비 비율					100.0%	88.4%

<표 3> 2004년 말뭉치 목록

위의 표에서 “O” 기호는 말뭉치의 저작권 활용 승낙서를 받았거나 또는 저작권이 자동말소된 경우로 저작권문제가 해결된 경우를 뜻하며, “X” 기호는 말뭉치의 저작권 활용 승낙서를 받지 못한 경우, “-” 기호는 정렬 문서로서 저작권과는 무관함을 뜻한다.

3.2 말뭉치 구성

3.2.1 언어별 말뭉치 구성

말뭉치 전체 480,662 어절
언어별 한글 텍스트 194,473 어절(40.46%)
 영문 텍스트 286,189 어절(59.54%)

3.2.2 원본별 말뭉치 구성

말뭉치 전체	480,662 어절
원본별	영한 대역 119,720 어절(41.8%)
	한영 대역 262,164 어절(54.6%)
	성경 98,778 어절(20.6%)

3.2.3 매체·장르별 말뭉치 구성

말뭉치 전체		480,662 어절
매체·장르별	잡지-총류	134,551 어절(28.0%)
	책-소설	94,252 어절(19.6%)
	책-교육자료 일반	69,499 어절(14.5%)
	책-기타 상상적 텍스트	98,778 어절(20.6%)
	기타 출판물-국방	58,114 어절(12.1%)
	기타 비출판물-연설문	25,468 어절(5.3%)

매체분류	내용분류	세부분류	장르별 어절수 (1999~2004)	비율 (1999~2003)	비율 (1999~2004)
신문			0	0.0%	0.0%
잡지	총류		258,666	3.4%	6%
책	교육 자료		365,543	8.2%	9%
	상상적 텍스트	소설	545,906	12.5%	13%
		성경	1,138,716	28.9%	28%
	기타 비상상적 텍스트		1,402,168	39.0%	34%
기타 출판물	사회, 생활	설명서 등	124,758	1.9%	3%
기타 비출판물	사회	연설문	244,729	6.1%	6%
	합계		4,080,446	100%	100%

<표4> 1999~2004년 매체·내용별 말뭉치 지도

- <표4>의 <비율(1999~2004)>칼럼에서 매체·내용별 말뭉치 구성이 2003년까지의 비율에 비해 미약하나마 조금씩 조정된 것을 볼 수 있으며 이는 균형성 확보를 위한 노력의 결과라 할 수 있다.

3.2.4 저작권 해결

저작권 해결 비율 총 88.4%(성경 포함)
 영문 텍스트 80.4%
 한글 텍스트 100%

3.3 텍스트별 문장 정렬 분포

3.3.1 2004년 구축된 말뭉치의 텍스트별 문장 정렬 분포

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

장르	파일명	0:1	1:0	1:1	1:2	1:3	1:4	1:5	1:6	1:7	1:8	2:1	2:2	2:3	2:4	3:1	3:2	3:3	3:4	4:1	4:2	4:3	4:5	5:1	5:2	5:3	5:5	6:1	6:2	7:1	8:1	합계
성경	rg0401			915																												915
	rg0402			222																												222
	rg0403			118																												118
	rg0404			1,292																												1292
소설	re0405	9	2	280	137	32	3	3			1	6	2	2								1										478
	re0406	16	3	223	104	44	13	7	5	5	1	6	1	1	1	2																432
	re0407			403	112	28	4	2				20	1	2		1	2										1					576
	re0408	23	3	301	121	53	14	8	2			8	4	5		1		1			1											545
	re0409	1	4	469	238	81	12	3	1	1		16	8	3	2	2		1		2												844
	re0410	6		218	133	29	11	2	2	1		6	1	5	1																	415
	re0411			104	30	11	3										1															149
문헌	rj0412			88	5																											93
	rj0413		1	82	2							2																				87
	rj0414			57	3							2				1																63
	rj0415			69	3							7				1																80
	rj0416		1	106	3							5				1																116
	rj0417		1	82	2							4																				89

제2장 한영 병렬 말뭉치 개발

[illegible]

3.3.2 1999~2004년 텍스트별 문장 정렬 유형 및 분포

한영			영한		
문장 정렬	정렬쌍	출현 확률	문장 정렬	정렬쌍	출현 확률
1:1	39583	0.846568428	1:1	64959	0.889983422
2:1	2927	0.062600252	1:2	4407	0.060378961
1:2	1799	0.038475522	2:1	1541	0.021111277
1:0	657	0.014051372	1:0	684	0.009371275
0:1	472	0.010094745	1:3	567	0.007768294
2:2	448	0.009581453	0:1	267	0.003658085
3:1	445	0.009517292	3:1	140	0.001918097
4:1	98	0.002095943	2:2	116	0.001589281
1:3	95	0.002031781	1:4	94	0.001287865
5:1	35	0.000748551	2:3	68	0.000931647
3:2	34	0.000727164	1:5	35	0.000479524
1:4	23	0.000491905	4:1	31	0.000424722
6:1	21	0.000449131	2:4	11	0.000150708
2:3	18	0.000384969	1:6	11	0.000150708
3:3	17	0.000363582	3:2	9	0.000123306
7:1	13	0.000278033	1:7	9	0.000123306
4:2	9	0.000192485	3:4	8	0.000109606
9:1	6	0.000128323	3:3	8	0.000109606
1:5	6	0.000128323	5:5	7	9.59049E-05
8:1	5	0.000106936	6:1	3	4.11021E-05
3:4	4	8.55487E-05	4:3	3	4.11021E-05
5:2	3	6.41615E-05	1:8	3	4.11021E-05
4:3	3	6.41615E-05	5:1	2	2.74014E-05
2:5	3	6.41615E-05	4:5	2	2.74014E-05
6:7	2	4.27743E-05	5:6	1	1.37007E-05
6:5	2	4.27743E-05	4:2	1	1.37007E-05
6:2	2	4.27743E-05	2:5	1	1.37007E-05
5:3	2	4.27743E-05	1:24	1	1.37007E-05
4:4	2	4.27743E-05			
2:4	2	4.27743E-05			
15:1	2	4.27743E-05			
11:1	2	4.27743E-05			
8:3	1	2.13872E-05			
7:9	1	2.13872E-05			
7:6	1	2.13872E-05			
7:5	1	2.13872E-05			
7:4	1	2.13872E-05			
7:2	1	2.13872E-05			
6:6	1	2.13872E-05			
5:7	1	2.13872E-05			
5:5	1	2.13872E-05			
4:6	1	2.13872E-05			
4:5	1	2.13872E-05			
25:1	1	2.13872E-05			

10:2	1	2.13872E-05	
10:1	1	2.13872E-05	
1:7	1	2.13872E-05	
1:6	1	2.13872E-05	
1:16	1	2.13872E-05	

한영 문장 정렬쌍	46,757개
영한 문장 정렬쌍	74,298개
성경 문장 정렬쌍	21,374개
합 계	142,429개

- 성경의 경우 텍스트 특성상 병렬 말뭉치의 문장 분할 및 문장 정렬 방법과는 달리, 성경의 절(시편의 경우 '편')을 문장 단위로 분할하며 정렬의 단위 역시 절을 기준으로 하며 1:1 정렬의 유형만 출현한다. 따라서
- 문장 정렬쌍 유형은, 영한 말뭉치가 한영 말뭉치보다 단일하다.
- 1:1 정렬 비율은 한영의 경우 약 85%, 영한의 경우 약 89%의 높은 비율을 기록하고 있어 1:1 정렬을 보다 선호하는 기계번역, 이개어사전의 자동추출 등 언어공학 분야에도 유용하게 활용될 수 있을 것으로 판단된다.

4. 한·영 병렬 말뭉치의 구축 실례

(예 1) 영한 병렬 말뭉치의 예

- 영어문서: re0405ee.hwp

<cesDoc version=3.24>

<cesHeader version="2.3">

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

```
<fileDesc>
  <titleStmt>
    <h.title>The Desert Islander</h.title>
    <respStmt>
      <respType>문헌 입력, 표준화, 헤더붙임</respType>
      <respName>연세대학교 언어정보연구원</respName>
    </respStmt>
  </titleStmt>
  <extent>
    <wordcount>7075</wordcount>
  </extent>
  <publicationStmt>
    <contributor>국립국어원</contributor>
    <pubAddress>서울시 강서구 방화3동 827번지</pubAddress>
    <availability status=restricted>배포 불가</availability>
    <PubDate>2004</PubDate>
  </publicationStmt>
  <sourceDesc>
    <biblStruct>
      <monogr>
        <h.title>The Desert Islander</h.title>
        <h.author>Stella Benson</h.author>
        <imprint>
          <pubPlace></pubPlace>
          <publisher></publisher>
          <pubDate>1930</pubDate>
        </imprint>
      </monogr>
    </biblStruct>
  </sourceDesc>
</fileDesc>
<encodingDesc>
  <projectDesc>21세기 세종계획 3단계 1차년도 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>2004/07/01</date></creation>
    <langUsage>
        <language id="en" iso639="en"></language>
    </langUsage>
    <wsdUsage>
        <writingSystem id="iso8859-1"></writingSystem>
    </wsdUsage>
    <textClass>
        <catRef scheme='SJ21' target="R1331">책: 상상적 텍스트,
소설/일반</catRef>
    </textClass>
    <translations>
        <translation trans.loc="re0405다.hwp" lang=kr
wsd="EUC-KR(KSC5601)">
            <translator>(주)아이브러리</translator>
        </translations>
</profileDesc>
<revisionDesc>
    <change>
        <changeDate>2004/12/07</changeDate>
        <respName>황은하, 박유현</respName>
        <respName>이석재</respName>
        <h.item>본문 마크업</h.item>
    </change>
</revisionDesc>
</cesHeader>

```

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

```
<text>
<body id=1>
<div id=1.1>
<head id=1.1.h1>THE DESERT ISLANDER (1930)</head>
<head id=1.1.h2>by Stella Benson (1892 - 1933)</head>
<p id=1.1.p1><s id=1.1.pl.s1>CONSTANTINE hopefully followed the Chinese
servant through the unknown house.</s><s id=1.1.pl.s2>He felt hopeful of
success in his plan of begging this Englishman for help, for he knew that
an Englishman, alone among people of a different colour (as this Englishman
was alone in this South China town), treated the helping of stray white men
almost as part of the White Man's Burden.</s><s id=1.1.pl.s3>But even
without his claim of one lonely white man upon another, Constantine would
have felt hopeful.</s><s id=1.1.pl.s4>He knew himself to be a man of
compelling manner in spite of his ugly, too long face, and his ugly, too
short legs.</s></p>
<p id=1.1.p2><s id=1.1.p2.s1>As Constantine stumped in on his hobnailed
soles, Mr. White--who was evidently not a very tactful man--said, "Oh, are
you another deserter from the Foreign Legion?"</s></p>
.
.
.
</div>
</body>
</text>
</cesDoc>
```

- 한글문서: re0405ek.hwp

```
<cesDoc version=3.24>
<cesHeader version="2.3">
  <fileDesc>
    <titleStmt>
      <h.title>무인도에 사는 사람</h.title>
```

```

    <respStmt>
      <respType>문헌 입력, 표준화, 헤더붙임</respType>
      <respName>연세대학교 언어정보연구원</respName>
    </respStmt>
  </titleStmt>
  <extent>
    <wordcount>6043</wordcount>
  </extent>
  <publicationStmt>
    <distributor>국립국어원</distributor>
    <pubAddress>서울시 강서구 방화3동 827번지</pubAddress>
    <availability status=restricted>배포 불가</availability>
    <PubDate>2004</PubDate>
  </publicationStmt>
  <sourceDesc>
    <biblStruct>
      <monogr>
        <h.title>무인도에 사는 사람</h.title>
        <h.author>Stella Benson</h.author>
        <imprint>
          <pubPlace>서울</pubPlace>
          <publisher>(주)아이브러리</publisher>
          <pubDate>2004</pubDate>
        </imprint>
      </monogr>
    </biblStruct>
  </sourceDesc>
</fileDesc>
<encodingDesc>
  <projectDesc>21세기 세종계획 3단계 1차년도 한영 병렬 말뭉치
구축</projectDesc>
  <editorialDecl>21세기 세종계획 말뭉치 문헌 입력 지침에
따름</editorialDecl>
  <samplingDecl></samplingDecl>

```

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

<refsDecl>본문, 구획, 문단, 문장 단위로 일련번호를 붙여
병렬텍스트와 정렬

```
</refsDecl>
</encodingDesc>
<profileDesc>
  <creation><date>2004/07/01</date></creation>
  <langUsage>
    <language id="kr" iso639="kr"></language>
  </langUsage>
  <wsdUsage>
    <writingSystem id="EUC-KR(KSC5601)"></writingSystem>
  </wsdUsage>
  <textClass>
    <catRef scheme='SJ21' target="R1331">책: 상상적 텍스트,
소설/일반</catRef>
  </textClass>
  <translations>
    <translation trans.loc="re0405ee.hwp" lang=en wsd="iso8859-1">
      <translator>(주)아이브러리</translator>
    </translation>
  </translations>
</profileDesc>
<revisionDesc>
  <change>
    <changeDate>2004/12/07</changeDate>
    <respName>황은하, 박유현</respName>
    <respName>이석재</respName>
    <h.item>본문 마크업</h.item>
  </change>
</revisionDesc>
</cesHeader>
<text>
<body id=1>
<div id=1.1>
<head id=1.1.h1>무인도에 사는 사람</head>
```

```

<head id=1.1.h2>스텔라 벤슨(Stella Benson, 1892-1933)</head>
<head id=1.1.h3>[목차]</head>
<p id=1.1.p1><s id=1.1.p1.s1>1. 쇼팡과 발라라이카</s></p>
<p id=1.1.p2><s id=1.1.p2.s1>2. 갑자기 다리를 잘라야한다니!</s></p>
<p id=1.1.p3><s id=1.1.p3.s1>3. 두 사람이 여행을 떠나다</s></p>
<p id=1.1.p4><s id=1.1.p4.s1>4. 모욕을 받느니 차라리...</s></p>
<head id=1.1.h4>1. 쇼팡과 발라라이카</head>
<p id=1.1.p5><s id=1.1.p5.s1>콘스탄틴은 희망을 품고 중국인 하인의 뒤를
따라 낯선 집 안으로 들어갔다.</s><s id=1.1.p5.s2>이 영국인을 친구로 삼는
계획은 아무래도 성공할 것 같다.</s><s id=1.1.p5.s3>주위에 백인이라곤 찾아볼
수 없는 이곳 중국 남부 지방에서 이 영국인은 떠돌이 백인을 돕는 것을 같은
백인으로서의 일종의 의무로 여기고 있는 것이다.</s><s id=1.1.p5.s4>설혹 그
영국인이 그 따위 피부색에 따른 의무감 같은 게 전혀 없다고 해도 콘스탄틴은
역시 희망을 잃지 않았을 것이다.</s><s id=1.1.p5.s5>긴 얼굴은 못생겼고,
다리는 짧고 보기 흉했지만, 그래도 자기 태도에는 다른 사람의 마음을 움직이는
것이 있다고 그는 자신하고 있었다.</s></p>
<p id=1.1.p6><s id=1.1.p6.s1>징을 박은 구두를 신고 콘스탄틴이 방에
들어서자 화이트 씨는 "아, 당신도 또 그 외인부대 탈주병이오?"하고
물었다.</s><s id=1.1.p6.s2>아무리 봐도 그렇게 눈치 없는 사람 같지는
않은데...</s></p>
.
.
.
.
.
.
.
</div>
</body>
</text>
</cesDoc>

```

- 정렬문서: re0405al.hwp

```

<cesAlign type=sent version=1.6>
<cesHeader version="2.3">
  <fileDesc>
    <titleStmt>

```


21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

```
<h.title>무인도에 사는 사람</h.title>
</titleStmt>
<publicationStmt>
  <distributor>국립국어원</distributor>
  <pubAddress>서울시 강서구 방화3동 827번지</pubAddress>
  <availability status=restricted>배포 불가</availability>
  <PubDate>2004</PubDate>
</publicationStmt>
<sourceDesc>
  <biblStruct>
    <monogr>
      <h.title>무인도에 사는 사람</h.title>
      <h.author>Stella Benson</h.author>
      <imprint>
        <pubPlace></pubPlace>
        <publisher></publisher>
        <pubDate>1930</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">
        <link xtargets="1.1.h2 ; 1.1.h2">
          <link xtargets=" ; 1.1.h3">
            <link xtargets=" ; 1.1.p1.s1">
              <link xtargets=" ; 1.1.p2.s1">
                <link xtargets=" ; 1.1.p3.s1">
                  <link xtargets=" ; 1.1.p4.s1">
                    <link xtargets=" ; 1.1.h4">
```

```

</linkGrp>
<linkGrp domains="1.1.pl : 1.1.p5">
<link xtargets="1.1.pl.s1 : 1.1.p5.s1">
<link xtargets="1.1.pl.s2 : 1.1.p5.s2 1.1.p5.s3">
<link xtargets="1.1.pl.s3 : 1.1.p5.s4">
<link xtargets="1.1.pl.s4 : 1.1.p5.s5">
    .
    .
    .
    .
    .
</linkGrp>
</linkGrp>
</linkGrp>
</linkList>
</cesAlign>

```

(예 2) 한영 병렬 말뭉치의 예

- 한글문서: rb0420kk.hwp

```

<cesDoc version=3.24>
<cesHeader version="2.3">
  <fileDesc>
    <titleStmt>
      <h.title>한국국제교류재단 소식지 2002년 2월호</h.title>
      <respStmt>
        <respType>문헌 입력, 표준화, 헤더붙임</respType>
        <respName>연세대학교 언어정보연구원</respName>
      </respStmt>
    </titleStmt>
    <extent>
      <wordcount>4187</wordcount>
    </extent>

```

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

```

<publicationStmt>
  <distributor>국립국어원</distributor>
  <pubAddress>서울시 강서구 방화3동 827번지</pubAddress>
  <availabilitystatus=restricted>배포 불가</availability>
  <PubDate>2004</PubDate>
</publicationStmt>
<sourceDesc>
  <biblStruct>
    <monogr>
      <h.title>한국국제교류재단 소식지 2002년
2월호</h.title>
      <h.author></h.author>
      <imprint>
        <pubPlace>서울</pubPlace>
        <publisher>한국국제교류재단</publisher>
        <pubDate>2002</pubDate>
      </imprint>
    </monogr>
  </biblStruct>
</sourceDesc>
</fileDesc>
<encodingDesc>
  <projectDesc>21세기 세종계획 3단계 1차년도 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>2004/07/01</date></creation>
  <langUsage>

```

```

    <language id="kr" iso639="kr"></language>
  </langUsage>
  <wsdUsage>
    <writingSystem id="EUC-KR(KSC5601)"></writingSystem>
  </wsdUsage>
  <textClass>
    <catRef scheme='SJ21' target="R1200">잡지: 총류</catRef>
  </textClass>
  <translations>
    <translation trans.loc="rb0420ee.hwp" lang=en wsd="iso8859-1">
      <translator></translator>
    </translation>
  </translations>
</profileDesc>
<revisionDesc>
  <change>
    <changeDate>2004/12/07</changeDate>
    <respName>황은하, 김수경</respName>
    <respName>이석재</respName>
    <h.item>본문 마크업</h.item>
  </change>
</revisionDesc>
</cesHeader>
</cesDoc>
<text>
  <body id=1>
    <div id=1.1>
      <head id=1.1.h1>한국국제교류재단소식</head>
      <head id=1.1.h2>2002.2[10권 1호]</head>
      <head id=1.1.h3>목차</head>
      <p id=1.1.p1><s id=1.1.p1.s1>신년사 2</s></p>
      <p id=1.1.p2><s id=1.1.p2.s1>새로운 10년을 시작하며</s></p>
      <p id=1.1.p3><s id=1.1.p3.s1>2002년도 사업계획 3</s></p>
      <p id=1.1.p4><s id=1.1.p4.s1>뉴스 포커스 4</s></p>
      <p id=1.1.p5><s id=1.1.p5.s1>재단창립 10주년 기념행사 성황리에

```

마쳐</s></p>

<p id=1.1.p6><s id=1.1.p6.s1>KF 논단 6</s></p>

<p id=1.1.p7><s id=1.1.p7.s1>한국과 세계를 잇는 민간외교의 초석</s></p>

<p id=1.1.p8><s id=1.1.p8.s1>해외의 한국학 8</s></p>

<p id=1.1.p9><s id=1.1.p9.s1>세필드대 한국학연구소</s></p>

<p id=1.1.p10><s id=1.1.p10.s1>펠로에세이 10</s></p>

<p id=1.1.p11><s id=1.1.p11.s1>한 나이 든 외국인 학자의 체한기</s></p>

<p id=1.1.p12><s id=1.1.p12.s1>직원 해외연수기 13</s></p>

<p id=1.1.p13><s id=1.1.p13.s1>문화는 답보, 경제는 진보?</s></p>

<p id=1.1.p14><s id=1.1.p14.s1>국제교류사업 캘린더 13</s></p>

<p id=1.1.p15><s id=1.1.p15.s1>재단 사업실적 14</s></p>

<head id=1.1.h4>이사장 신년사</head>

<head id=1.1.h5>새로운 10년을 시작하며</head>

<head id=1.1.h6>이인호</head>

<head id=1.1.h7>한국국제교류재단 이사장</head>

<p id=1.1.p16><s id=1.1.p16.s1>2001년은 지구촌 주민 전체에게 특별한 의미를 가지는 한 해였으며, 특히 우리 한국국제교류재단에게는 여러 가지로 뜻깊은 해였습니다.</s><s id=1.1.p16.s2>새 천년의 첫 걸음을 내딛는 해로서 그처럼 큰 기대를 모아왔던 2001년이 9월 11일의 테러 같은 경악스러운 사건과 그 뒤를 이은 처참한 아프간 전쟁, 그리고 그에 따르는 후유증을 유산으로 남겼지만 우리 재단으로서는 2001년이 재단 창립 10년의 역사를 중간 결산하는 해였고, 재단의 활동이 그 언제보다도 중요해진 변화된 세계환경에서 새로운 도약을 다짐하는 계기가 되기도 했습니다.</s></p>

. . .
. . .
. . .

</div>

</body>

</text>

</cesDoc>

- 영어문서: rb0420ke.hwp

```

<cesDoc version=3.24>
<cesHeader version="2.3">
  <fileDesc>
    <titleStmt>
      <h.title>Korea Foundation News Letter, Feb. 2002</h.title>
      <respStmt>
        <respType>문헌 입력, 표준화, 헤더붙임</respType>
        <respName>연세대학교 언어정보연구원</respName>
      </respStmt>
    </titleStmt>
    <extent>
      <wordcount>7110</wordcount>
    </extent>
    <publicationStmt>
      <distributor>국립국어원</distributor>
      <pubAddress>서울시 강서구 방화3동 827번지</pubAddress>
      <availability status=restricted>배포 불가</availability>
      <PubDate>2004</PubDate>
    </publicationStmt>
    <sourceDesc>
      <biblStruct>
        <monogr>
          <h.title>Korea Foundation News Letter, Feb.
2002</h.title>
          <h.author></h.author>
          <imprint>
            <pubPlace>서울</pubPlace>
            <publisher>한국국제교류재단</publisher>
            <pubDate>2002</pubDate>
          </imprint>
        </monogr>
      </biblStruct>
    </sourceDesc>
  </fileDesc>

```

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

```
<encodingDesc>
  <projectDesc>21세기 세종계획 3단계 1차년도 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>2004/07/01</date></creation>
  <langUsage>
    <language id="en" iso639="en"></language>
  </langUsage>
  <wsdUsage>
    <writingSystem id="iso8859-1"></writingSystem>
  </wsdUsage>
  <textClass>
    <catRef scheme='SJ21' target="R1200">잡지: 종류</catRef>
  </textClass>
  <translations>
    <translation trans.loc="rb0420ek.hwp" lang=kr
wsd="EUC-KR(KSC5601)">
      <translator></translator>
    </translation>
  </translations>
</profileDesc>
<revisionDesc>
  <change>
    <changeDate>2004/12/07</changeDate>
    <respName>황은하, 김수경</respName>
    <respName>이석재</respName>
    <h.item>본문 마크업</h.item>
  </change>
```

```
</revisionDesc>
</cesHeader>
<text>
<body id=1>
<div id=1.1>
<head id=1.1.h1>Korea Foundation Newsletter</head>
<head id=1.1.h2>MARCH 2002[Vol.11,No.1]</head>
<head id=1.1.h3>Contents</head>
<p id=1.1.p1><s id=1.1.p1.s1>President's New Year Message 1</s></p>
<p id=1.1.p2><s id=1.1.p2.s1>Entering a New Decade</s></p>
<p id=1.1.p3><s id=1.1.p3.s1>News Focus 3</s></p>
<p id=1.1.p4><s id=1.1.p4.s1>KF's Major Program for 2002</s></p>
<p id=1.1.p5><s id=1.1.p5.s1>Korea Foundation Begins Second Decade by
Fondly Recalling First</s></p>
<p id=1.1.p6><s id=1.1.p6.s1>KF Forum 7</s></p>
<p id=1.1.p7><s id=1.1.p7.s1>Ten Years of the Korea Foundation</s></p>
<p id=1.1.p8><s id=1.1.p8.s1>Mexico and Korea, 40 Years of
Friendship</s></p>
<p id=1.1.p9><s id=1.1.p9.s1>Korean Studies Abroad 10</s></p>
<p id=1.1.p10><s id=1.1.p10.s1>Fellow Essay 13</s></p>
<p id=1.1.p11><s id=1.1.p11.s1>An Old Man in Korea</s></p>
<p id=1.1.p12><s id=1.1.p12.s1>On the Spot 14</s></p>
<p id=1.1.p13><s id=1.1.p13.s1>Is Cultural Sector Non-marching while the
Economy Is Marching?</s></p>
<p id=1.1.p14><s id=1.1.p14.s1>New Fellowship Program 16</s></p>
<p id=1.1.p15><s id=1.1.p15.s1>Korea Foundation Activities Calender
17</s></p>
<p id=1.1.p16><s id=1.1.p16.s1>Korea Foundation Activities 18</s></p>
<head id=1.1.h4>New Year Message</head>
<head id=1.1.h5>Entering a New Decade</head>
<head id=1.1.h6>Lee In-ho</head>
<head id=1.1.h7>President</head>
<p id=1.1.p17><s id=1.1.p17.s1>In retrospect, 2001 was a momentous year
for every member of the global community.</s><s id=1.1.p17.s2>It also bore
```


21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

great significance for the Korea Foundation, as the year of the Foundation's 10th anniversary.</s><s id=1.1.pl7.s3>At the dawn of the new millennium, we shared high hopes and expectations for the future.</s><s id=1.1.pl7.s4>But we were all saddened by the heinous terrorist acts of September 11, 2001 and the ensuing Afghanistan war.</s></p>

· · ·
· · ·
· · ·

</div>
</body>
</text>
</cesDoc>

- 정렬문서: rb0420al.hwp

<cesAlign type=sent version=1.6>
<cesHeader version="2.3">
 <fileDesc>
 <titleStmt>
 <h.title>한국국제교류재단 소식지 2002년 2월호</h.title>
 </titleStmt>
 <publicationStmt>
 <distributor>국립국어원</distributor>
 <pubAddress>서울시 강서구 방화3동 827번지</pubAddress>
 <availability status=restricted>배포 불가</availability>
 <PubDate>2004</PubDate>
 </publicationStmt>
 <sourceDesc>
 <biblStruct>
 <monogr>
 <h.title>한국국제교류재단 소식지 2002년
2월호</h.title>
 <h.author></h.author>

```

    <imprint>
      <pubPlace>서울</pubPlace>
      <publisher>한국국제교류재단</publisher>
      <pubDate>2002</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">
      <link xtargets="1.1.h2 ; 1.1.h2">
      <link xtargets="1.1.h3 ; 1.1.h3">
      <linkGrp domains="1.1.pl ; 1.1.pl">
      <link xtargets="1.1.pl.s1 ; 1.1.pl.s1">
    </linkGrp>
    <linkGrp domains="1.1.p2 ; 1.1.p2">
      <link xtargets="1.1.p2.s1 ; 1.1.p2.s1">
    </linkGrp>
    <linkGrp domains="1.1.p2 ; 1.1.p3">
      <link xtargets=" ; 1.1.p3.s1">
    </linkGrp>
    <linkGrp domains="1.1.p3 ; 1.1.p4">
      <link xtargets="1.1.p3.s1 ; 1.1.p4.s1">
    </linkGrp>
    <linkGrp domains="1.1.p3 ; 1.1.p5">
      <link xtargets=" ; 1.1.p5.s1">
    </linkGrp>
    <linkGrp domains="1.1.p4 ; 1.1.p5">
      <link xtargets="1.1.p4.s1 ; ">
    </linkGrp>

```

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

```

<linkGrp domains="1.1.p5 ; 1.1.p5">
<link xtargets="1.1.p5.s1 ; ">
</linkGrp>
<linkGrp domains="1.1.p6 ; 1.1.p6">
<link xtargets="1.1.p6.s1 ; 1.1.p6.s1">
</linkGrp>
<linkGrp domains="1.1.p6 ; 1.1.p7">
<link xtargets=" ; 1.1.p7.s1">
</linkGrp>
<linkGrp domains="1.1.p7 ; 1.1.p7">
<link xtargets="1.1.p7.s1 ; ">
</linkGrp>
<linkGrp domains="1.1.p8 ; 1.1.p7">
<link xtargets="1.1.p8.s1 ; ">
</linkGrp>
.
.
.
.
</linkGrp>
</linkGrp>
</linkGrp>
</linkList>
</cesAlign>

```

(예 3) 성경 병렬 말뭉치의 예

- 영어문서: rg0401ee.hwp

```

<cesDoc version=3.24>
<cesHeader version="2.3">
  <fileDesc>
    <titleStmt>

```

```

<h.title>HOLY Bible THE GOSPEL ACCORDING TO Proverbs</h.title>
<respStmt>
  <respType>문헌 입력, 표준화, 헤더붙임</respType>
  <respName>연세대학교 언어정보연구원</respName>
</respStmt>
</titleStmt>
<extent>
  <wordcount>14154</wordcount>
</extent>
<publicationStmt>
  <distributor>국립국어원</distributor>
  <pubAddress>서울시 강서구 방화3동 827번지</pubAddress>
  <availability status=restricted>배포 불가</availability>
  <PubDate>2004</PubDate>
</publicationStmt>
<sourceDesc>
  <biblStruct>
    <monogr>
      <h.title>HOLY Bible THE GOSPEL ACCORDING TO
Proverbs</h.title>
      <h.author></h.author>
      <imprint>
        <pubPlace>New York</pubPlace>
        <publisher>AMERICAL BIBLE SOCIETY</publisher>
        <pubDate>1997</pubDate>
      </imprint>
    </monogr>
  </biblStruct>
</sourceDesc>
</fileDesc>
<encodingDesc>
  <projectDesc>21세기 세종계획 3단계 1차년도 한영 병렬 말뭉치
구축</projectDesc>
  <editorialDecl>21세기 세종계획 말뭉치 문헌 입력 지침에

```

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

```
따름</editorialDecl>
    <samplingDecl></samplingDecl>
    <refsDecl>본문, 구획, 문단, 문장 단위로 일련번호를 붙여
병렬텍스트와 정렬
    </refsDecl>
</encodingDesc>
<profileDesc>
    <creation><date>2004/07/01</date></creation>
    <langUsage>
        <language id="en" iso639="en"></language>
    </langUsage>
    <wsdUsage>
        <writingSystem id="iso8859-1"></writingSystem>
    </wsdUsage>
    <textClass>
        <catRef scheme='SJ21' target="R1332">책: 상상적 텍스트
역사</catRef>
    </textClass>
    <translations>
        <translation trans.loc="rg0401ek.hwp" lang=kr
wsd="EUC-KR(KSC5601)">
            <translator></translator>
        </translations>
</profileDesc>
<revisionDesc>
    <change>
        <changeDate>2004/12/07</changeDate>
        <respName>황은하, 박유현</respName>
        <respName>이석재</respName>
        <h.item>본문 마크업</h.item>
    </change>
</revisionDesc>
</cesHeader>
<text>
```

```

<body id=1>
<div id=1.1>
  <ver id=1.1.v1>The proverbs of Solomon son of David, king of
Israel:</ver>
  <ver id=1.1.v2>for attaining wisdom and discipline: for understanding
words of insight:</ver>
  <ver id=1.1.v3>for acquiring a disciplined and prudent life, doing what
is right and just and fair:</ver>
  <ver id=1.1.v4>for giving prudence to the simple, knowledge and
discretion to the young--</ver>
  <ver id=1.1.v5>let the wise listen and add to their learning, and let the
discerning get guidance--</ver>
  <ver id=1.1.v6>for understanding proverbs and parables, the sayings and
riddles of the wise.</ver>
  .
  .
  .
</div>
</body>
</text>
</cesDoc>

```

- 한글문서: rg0401ek.hwp

```

<cesDoc version=3.24>
<cesHeader version="2.3">
  <fileDesc>
    <titleStmt>
      <h.title>성경전서 표준새번역 잠언</h.title>
      <respStmt>
        <respType>문헌 입력, 표준화, 헤더붙임</respType>
        <respName>연세대학교 언어정보연구원</respName>
      </respStmt>
    </titleStmt>
  </fileDesc>

```

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

```
</titleStmt>
<extent>
  <wordcount>10892</wordcount>
</extent>
<publicationStmt>
  <distributor>국립국어원</distributor>
  <pubAddress>서울시 강서구 방화3동 827번지</pubAddress>
  <availability status=restricted>배포 불가</availability>
  <PubDate>2004</PubDate>
</publicationStmt>
<sourceDesc>
  <biblStruct>
    <monogr>
      <h.title>성경전서 표준새번역 잠언</h.title>
      <h.author></h.author>
      <imprint>
        <pubPlace>서울</pubPlace>
        <publisher>대한성서공회</publisher>
        <pubDate>1993</pubDate>
      </imprint>
    </monogr>
  </biblStruct>
</sourceDesc>
</fileDesc>
<encodingDesc>
  <projectDesc>21세기 세종계획 3단계 1차년도 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>2004/07/01</date></creation>
  <langUsage>
    <language id="kr" iso639="kr"></language>
  </langUsage>
  <wsdUsage>
    <writingSystem id="EUC-KR(KSC5601)"></writingSystem>
  </wsdUsage>
  <textClass>
    <catRef scheme='SJ21' target="R1332">책: 상상적 텍스트,
역사</catRef>
  </textClass>
  <translations>
    <translation trans.loc="rg0401ee.hwp" lang=en wsd="iso8859-1">
    <translator></translator>
  </translations>
</profileDesc>
<revisionDesc>
  <change>
    <changeDate>2004/12/07</changeDate>
    <respName>황은하, 박유현</respName>
    <respName>이석재</respName>
    <h.item>본문 마크업</h.item>
  </change>
</revisionDesc>
</cesHeader>
<text>
<body id=1>
<div id=1.1>
  <ver id=1.1.v1>[잠언의 목적과 주제] 이것은 다윗의 아들 이스라엘 왕
솔로몬의 잠언이다.</ver>
  <ver id=1.1.v2>이 잠언은 지혜와 훈계를 알게 하며, 명철의 말씀을 깨닫게
하며,</ver>
  <ver id=1.1.v3>정의와 공평과 정직을 지혜롭게 실행하도록 훈계를 받게

```


하며,</ver>

<ver id=1.1.v4>ㄱ) 어수룩한 사람을 슬기롭게 하여 주며, 젊은이들에게 지식과 분별력을 갖게 하여 주는 것이니, (ㄱ. 어수룩한 사람으로 번역된 히브리어 프타임은 도덕적 방향감각이 없어서 악으로 기울어질 수 있는 단순한 사람을 일컬음(22.23절을 참조))</ver>

<ver id=1.1.v5>지혜 있는 사람은 이 가르침을 듣고 학식을 더할 것이요, 명철한 사람은 지혜를 더 얻게 될 것이다.</ver>

<ver id=1.1.v6>잠언과 비유와 지혜 있는 사람의 말과 그 심오한 뜻을 깨달아 알 수 있을 것이다.</ver>

· · ·
· · ·
· · ·

</div>
</body>
</text>
</cesDoc>

- 정렬문서: rg0401al.hwp

<cesAlign type=sent version=1.6>

<cesHeader version="2.3">

<fileDesc>

<titleStmt>

<h.title>성경전서 표준새번역 잠언</h.title>

</titleStmt>

<publicationStmt>

<distributor>국립국어원</distributor>

<pubAddress>서울시 강서구 방화3동 827번지</pubAddress>

<availability status=restricted>배포 불가</availability>

<PubDate>2004</PubDate>

</publicationStmt>

<sourceDesc>

<biblStruct>

<monogr>

```

    <h.title>성경전서 표준새번역 잠언</h.title>
    <h.author></h.author>
    <imprint>
        <pubPlace>서울</pubPlace>
        <publisher>대한성서공회</publisher>
        <pubDate>1993</pubDate>
    </imprint>
</monogr>
</biblStruct>
</sourceDesc>
</fileDesc>
</cesHeader>
<linkList>
<linkGrp domains="1 ; 1">
<linkGrp domains="1.1 ; 1.1">
<link xtargets="1.1.v1 ; 1.1.v1">
<link xtargets="1.1.v2 ; 1.1.v2">
<link xtargets="1.1.v3 ; 1.1.v3">
<link xtargets="1.1.v4 ; 1.1.v4">
<link xtargets="1.1.v5 ; 1.1.v5">
<link xtargets="1.1.v6 ; 1.1.v6">
    .
    .
    .
</linkGrp>
</linkGrp>
</linkList>
</cesAlign>

```

여 백

제 3 장

한일 병렬 말뭉치 개발

연구책임자 : 전 형 식

여 백

제 3 장

한일 병렬 말뭉치 개발

1. 개요

본 연구는 21세기 세종계획 국어 기초자료 구축 사업의 일환으로 한일 병렬 말뭉치를 구축하는 것이다.

2004년도 한일 병렬 말뭉치 구축 분과에서는 22만 어절의 원시 말뭉치와, 12만 어절의 형태소 분석 말뭉치의 구축을 목표로 하였다. 2001년도부터 지난 3년 동안 축적된 노하우를 충분히 살려 다양한 장르의 자료를 구축하기 위하여 한일 병렬 말뭉치 구축 분과에 맞는 작업 기준의 개선 및 새로운 기준의 적용 등을 병행하여야 하는 작업이 필요하게 되었다. 이에 2004년도 작업에서는 세종계획 국어 특수자료 구축 분과의 기본 원칙을 따르되, 실용성이 있는 한일 병렬 말뭉치 구축을 위한 적극적인 연구를 추진하였다.

또한, 올해 한일 병렬 말뭉치 구축 분과에서는 기계화에 의한 작업 자동화의 실현이라는 차원에서 여러 프로그램을 사용하는 등 시범적인 방법으로 말뭉치 구축 작업이 이루어졌다. 그 결과 대체로 순조롭게 작업이 진행되었으며, 향후 이러한 방법의 효율성 및 가능성을 확인할 수 있는 많은 정보를 얻을 수 있었다. 이를 바탕으로 내년도부터는 조금 더 효율적인 작업이 이루어질 것이다.

2. 설계 원칙

- 현대 한국어와 현대 일본어의 대응 자료를 말뭉치 구축 대상으로 한다.
- 기계번역, 전자사전 구축, 자연언어 처리, 언어 교육 분야, 한일어의 대조 연구 등에 널리 활용 할 수 있도록 범용성을 갖도록 한다.
- 현대 한국어와 일본어의 대표성과 균형성을 갖도록 한다.
- 표준화 지침에 의거하여 부호와 문서 형식을 통일한다.
- 말뭉치의 효율적 관리와 이용에 필요한 헤더 정보와 마크업을 부여한다.
- 연구 결과물은 한글 2002 이상 파일(.hwp) 또는 Unicode 텍스트 파일(.txt)로 구축한다. 2004년도 작업부터는 모든 결과물을 Unicode 텍스트 파일(.txt)로 구축한다.
- 최종 오류율이 0.1% 이내가 되도록 한다.

3. 작업지침

3.1 기본 원칙

- 모든 작업은 작업 지침(21세기 세종계획의 기본 지침에 준하고 있음)을 작성하고 이에 준하여 작업을 진행시킨다. 다만, 새로 발견된 문제점을 해결하기 위한 새 지침은 한일 병렬 말뭉치 구축 분과 내 정기 회의에서 검토 후 적용하기로 한다.
- 데이터의 입력과 저장 방식은 국립국어원의 방침에 따르며 이를 작업 지침에 명시하여 혼란을 방지한다.
- 원문 유지 : 문헌의 경우 원문을 그대로 유지한다.
- 문서 정보 표기 방식 : SGML/CES 방식을 택한다.
- 헤더 : 헤더는 자세히 표기한다.

- 본문 마크업 : 본문에는 최소한의 정보 마크업을 부여한다.
- 일본어 문서의 어절수 : 2003년도 작업까지는 샘플조사(각 장르별로 하나의 파일에서 한 페이지를 조사)를 실시하여 한 행의 평균 어절수를 계산한 후, 전체 문서의 크기에 대비하여 추정치를 산출하였으나, 2004년도부터는 기계 처리를 이용하여 보다 정확한 어절수 산출 방식을 택한다.

3.2 구축 대상 및 방법

- 현대어의 언어 현상이 잘 나타나 있는 장르와 종류를 대상으로 구축한다.
- 구축 대상 자료는 신문 기사, 잡지 기사, 논문, 소설, 영화 및 애니메이션 시나리오이다.
- 한일 대역과 일한 대역문을 절반씩 구축하는 것을 목표로 한다.
- 구축 방법은 한영 병렬 말뭉치 구축 분과의 구축 방법과 연구 내용에 준하되, 일본어 문법 구조의 차이 등으로 그 적용이 곤란한 경우에는 한일 병렬 말뭉치 구축 분과 내에서 정하는 새로운 기준을 적용하도록 한다.

3.3 자료 선정

- 자료의 선정은 한일 병렬 말뭉치 구축 분과 내 자료 평가 위원회에서 추천된 것을 대상으로 한다.
- 각 원문 자료의 선정은 현대 한국어와 현대 일본어의 성격이 잘 반영되어 있는 것을 선정하는 기준으로 하여, 이에 대한 대역문 자료도 대역 내용의 신뢰성 및 타당성을 고려한 질이 높은 자료를 선정하도록 한다.
- 저작권 문제가 해결된 자료를 우선적으로 선정한다.

3.4 파일 이름

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

파일명의 부여는 21세기 세종계획의 작업 지침에 따르며, 8자리 수를 기본으로 한다.

첫 자리는 한일 병렬 말뭉치 고유 기호인 j를, 두 번째 자리는 매체와 장르 구분을 명시하는 기호(a-k)를, 3-6째 자리는 0401부터 자료의 수량에 따른 일련번호를 부여한다. 즉, 2004년도 자료를 나타내는 '04'와 나머지 두 자리에 동일 장르의 일련번호가 부여된다. 그러므로 동일 장르내의 일련번호는 0401부터 최대 0499가 된다.

끝의 두 자리는 각각 한국어 원본 kk, 한일 번역자료 kj, 일본어 원본 jj, 일한 번역자료 jk로 표시한다. 그리고 정렬문서는 al로 표시한다.

예) jj0401jj.txt - 일본어 논문 원본(언어 : 일본어)/ jj0401jk.txt - 일본어 논문 대역문(언어 : 한국어)

ja0401kk.txt - 한국어 신문 원본(언어 : 한국어)/ ja0401kj.txt - 한국어 신문 대역문(언어 : 일본어)

je0401al.txt - 소설 정렬 문서

<매체와 장르 분류 코드>

a-신문

b-잡지

c-책-교과서

d-책-사전

e-책-소설

f-책-시

g-책-기타 상상적 텍스트

h-책-정보(인문, 사회, 자연, 예술), 기타 비상상적 텍스트

i-기타 출판물

j-기타 비출판물

k-시나리오

3.5 입력 환경

훈글 2002 이상 혹은 Unicode 대응 텍스트 에디터에서 입력 작업을 한다.

<훈글 2002 이상의 경우>

전체를 바탕 스타일로 하고, 그 속성은 다음과 같다.

- 글자 모양
 - 명조(한국어) / 신명조 약자(일본어)
 - 글자 크기 10 pt
 - 글자 간격 0%
 - 장평 100%
- 문단 모양
 - 왼쪽, 오른쪽, 문단 간격, 낱말 간격 모두 0
 - 들여쓰기 2
 - 줄 간격 150%
 - 정렬방식 왼쪽
- 편집 용지
 - A4
 - 용지방향 좁게
 - 위, 아래, 왼쪽, 오른쪽, 머리말, 꼬리말, 모두 15.0mm, 제본0mm
- 페이지 : 표시하지 않음
- 머리말/꼬리말 : 표시하지 않음

<Unicode 대응 텍스트 에디터의 경우>

한국어, 일본어 양 언어를 지원할 수 있는 환경으로 설정한 후 입력한다. 다만, 이 경우 OS 기반은 Windows 2000 혹은 Windows XP 환경에서 작업을 한다.

Unicode 대응 텍스트 에디터로는 Maruo(마루오), UltraEdit 등이 있으며, 훈글 2002와 같은 워드프로세서와 달리 순수하게 텍스트만 다루므로 별도의 문서 속성 지정을 할 필요가 없다.

3.6 저장

원칙적으로 21세기 세종계획의 작업 지침에 따른다. 문서 저장 형식의 지침이 Unicode로 통일됨에 따라 2004년도 작업부터는 파일 편집의 용이성 및 효율성을 고려하여 모든 파일의 저장 형식을 Unicode 텍스트 문서(.txt)로 통일하였다.

저장 형식이 한글 2002 이상 파일(.hwp)이 아닌 Unicode 텍스트 문서(.txt)로 한 이유로서는 파일을 편집함에 있어서 범용성이 높다는 점을 들 수 있다. 자료를 입력하여 마크업, 헤더 부여, 검수 작업 등 각 과정에서 거쳐야할 작업을 정확하게 그리고 효율적으로 수행하기 위해서는 한글 2002의 기능만으로는 어려운 점이 많다. 따라서 범용성이 높은 Unicode 텍스트 문서(.txt)를 택함으로서 각 작업 과정에 맞는 다양한 보조 프로그램을 사용할 수 있다는 것이 큰 이유가 된다.

워드프로세서 한글 2002 이상으로부터는 기본 코드가 Unicode로 되어 있어 Unicode 텍스트 문서(.txt)와의 호환성에 대해서도 문제가 없다.

3.7 헤더

병렬 말뭉치의 헤더는 CES 표준을 채택한다. CES표준은 TEI를 완전히 포함하고 있기 때문에 기존의 세종계획 말뭉치 표준 헤더와의 호환에는 문제가 없다.

기본적인 기술 사항은 한국어로 표기하되, 일본어 원문 자료 중 제목, 저자 이름, 출판사 등, 한국어 표기하기 어려운 부분은 원문 자료의 언어에 맞추어 표기한다. 정렬 문서의 경우 기본 기술 사항을 제외하고 원문 자료의 언어에 맞추어 표기하는 것을 원칙으로 한다. 또한, 정렬 문서의 헤더는 정보가 중복되므로 헤더의 모든 항목을 표기하지 않고 최소한의 서지 정보만을 나타내도록 한다.

헤더 항목의 어절수 부분은 일본어의 경우 띄어쓰기를 하지 않은 언어이므로, 다음 <표1>과 같이 우선 형태소 단위로 기계적으로 분할한 뒤, 한국어 띄어쓰기 단위로 다시 연결시키는 매크로를 실행하여 어절수를 산출하기로 한다.

<표1> 일본어 어절수 산출 방법

	분할 기준에 따른 분할수 차이	처리 방법
일본어 원문 <분할수 0 >	およそ30人と国務省役人らの厳しい護衛が登場したのは	
일본어 형태소 단위 기준 <분할수 16 >	およそ 30 人 と 国務省 役人 ら の 厳しい 護衛 が 登場 し た の は	일본어 형태소 분석기를 실행하여 일본어 형태소 단위로 자동 분할
한국어 띄어쓰기 단위 기준 <분할수 8 >	およそ 30人と 国務省 役人らの 厳しい 護衛が 登場した のは	일본어 형태소 단위를 한 국어 띄어쓰기 단위로 재 편집하는 매크로를 실행 후 어절수를 산출

이와 같은 방법으로 일본어 문서를 한국어 띄어쓰기 단위로 편집 후 어절수를 산출한다. 이 방법으로 일본어 문서를 한국어 띄어쓰기 단위로 편집하는 정확도는 파일 내용에 따라 약간의 차이는 보였으나 95% 이상의 일치도가 확인되었으며, 기존의 어절수 산출 방법(각 장르별로 하나의 파일에서 한 페이지를 조사하여 한 행의 평균 어절수를 계산한 후, 전체 문서의 크기에 대비하여 추정치를 산출함)보다 훨씬 신빙성 있는 수치를 도출할 수 있게 되었다.

3.8 마크업

기본적인 마크업 방식은 2003년도 작업 과정과 동일하다. 다만, 2004년도부터 새로 시작된 시나리오 자료의 마크업 방법은 다음과 같이 발화자 이름을 센텐스 단위로 나누고, 이하에 나타나는 발화 내용을 마침표나 물음표, 느낌표의 개수에 관계없이 하나의 센텐스로 묶기로 하였다.

이에 따라 소설이나 잡지 기사 등 다른 장르의 대화를 나타낼 수 있는 부분에서도 같은 기준을 적용하여 따옴표와 따옴표 사이에 있는 대화문은 하나의 센텐스로 처리하였다.

【시나리오 대화문 마크업 방법】

<p id=1.1.p124><s id=1.1.p124.s1>담당의:</s><s id=1.1.p124.s2>이건 늘 드시던 약입니다. 일시적인 효과밖엔 없습니다만. 어쨌든 너무 심각하게 생각하지 마시고 부인과 함께 계실 때는 가능한 한 대화를 많이 하세요. 사정이 허락된다면 여행을 가시는 것도 괜찮고요. 무슨 일이 생기면 연락주세요.</s></p>

<p id=1.1.p125><s id=1.1.p125.s1>니시:</s><s id=1.1.p125.s2>....</s></p>

<p id=1.1.p126><s id=1.1.p126.s1>담당의:</s><s id=1.1.p126.s2>그리고 이 셔츠. 좀 크려나. 괜찮으시면 입으세요.</s></p>

【소설 대화문 마크업 방법】

<p id=1.1.p31><s id=1.1.p31.s1>“어떻게……. 어떻게…….”</s></p>

<p id=1.1.p32><s id=1.1.p32.s1>선생님은 같은 말을 반복하셨다.</s><s id=1.1.p32.s2>그 말은 인적이 뜬 한낮에 묘한 뉘앙스를 풍기며 울렸다.</s><s id=1.1.p32.s3>나는 즉시 뭐라고 대답할 수 없었다.</s></p>

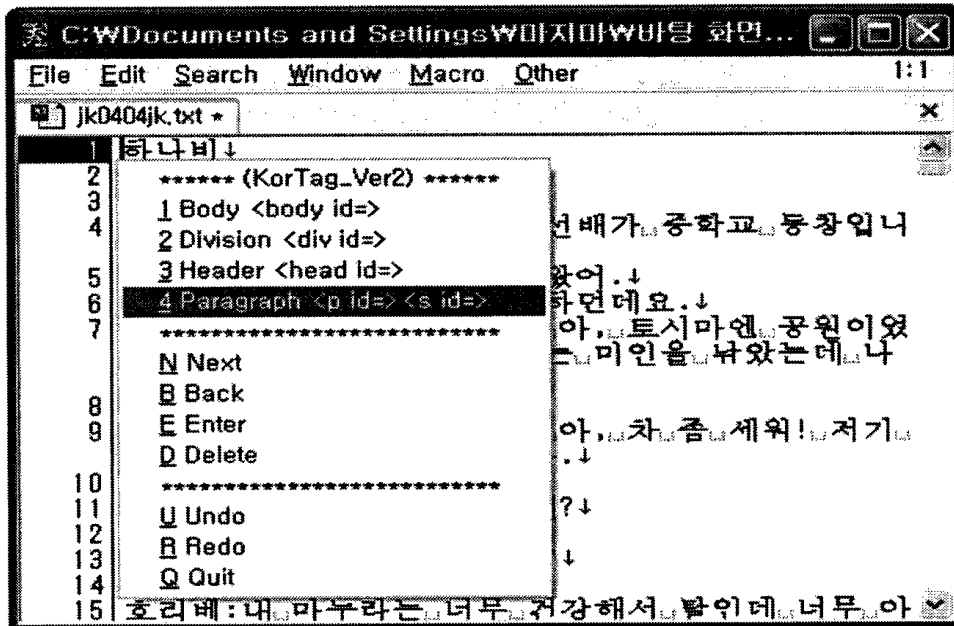
<p id=1.1.p33><s id=1.1.p33.s1>“내 뒤를 따라온 겁니까? 어떻게…….”</s></p>

또한 2004년도 작업부터는 마크업 과정을 반자동화 하는 매크로를 작성하여 태그 부분의 기술 사항 오류를 최소화하도록 하였다. 이 매크로는 텍스트 에디터 “Maruo(마루오)”¹⁾를 이용해 사용하는 것으로 본 보고서에서 설명하는 매크로는 모두 텍스트 에디터 “Maruo(마루오)” 전용이다.

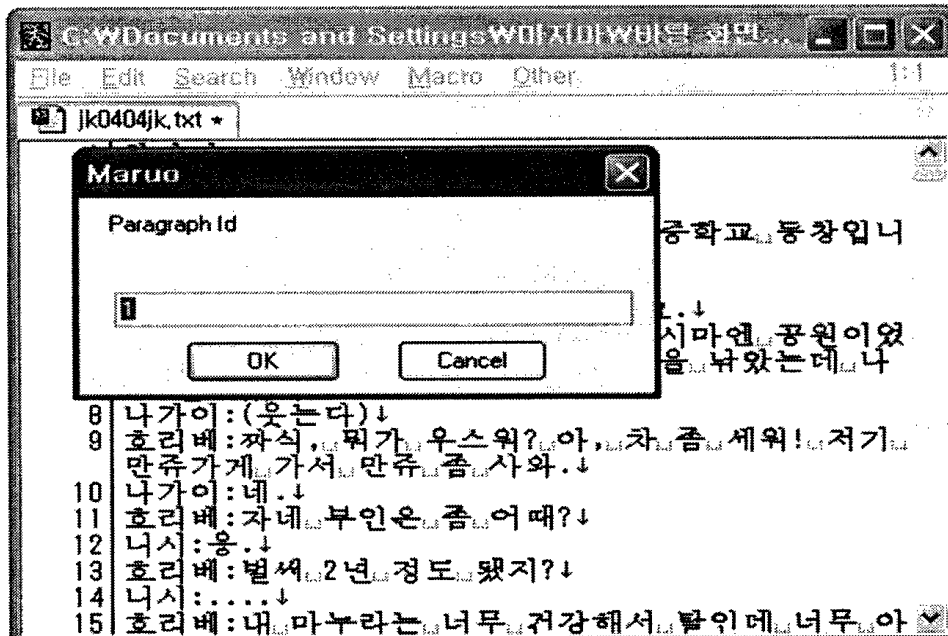
이하의 마크업 작업용 매크로의 동작 화면이다. 이 매크로의 활용으로 올해 마크업 작업 시에는 id 번호의 순서 오류나 태그 부분에 사용되는 부호의 누락 등 단순 입력 오류를 완벽하게 막을 수 있게 되었다. 이는 결과물의 최종 오류율을 줄이는 데에도 크게 기여할 것으로 생각된다. 또한 단순 작업에 의한 연구원의 정신적 부담을 해소하고 반드시 사람의 손을 거쳐야 할 검수 작업에 전념할 수 있게 한다는 점에도 큰 의미가 있을 것이다.

1) 일본 텍스트 에디터 히데마루(秀丸)의 영문판. <http://hide.maruo.co.jp/>에서 다운로드 가능한 셰어 프로그램이다.

【마크업 작업용 매크로의 동작 1】



【마크업 작업용 매크로의 동작 2】

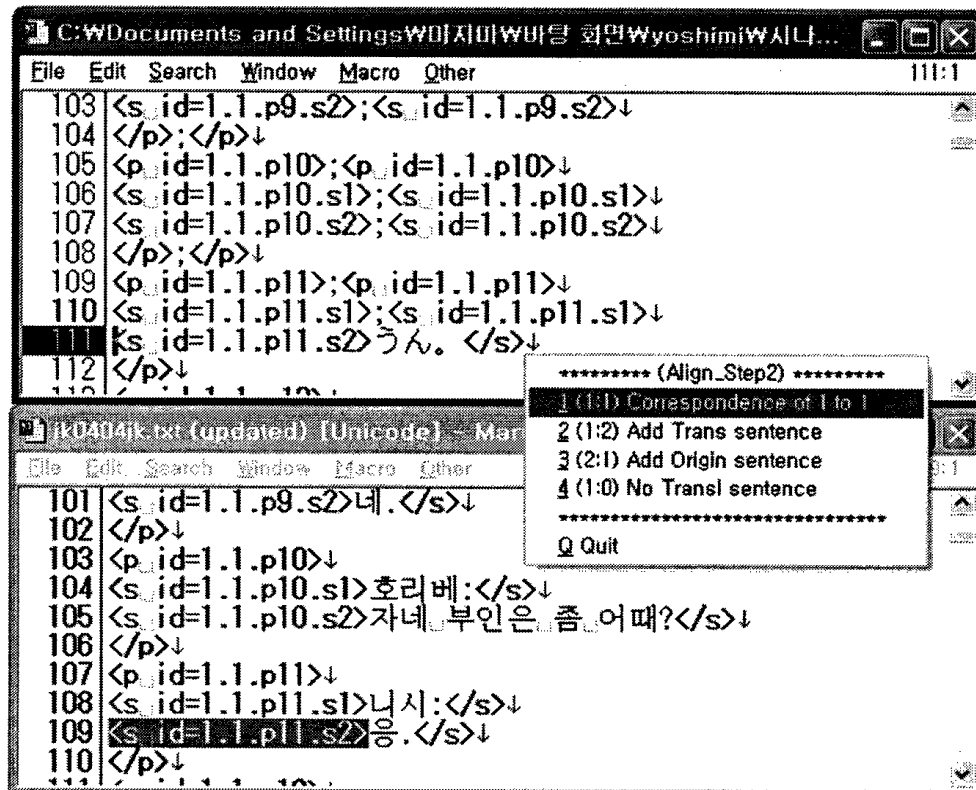


3.9 문장 정렬

병렬 말뭉치의 정렬은 문장 대 문장 정렬을 기본으로 한다. 역시 2003년도 작업 형식과 동일하다.

이 정렬 작업도 수작업에 의존했던 부분을 대신해 최대한 기계화를 구현한 반 자동 매크로를 작성하여 작업을 진행하였다.

【정렬 작업용 매크로의 동작】



문장 정렬용 매크로는 한 줄에 한 센텐스만 나타낼 수 있게 설계되었으므로, 시각적으로도 작업하기가 용이하며 마크업 작업 시에 생긴 오류도 발견하기 쉽다. 그러므로 전체적인 검수 작업 전에 각 과정별 매크로를 실행하면서도 검수를 할 수

있게 되는 셈이다. 이는 최종 오류율을 줄이는 효과도 가져올 것이다.

3.10 검수

검수 작업은 우선 두 명으로 조를 편성하여 같은 장르의 자료를 작업한 뒤, 자신의 파일을 검수하는 1차 검수, 상대의 파일을 검수하는 2차 검수, 마지막으로 전체적으로 팀장의 검수를 받는 3차 검수 과정을 거치도록 되어 있다.

입력 과정에서의 오류는 그 다음 작업 과정인 마크업 작업에도 영향을 미치며, 마찬가지로 마크업 과정에서의 오류는 정렬 작업에도 영향을 미친다. 따라서 다음 작업 과정으로 가기 전에 이루어지는 검수 작업 시에는 특히 세심한 검수 작업이 필요하게 될 것이다.

문제점 발견 시에는 즉시 팀장에게 연락하여 정기회의 때 해결 방침을 정하도록 하였다.

4. 작업 파일 목록

4.1 원시 말뭉치 구축량

2004년도 한일 병렬 말뭉치 구축 분과에서 작업한 원시 말뭉치의 구축량을 나타내면 다음과 같다.

- 총 220,339 어절 (100%)
- 한국어 텍스트 110,331 어절 (50.1%) / 일본어 텍스트 110,008 어절 (49.9%)
- 한국어 파일수 64 / 일본어 파일수 64
- 한국어 원문 파일수 74 / 일본어 원문 파일수 54

4.2 원시 말뭉치 전체 목록

전체 목록은 다음과 같다. 저작권 부분의 “○” 표시는 저작권 승인을 받을 것, “x” 표시는 저작권 승인을 받지 못 한 것, “-” 표시는 저작권과 관련 없는 자료임을 나타낸다.

원본 언어	파일명	제목	텍스트 언어	장르	출판일	어절수	저작권
일본어	je0401jj	こころ-先生と私	일본어	소설	2002/12/06	14,859	-
일본어	je0401jk	마음-선생님과 나	한국어	소설	2002/12/06	16,270	○
일본어	je0402jj	こころ-両親と私	일본어	소설	2002/12/06	7,833	-
일본어	je0402jk	마음-양친과 나	한국어	소설	2002/12/06	8,574	○
일본어	je0403jj	こころ-先生と遺書	일본어	소설	2002/12/06	25,821	-
일본어	je0403jk	마음-선생님과 유서	한국어	소설	2002/12/06	28,065	○
소설 종합						101,422	
일본어	jj0401jj	生成語彙論と対訳辞書	일본어	논문	2000/05	3,581	○
일본어	jj0401jk	생성어휘론과 대역사전	한국어	논문	2000/05	3,294	○
일본어	jj0402jj	韓国における専門用語の歴史と現状	일본어	논문	2000/02	3,606	○
일본어	jj0402jk	한국에 있어서의 전문용어 역사와 현상	한국어	논문	2000/02	3,603	○
일본어	jj0403jj	韓日対訳構文辞書における文型の意味的特徴	일본어	논문	2002/11	3,383	○
일본어	jj0403jk	한일대역구문사전에 있어서 문형의 의미적 특징	한국어	논문	2002/11	3,006	○
일본어	jj0404jj	共起關係を用いた同訓異字の意味弁別	일본어	논문	2003/11	3,384	○
일본어	jj0404jk	공기 관계를 이용한 동훈이자(同訓異字)의 의미 변별	한국어	논문	2003/11	3,125	○
일본어	jj0405jj	日韓比較語彙研究における意味の問題について	일본어	논문	2003/03/31	2,601	○
일본어	jj0405jk	일한비교어휘연구에 있어서 의미의 문제에 대하여	한국어	논문	2003/03/31	2,490	○
한국어	jj0411kj	日本の国語情報化	일본어	논문	2002/05/30	4,041	○
한국어	jj0411kk	일본의 국어 정보화	한국어	논문	2002/05/30	3,670	○

제3장 한일 병렬 말뭉치 개발

한국어	jj0412kj	第二外国語の習得と対訳辞書	일본어	논문	2003/06	3,486	○
한국어	jj0412kk	제2외국어 습득과 대역사전	한국어	논문	2003/06	3,221	○
한국어	jj0413kj	コーパスの意味情報を利用した語彙の対照研究	일본어	논문	2001/12/09	2,496	○
한국어	jj0413kk	코퍼스의 의미정보를 이용한 어휘의 대조연구	한국어	논문	2001/12/09	2,888	○
논문 종합						51,875	
일본어	jk0401jj	もののけ姫	일본어	시나리오	1997/07	7,327	×
일본어	jk0401jk	모노노케 히메(원령 공주)	한국어	시나리오	1997/07	6,084	×
일본어	jk0402jj	となりのトトロ	일본어	시나리오	1988/04/06	2,869	×
일본어	jk0402jk	이웃의 토토로	한국어	시나리오	1988/04/06	3,313	×
일본어	jk0403jj	四月物語	일본어	시나리오	1999/03/17	2,347	×
일본어	jk0403jk	4월 이야기	한국어	시나리오	1999/03/17	2,589	×
일본어	jk0404jj	花火	일본어	시나리오	1986/06/21	2,791	×
일본어	jk0404jk	하나비	한국어	시나리오	1986/06/21	2,252	×
시나리오 종합						29,572	
일본어	jb0401jj	映像革命、そして押井守の世界	일본어	잡지	2004/05	417	×
일본어	jb0401jk	영상혁명 그리고 오시이 마모루의 세계...	한국어	잡지	2004/05	431	×
일본어	jb0402jj	韓国社会のトレンドを読むならホームショッピング	일본어	잡지	2004/05	656	×
일본어	jb0402jk	한국 사회의 트렌드를 읽으려면 홈쇼핑을 보라	한국어	잡지	2004/05	688	×
일본어	jb0403jj	クイズジャパン	일본어	잡지	2004/05	278	×
일본어	jb0403jk	퀴즈재팬	한국어	잡지	2004/05	243	×
일본어	jb0404jj	テレビ・ラジオ・照明・舞台美術の世界へ	일본어	잡지	2004/05	493	×
일본어	jb0404jk	텔레비전, 라디오, 조명, 무대미술의 세계로	한국어	잡지	2004/05	448	×
일본어	jb0405jj	本質を見つめ、学問を追究し、真の現代人を育成する	일본어	잡지	2004/05	395	×
일본어	jb0405jk	본질의 응시와 깊은 학문의 연구를 통해 여러분을 참된 현대인으로 육성시키는	한국어	잡지	2004/05	390	×

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

일본어	jb0406jj	おにぎり!	일본어	잡지	2004/05	541	×
일본어	jb0406jk	오니기리!	한국어	잡지	2004/05	537	×
일본어	jb0407jj	韓国ドラマ	일본어	잡지	2004/05	914	×
일본어	jb0407jk	한국드라마	한국어	잡지	2004/05	812	×
일본어	jb0408jj	フリマやろうよ	일본어	잡지	2004/05	740	×
일본어	jb0408jk	프리마켓에 나가자	한국어	잡지	2004/05	634	×
일본어	jb0409jj	セルレンダリング・フル 3 D アニメーションの限界と機能 性	일본어	잡지	2004/06	522	×
일본어	jb0409jk	셀 렌더링 풀 3D 애니매 이션의 한계와 가능성	한국어	잡지	2004/06	501	×
일본어	jb0410jj	韓国の道路は工事中	일본어	잡지	2004/06	912	×
일본어	jb0410jk	한국의 거리는 공사 중	한국어	잡지	2004/06	820	×
일본어	jb0411jj	クイズジャパン	일본어	잡지	2004/06	418	×
일본어	jb0411jk	퀴즈재팬	한국어	잡지	2004/06	343	×
일본어	jb0412jj	映画・テレビ業界を目指す 人のための全く新しい学校	일본어	잡지	2004/06	420	×
일본어	jb0412jk	영화·텔레비전 업계를 목 표로 하는 사람을 위한 완전히 새로운 학교	한국어	잡지	2004/06	390	×
일본어	jb0413jj	多文化コミュニケーションを学 ぶ場	일본어	잡지	2004/06	462	×
일본어	jb0413jk	다양한 문화 커뮤니케이 션을 배우는 장	한국어	잡지	2004/06	372	×
일본어	jb0414jj	Only LIVE! LIVE! LIVE!	일본어	잡지	2004/06	974	×
일본어	jb0414jk	Only LIVE! LIVE! LIVE!	한국어	잡지	2004/06	902	×
일본어	jb0415jj	特集	일본어	잡지	2004/06	2,196	×
일본어	jb0415jk	특집	한국어	잡지	2004/06	2,166	×
잡지 종합						20,015	
한국어	ja0401kj	チャイナショック	일본어	신문	2004/05/02	267	-
한국어	ja0401kk	차이나 쇼크	한국어	신문	2004/05/02	239	-
한국어	ja0402kj	マックレーカー	일본어	신문	2004/05/03	252	-
한국어	ja0402kk	머크레이커	한국어	신문	2004/05/03	226	-
한국어	ja0403kj	国益と同盟	일본어	신문	2004/05/04	284	-
한국어	ja0403kk	국익과 동맹	한국어	신문	2004/05/04	251	-
한국어	ja0404kj	宮廷政治	일본어	신문	2004/05/05	287	-
한국어	ja0404kk	궁정정치	한국어	신문	2004/05/05	239	-
한국어	ja0405kj	アテネ・オリンピック	일본어	신문	2004/05/06	261	-
한국어	ja0405kk	아테네 올림픽	한국어	신문	2004/05/06	226	-

제3장 한일 병렬 말뭉치 개발

한국어	ja0406kj	ダイヤモンド	일본어	신문	2004/05/07	237	-
한국어	ja0406kk	다이아몬드	한국어	신문	2004/05/07	215	-
한국어	ja0407kj	墮胎と犯罪	일본어	신문	2004/05/09	286	-
한국어	ja0407kk	낙태와 범죄	한국어	신문	2004/05/09	239	-
한국어	ja0408kj	悪妻分析	일본어	신문	2004/05/10	268	-
한국어	ja0408kk	惡妻 분석	한국어	신문	2004/05/10	253	-
한국어	ja0409kj	8月の砲声	일본어	신문	2004/05/11	266	-
한국어	ja0409kk	8월의 대포	한국어	신문	2004/05/11	261	-
한국어	ja0410kj	アトラス	일본어	신문	2004/05/12	242	-
한국어	ja0410kk	아틀라스	한국어	신문	2004/05/12	242	-
한국어	ja0411kj	性惡説	일본어	신문	2004/05/13	246	-
한국어	ja0411kk	性惡説	한국어	신문	2004/05/13	228	-
한국어	ja0412kj	ヘッジファンド	일본어	신문	2004/05/14	262	-
한국어	ja0412kk	헤지펀드	한국어	신문	2004/05/14	242	-
한국어	ja0413kj	逆境指数 (A Q)	일본어	신문	2004/05/16	267	-
한국어	ja0413kk	역경지수(AQ)	한국어	신문	2004/05/16	233	-
한국어	ja0414kj	自殺予防の音楽	일본어	신문	2004/05/17	284	-
한국어	ja0414kk	자살 예방 음악	한국어	신문	2004/05/17	256	-
한국어	ja0415kj	在韓米軍のイラク派兵	일본어	신문	2004/05/18	307	-
한국어	ja0415kk	미군 차출	한국어	신문	2004/05/18	264	-
한국어	ja0416kj	発明家大統領	일본어	신문	2004/05/19	300	-
한국어	ja0416kk	발명가 대통령	한국어	신문	2004/05/19	264	-
한국어	ja0417kj	シク教 (S i k h)	일본어	신문	2004/05/20	275	-
한국어	ja0417kk	시크(Sikh)	한국어	신문	2004/05/20	232	-
한국어	ja0418kj	サルトシェバデン協約	일본어	신문	2004/05/21	270	-
한국어	ja0418kk	살트셰바덴 협약	한국어	신문	2004/05/21	228	-
한국어	ja0419kj	ほどほどに務められた方	일본어	신문	2004/05/23	286	-
한국어	ja0419kk	어지간히 하신 분	한국어	신문	2004/05/23	233	-
한국어	ja0420kj	ウォッチタワー	일본어	신문	2004/05/24	259	-
한국어	ja0420kk	워치타워	한국어	신문	2004/05/24	250	-
한국어	ja0421kj	日本の二重良心	일본어	신문	2004/05/25	296	-
한국어	ja0421kk	일본의 이중양심	한국어	신문	2004/05/25	271	-
한국어	ja0422kj	移民政策	일본어	신문	2004/05/26	285	-
한국어	ja0422kk	이민 정책	한국어	신문	2004/05/26	242	-
한국어	ja0423kj	チャラビ	일본어	신문	2004/05/27	291	-
한국어	ja0423kk	찰라비	한국어	신문	2004/05/27	229	-
한국어	ja0424kj	トヨタとホンダ	일본어	신문	2004/05/28	266	-
한국어	ja0424kk	도요타와 혼다	한국어	신문	2004/05/28	242	-
한국어	ja0425kj	食教育	일본어	신문	2004/05/30	262	-

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

한국어	ja0425kk	食교육	한국어	신문	2004/05/30	219	-
한국어	ja0426kj	メソの秘訣	일본어	신문	2004/05/31	251	-
한국어	ja0426kk	매추의 비결	한국어	신문	2004/05/31	226	-
한국어	ja0427kj	ある同胞愛	일본어	신문	2004/06/01	294	-
한국어	ja0427kk	어떤 동포애	한국어	신문	2004/06/01	253	-
한국어	ja0428kj	ワハビズム	일본어	신문	2004/06/03	247	-
한국어	ja0428kk	와하비즘	한국어	신문	2004/06/03	228	-
한국어	ja0429kj	イタイタイ病	일본어	신문	2004/06/04	269	-
한국어	ja0429kk	이타이 이타이병	한국어	신문	2004/06/04	255	-
한국어	ja0430kj	美ぼうのプレミアム	일본어	신문	2004/06/06	276	-
한국어	ja0430kk	미모 프리미엄	한국어	신문	2004/06/06	240	-
한국어	ja0431kj	土壤社会	일본어	신문	2004/06/07	253	-
한국어	ja0431kk	토양 사회	한국어	신문	2004/06/07	238	-
한국어	ja0432kj	徐熙將軍	일본어	신문	2004/06/08	306	-
한국어	ja0432kk	서희 장군	한국어	신문	2004/06/08	253	-
한국어	ja0433kj	ノルマンディーのプーチン	일본어	신문	2004/06/09	270	-
한국어	ja0433kk	노르망디의 푸틴	한국어	신문	2004/06/09	252	-
한국어	ja0434kj	レーガン元大統領	일본어	신문	2004/06/10	273	-
한국어	ja0434kk	레이건	한국어	신문	2004/06/10	241	-
신문 종합						17,455	
전체 종합						220,339	

주) ○ : 저작권 승인, × : 저작권 승인 안 됨, - : 저작권과 관련 없는 자료

5. 말뭉치 지도

2001년도부터 금년도 2004년도까지 한일병렬 말뭉치 구축 분과에서 구축한 말뭉치 구축량을 도표로 나타내면 다음과 같다.

말뭉치 자료의 선정에는 원본 자료의 입수 및 저작권 해결 등 여러 가지 문제들이 있는 것도 사실이다. 앞으로의 한일병렬 말뭉치 구축 시에는 이 말뭉치 지도의 구축량 분포를 토대로 보다 균형성이 있는 다양한 장르의 말뭉치를 구축할 계획이다.

제3장 한일 병렬 말뭉치 개발

구축연도	2단계 (2001-2003)												3단계 (2004-2007)				장르별 합계	장르별 비율
	2001				2002				2003				2004					
	원본		대역본		원본		대역본		원본		대역본		원본		대역본			
언어 구분	K	J	K	J	K	J	K	J	K	J	K	J	K	J	K	J		
(a)신문	20,352			32,259	18,471			26,745	17,319			14,104	8,210			9,245	146,705	17.75%
(b)잡지					17,999			20,220	13,433			10,388		10,338	9,677		82,055	9.93%
(c)책-교과서																	0	0%
(d)책-사전																	0	0%
(e)책-소설		27,847	14,104		22,658	2,010	1,592	28,777		41,401	37,171			48,513	52,909		276,982	33.52%
(f)책-시																	0	0%
(g)책-기타		1,347	870														2,217	0.27%
(h)책-정보		200	111							9,081	9,596						18,988	2.30%
(i)기타 출판물	37,796			40,364		24,027	18,842			22,258	22,234						165,521	20.03%
(j)기타 비출판물		4918	3,931						15,335	4,898	3,363	19,974	25,297			26,578	104,294	12.62%
(k)시나리오														15,334	14,238		29,572	3.58%
언어별 합계	58,148	34,312	19,016	72,623	59,128	26,037	20,434	75,742	46,087	77,638	72,364	44,466	33,507	74,185	76,824	35,823	826,334	100%
총합	184,099				181,341				240,555				220,339				826,334	

6. 한일 병렬 말뭉치 구축의 실례

6.1 일본어 원시 말뭉치 (잡지 jb0406jj.txt)

```
<cesDoc version=3.24>
<cesHeader version="2.3">
  <fileDesc>
    <titleStmt>
      <h.title>잡지/おにぎり! </h.title>
      <respStmt>
        <respType>문헌입력, 표준화, 헤더붙임</respType>
        <respName>고려대학교 일어일문학과 한일병렬말뭉치 구축팀
      </respName>
    </respStmt>
  </titleStmt>
  <extent>
    <wordcount>541</wordcount>
    <bytecount>4096</bytecount>
  </extent>
  <publicationStmt>
    <contributor>국립국어원</contributor>
    <pubAddress>서울시 강서구 방화3동 827번지</pubAddress>
    <availability status=restricted>배포불가</availability>
    <PubDate>2004</PubDate>
  </publicationStmt>
  <sourceDesc>
    <biblStruct>
      <monogr>
```

```

<h.title>おにぎり！</h.title>
<h.author></h.author>
<imprint>
  <pubPlace>日本</pubPlace>
  <publisher>月刊留学生</publisher>
  <pubDate>2004/05</pubDate>
</imprint>
</monogr>
</biblStruct>
</sourceDesc>
</fileDesc>
<encodingDesc>
  <projectDesc>21세기 세종계획 3단계 1차년도 한일병렬 말뭉치구축
</projectDesc>
  <editorialDecl>21세기 세종계획 말뭉치문헌입력 지침에 따름
</editorialDecl>
  <samplingDecl>おにぎり！를 전자파일화</samplingDecl>
  <refsDecl></refsDecl>
<encodingDesc>
<profileDesc>
  <creation><date>2004/07</date></creation>
  <langUsage>
    <language id="kr" iso639="kr"></language>
  </langUsage>
  <wsdUsage>
    <writingSystem id="UTF-16"></writingSystem>
  </wsdUsage>
  <textClass>
    <catRef scheme='SJ21' target="R1131">잡지</catRef>

```



```
</textClass>
<translation>
  <translation trans.loc="jb0406jk.txt" lang="kr" wsd="UTF-16">
    <translator></translator>
  </translation>
</profileDesc>
<revisionDesc>
  <change>
    <changeDate>2004/07</changeDate>
    <respName>최장원</respName>
    <h.item>본문 마크업</h.item>
  </change>
</revisionDesc>
</cesHeader>
<text>
<body id=1>
<div id=1.1>
<head id=1.1.h1>おにぎり! </head>
<p id=1.1.p1><s id=1.1.p1.s1>日本のどのコンビニに行っても、食品コーナーには必ず
おにぎりが陳列されている。</s><s id=1.1.p1.s2>留学生も、誰もが1度はコンビニでお
にぎりを買って食べた経験があるだろう。</s><s id=1.1.p1.s3>しかし、日本に来て初め
ておにぎりを買った時、韓国と違った多様な包装といくつもの種類を前にして、何を食べたら
いいのか、選ぶのにずいぶん迷った記憶もあるだろう。</s></p>
<p id=1.1.p2><s id=1.1.p2.s1>忙しい現代社会を生きる人達にとって、いつでも栄養と
味を提供してくれ、簡単に食べられる便利さまでプラスしてくれるおにぎりは、これからもずっと
無窮無尽の発展の可能性を持つ、お弁当の代表的存在であると信じて疑わない。
</s></p>
<p id=1.1.p3><s id=1.1.p3.s1>さあ、今回は、私達の日常の中で、常に親しい存在と
してお腹を満たしてくれるおにぎりについて、考えてみてはどうだろうか? </s></p>
```

<p id=1.1.p4><s id=1.1.p4.s1><おにぎりの種類></s></p>
<p id=1.1.p5><s id=1.1.p5.s1>梅</s></p>
<p id=1.1.p6><s id=1.1.p6.s1>梅干しを食べてみたことがある人なら、言わなくても分かるだろう。</s><s id=1.1.p6.s2>最も日本の香りがすると言える。</s><s id=1.1.p6.s3>酸っぱくて刺激的な味は、食欲のない人にうってつけ。</s></p>
<p id=1.1.p7><s id=1.1.p7.s1>昆布</s></p>
<p id=1.1.p8><s id=1.1.p8.s1>昆布を醤油と砂糖で佃煮にしたもので、甘くて美味しい。</s><s id=1.1.p8.s2>老若男女誰からも好まれる、人気メニュー！</s></p>
<p id=1.1.p9><s id=1.1.p9.s1>おかか</s></p>
<p id=1.1.p10><s id=1.1.p10.s1>かつお節という、うどんにも入れているものが入っている。</s><s id=1.1.p10.s2>少しパサパサしているが、日本的な味が感じられる。</s></p>
<p id=1.1.p11><s id=1.1.p11.s1>ツナ</s></p>
<p id=1.1.p12><s id=1.1.p12.s1>韓国で売っているおにぎり（チュモクパプ）のチャムチ & マヨネーズと味が似ていて、一番口にあう。</s><s id=1.1.p12.s2>シーチキンと表記されているものもある。</s></p>
<p id=1.1.p13><s id=1.1.p13.s1>タラコ</s></p>
<p id=1.1.p14><s id=1.1.p14.s1>基本的に魚の卵で作ってあるものは値段が高い。</s><s id=1.1.p14.s2>しかし高いだけに美味しい。</s><s id=1.1.p14.s3>一般的に「タラコおにぎり」のタラコは焼いたものが入っている。</s></p>
<p id=1.1.p15><s id=1.1.p15.s1>明太子</s></p>
<p id=1.1.p16><s id=1.1.p16.s1>明太子はいろいろな料理に使われていて、独特なうま味がある。</s><s id=1.1.p16.s2>ピリッとして刺激的だが、一度好きになったら止まらない！</s><s id=1.1.p16.s3>特にミョンランチョッ（明太子の塩辛）が好きな人にオススメ。</s></p>
<p id=1.1.p17><s id=1.1.p17.s1>鮭</s></p>
<p id=1.1.p18><s id=1.1.p18.s1>鮭は日本人がよく食べる魚のひとつ。</s><s id=1.1.p18.s2>鮭おにぎりには焼いた鮭がごろっと入っているものと、鮭フレークが入っているものがあり、安い鮭おにぎりの場No.1大抵フレークだ。</s></p>

<p id=1.1.p19><s id=1.1.p19.s1>赤飯</s></p>

<p id=1.1.p20><s id=1.1.p20.s1>あずきともち米で炊いたご飯で、特別な味付けはせず、それだけでとても美味しい。</s></p>

<p id=1.1.p21><s id=1.1.p21.s1>おにぎりの正体</s></p>

<p id=1.1.p22><s id=1.1.p22.s1>・ 名前の正体から知っておこう</s></p>

<p id=1.1.p23><s id=1.1.p23.s1>日本語の「握る」という動詞が名詞化して、「握り」となったもの。</s><s id=1.1.p23.s2>そして、食べ物の前によく使われる接頭語「お」を付けて「おにぎり」となった。</s><s id=1.1.p23.s3>同じ意味で「おむすび」とも呼ばれるが、やはり同様に「むすぶ」という動詞が名詞になったもの。</s><s id=1.1.p23.s4>「おにぎり」と「おむすび」、意味的には全く違いはなく、ただ地方によって呼び方が違うだけだ。</s><s id=1.1.p23.s5>主に関東地方（東京側）では「おにぎり」と呼び、関西地方（大阪側）では「おむすび」と呼ぶ場合が多いという。</s><s id=1.1.p23.s6>また、千葉県では大きく作ったものをおむすび、小さいものをおにぎりと言うのだとか。</s></p>

<p id=1.1.p24><s id=1.1.p24.s1>・ 果たして味の正体は？</s></p>

<p id=1.1.p25><s id=1.1.p25.s1>握りこぶしくらいの大きさに米を握りかため、塩で味をつけたり、海苔を巻いて作る簡単な料理方法の食品で、種類は数十種に及び、今も新しいメニューが登場し続けている。</s></p>

<p id=1.1.p26><s id=1.1.p26.s1>おにぎりの歴史</s></p>

<p id=1.1.p27><s id=1.1.p27.s1>おにぎりは、約1000年前、京都が都であった平安時代から存在していたと推定されているが、本格的なおにぎりの歴史は、豊臣秀吉が天下を統一した戦国時代からだという。</s><s id=1.1.p27.s2>武士達が携帯食品と</s></p>

<p id=1.1.p28><s id=1.1.p28.s1>して、また、非常食として利用したものが、今日のおにぎりの起源だと言えるのだが、今とは異なりその当時は、長期の保存を目的として、炒めたご飯や乾燥させたご飯を使ったという。</s><s id=1.1.p28.s2>今のように白米のご飯でおにぎりを作るようになったのは明治時代（19世紀初め）、厳密に言えば戦後である1950年代初期と、最近のことである。</s></p>

<p id=1.1.p29><s id=1.1.p29.s1>1978年には現在一般的に見られる、パリパリの手巻

きおにぎりが誕生、1980年には初めてツナマヨネーズが登場し、おにぎりの中でも最高の人気となる。</s></p>

<p id=1.1.p30><s id=1.1.p30.s1>最近では外食産業として、その規模がますます大きくなり、百貨店や専門食堂などでも活発に販売されており、一般家庭で作って食べる、お弁当の概念を越えている。</s></p>

<p id=1.1.p31><s id=1.1.p31.s1>日本のおにぎり事情</s></p>

<p id=1.1.p32><s id=1.1.p32.s1>市販されているおにぎりは、コンビニだけでも年間約20億個！</s><s id=1.1.p32.s2>1日のおにぎりの販売数は、東京だけでも平均700万個以上と、日本人の食生活において絶対的な存在だ。</s><s id=1.1.p32.s3>コンビニの間でも互いに競争が起こり、包装を変えた高級おにぎりや、独特な材料のおにぎりを開発・発売している。</s><s id=1.1.p32.s4>また、お菓子やスナック類にも登場している。</s><s id=1.1.p32.s5>一時はおにぎりの好みで性格や将来まで占う、おにぎり占いが若者の間で流行したことも。</s></p>

<p id=1.1.p33><s id=1.1.p33.s1>日頃は何も考えずに買って食べていたおにぎりの種類や形など、その多彩ぶりを知てみると、何故かますます食べたくなる。</s><s id=1.1.p33.s2>今までにおにぎりとは仲良くしていなかった留学生達は、今日、コンビニに寄って1つ買ってみよう。</s><s id=1.1.p33.s3>好みのおにぎりはどれだろう？</s></p>

</div>

</body>

</text>

</cesDoc>

6.2 한국어 원시 말뭉치 (잡지 jb0406jk.txt)

<cesDoc version=3.24>

<cesHeader version="2.3">

<fileDesc>

<titleStmt>

<h.title>잡지/오니기리!</h.title>

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

```
<respStmt>
  <respType>문헌입력, 표준화, 헤더붙임</respType>
  <respName>고려대학교   일어일문학과   한일병렬말뭉치   구축팀
</respName>
</respStmt>
</titleStmt>
<extent>
  <wordcount>537</wordcount>
</extent>
<publicationStmt>
  <distributor>국립국어원</distributor>
  <pubAddress>서울시 강서구 방화3동 827번지</pubAddress>
  <availability status=restricted>배포불가</availability>
  <PubDate>2004</PubDate>
</publicationStmt>
<sourceDesc>
  <biblStruct>
    <monogr>
      <h.title>오니기리!</h.title>
      <h.author></h.author>
      <imprint>
        <pubPlace>일본</pubPlace>
        <publisher>월간유학생</publisher>
        <pubDate>2004/05</pubDate>
      </imprint>
    </monogr>
  </biblStruct>
</sourceDesc>
</fileDesc>
```

```

<encodingDesc>
  <projectDesc>21세기 세종계획 3단계 1차년도 한일병렬 말뭉치구축
</projectDesc>
  <editorialDecl>21세기 세종계획 말뭉치문헌입력 지침에 따름
</editorialDecl>
  <samplingDecl>오니거리!를 전자파일화</samplingDecl>
  <refsDecl></refsDecl>
<encodingDesc>
<profileDesc>
  <creation><date>2004/07</date></creation>
  <langUsage>
    <language id="kr" iso639="kr"></language>
  </langUsage>
  <wsdUsage>
    <writingSystem id="UTF-16"></writingSystem>
  </wsdUsage>
  <textClass>
    <catRef scheme='SJ21' target="R1131">잡지</catRef>
  </textClass>
  <translation>
    <translation trans.loc="jb0406jj.txt" lang="jp" wsd="UTF-16">
      <translator></translator>
    </translation>
  </translation>
</profileDesc>
<revisionDesc>
  <change>
    <changeDate>2004/07</changeDate>
    <respName>최장원</respName>
    <h.item>본문 마크업</h.item>
  </change>
</revisionDesc>

```

```
</change>
</revisionDesc>
</cesHeader>
<text>
<body id=1>
<div id=1.1>
<head id=1.1.h1>오니기리!</head>
<p id=1.1.p1><s id=1.1.p1.s1>일본의 어느 편의점에도 가더라도 식품코너에는 반드시 오니기리가 진열되어 있다.</s><s id=1.1.p1.s2>그렇기 때문에 유학생 누구나 한 번 이상은 편의점에서 오니기리를 사 먹어 본 경험이 있을 것이다.</s><s id=1.1.p1.s3>그러나 일본에 와서 맨 처음 오니기리를 샀을 때, 한국과 달리 다양한 포장과 여러 종류 앞에 무엇을 먹어야 할 지 고르는데 무척 망설였던 기억이 있을 것이다.</s></p>
<p id=1.1.p2><s id=1.1.p2.s1>바쁜 현대 사회를 살아가는 사람들에게 언제나 영양과 맛을 제공해 주고 간편하게 먹을 수 있어 편리함까지 더해주는 오니기리는 앞으로도 계속해서 무궁무진한 발전의 가능성을 가진 도시락의 대표적 존재임을 믿어 의심치 않는데...</s></p>
<p id=1.1.p3><s id=1.1.p3.s1>자, 이번 월간 유학생 5월호 [ 좀더 알고 싶어요 ]에서 우리 일상 속에서 항상 친근하게 배를 채워주는 오니기리에 대해 생각해 보면 어떨까?</s></p>
<p id=1.1.p4><s id=1.1.p4.s1><오니기리의 종류></s></p>
<p id=1.1.p5><s id=1.1.p5.s1>우메(매실 짬아치)</s></p>
<p id=1.1.p6><s id=1.1.p6.s1>우메보시를 먹어 본 사람은 말 안해도 알 것이다.</s><s id=1.1.p6.s2>가장 일본냄새가 풍긴다고 할 수 있다.</s><s id=1.1.p6.s3>새콤하고 자극적인 맛이 특히 입맛이 없는 사람에 좋다.</s></p>
<p id=1.1.p7><s id=1.1.p7.s1>콘부(다시마)</s></p>
<p id=1.1.p8><s id=1.1.p8.s1>다시마를 간장과 설탕에 조린 것으로 달콤하고 맛있다.</s><s id=1.1.p8.s2>남녀노소 인기 메뉴!</s></p>
<p id=1.1.p9><s id=1.1.p9.s1>오카카(카쯔오)</s></p>
```

<p id=1.1.p10><s id=1.1.p10.s1>‘카즈오부시’라고 하는, 가끔 우동에도 넣어 먹는 가다랭이포가 들어있다.</s><s id=1.1.p10.s2>조금 쫄쫄하지만 일본적인 맛을 느낄 수 있다.</s></p>

<p id=1.1.p11><s id=1.1.p11.s1>쯔나(참치)</s></p>

<p id=1.1.p12><s id=1.1.p12.s1>한국에서 파는 주먹밥 <참치&마요네즈>와 맛이 비슷하여 제일 입맛에 맞다.</s></p>

<p id=1.1.p13><s id=1.1.p13.s1>타라코 (대구알)</s></p>

<p id=1.1.p14><s id=1.1.p14.s1>기본적으로 생선알로 만든 것은 비싼 편이다.</s><s id=1.1.p14.s2>그러나 비싼 만큼 맛있다.</s><s id=1.1.p14.s3>일반적으로 「타라코 주먹밥」의 타라코는 구운 것이 들어가 있다.</s></p>

<p id=1.1.p15><s id=1.1.p15.s1>멘타이코(명태알)</s></p>

<p id=1.1.p16><s id=1.1.p16.s1>멘타이코는 피자, 스파게티 소스 등 많은 요리에 사용되는 만큼 독특한 감칠맛이 있다.</s><s id=1.1.p16.s2>조금 씹쓸하고 자극적이지만 한 번 좋아하게 되면 멈출 수 없다!</s><s id=1.1.p16.s3>특히 명란젓을 좋아하는 사람들에게 추천.</s></p>

<p id=1.1.p17><s id=1.1.p17.s1>사케·샤케(연어)</s></p>

<p id=1.1.p18><s id=1.1.p18.s1>연어는 일본인들이 자주 먹는 생선 중 하나로 고소하고 담백하다.</s></p>

<p id=1.1.p19><s id=1.1.p19.s1>세키항(팔밥)</s></p>

<p id=1.1.p20><s id=1.1.p20.s1>팔과 찹쌀로 지은 밥으로 특별한 양념없이 그 자체로도 꽤 맛있다.</s></p>

<p id=1.1.p21><s id=1.1.p21.s1>오니기리의 정체</s></p>

<p id=1.1.p22><s id=1.1.p22.s1>①이름의 정체부터 알아보자.</s></p>

<p id=1.1.p23><s id=1.1.p23.s1>일본어로 “취다”, “잡다”란 뜻의 “니기루”라는 동사가 명사화 되어 “니기리”가 되었다.</s><s id=1.1.p23.s2>그리고 음식 앞에 자주 사용하는 접두어 “오”를 붙여 “오니기리”가 되었다.</s><s id=1.1.p23.s3>한 편 똑같은 의미로 “오무스비”라고도 불리우는데 역시 마찬가지로 “무스부”라는 동사가 명사화 된 것으로, 즉 주먹밥이란 뜻이다.</s><s id=1.1.p23.s4>“오니기리”와 “오무스비”의 차이는 의미적인 면에서는 전혀 차이

가 없고 단지 지방에 따라 호칭 방법이 다른 것 뿐이다.</s><s id=1.1.p23.s5> 주로 관동지방 (도쿄쪽)에서는 “오니기리”라 부르고 관서지방(오사카쪽)에서는 “오무스비”라고 부르는 경우가 많다고 한다.</s><s id=1.1.p23.s6>또한 치바현에서는 크게 만든 것을 무스비, 작은 것을 오니기리라 한다고 한다.</s></p>

<p id=1.1.p24><s id=1.1.p24.s1>②과연 맛의 정체는?</s></p>

<p id=1.1.p25><s id=1.1.p25.s1>주먹만한 크기로 밥을 주물러 소금으로 간을 하거나 김을 말아서 만드는 간편한 요리방법의 종류는 수십 개에 달하고 지금도 계속해서 새로운 메뉴가 등장하고 있다.</s></p>

<p id=1.1.p26><s id=1.1.p26.s1>오니기리의 역사</s></p>

<p id=1.1.p27><s id=1.1.p27.s1>오니기리는 약 1000년 전 교토가 수도였던 헤이안 시대부터 존재했던 것으로 추정되고 있는데, 본격적인 오니기리의 역사는 도요토미 히데요시가 전국을 통일했던 전국시대(무사시대)부터라고 한다.</s><s id=1.1.p27.s2>무사들이 휴대식품으로,</s></p>

<p id=1.1.p28><s id=1.1.p28.s1>또는 비상식량으로 이용 한 것이 오늘날 오니기리의 기원이라 할 수 있는데 지금과는 달리 그 당시엔 오랜 저장을 목적으로 볶은 밥이나 말린 밥을 사용하였다고 한다.</s><s id=1.1.p28.s2>지금처럼 흰 쌀밥으로 오니기리를 만들기 시작한 것은 메이지 시대(19세기 초)이며, 엄밀히 말해 전쟁 이후 인 1950년대 초기로 최근의 일이다.</s></p>

<p id=1.1.p29><s id=1.1.p29.s1>1978년에는 현재 일반적으로 볼 수 있는 바삭 바삭한 테마키 오니기리가 탄생하였고, 1980년에는 최초로 [쯔나 마요네즈]가 등장하여 오니기리 중에서도 최고의 인기를 얻고 있다.</s></p>

<p id=1.1.p30><s id=1.1.p30.s1>최근에는 외식사업으로써 그 규모가 더욱 커져 백화점이나 전문 식당 등에서도 활발히 판매되고 있어 보통 집에서 만들어 먹는 도시락의 개념을 넘어서고 있다.</s></p>

<p id=1.1.p31><s id=1.1.p31.s1>일본의 오니기리 사정</s></p>

<p id=1.1.p32><s id=1.1.p32.s1>편의점에서만 팔리는 오니기리는 연간 약 20억 개!</s><s id=1.1.p32.s2>도쿄에서만 하루 오니기리의 판매수는 평균 700만 개 이상으로 일본인 식생활에 절대적인 존재이다.</s><s id=1.1.p32.s3>편의점들 사이에서도 서로 경쟁이 붙어 포장을 달리 한 고급 오니기리와 독특한 재료의

오니기리를 개발하여 내놓고 있으며 심지어 과자와 스낵류로까지 나오고 있다.</s><s id=1.1.p32.s4>한 때 오니기리 취향에 따라 성격과 장래까지 점치는 오니기리점(우라나이)이 젊은이들 사이에서 유행하기도 했다.</s></p>

<p id=1.1.p33><s id=1.1.p33.s1>평소에는 아무 생각 없이 사먹던 오니기리의 종류나 모양 등의 그 다양함을 알게 되니 왠지 더욱 먹고 싶어진다.</s><s id=1.1.p33.s2>아직까지 오니기리와 친하지 않은 유학생들은 오늘 편의점에 들러 한두 개 사 보자.</s><s id=1.1.p33.s3>과연 내 입맛에 맞는 오니기리는?</s></p>

</div>

</body>

</text>

</cesDoc>

6.3 정렬 파일 (잡지 jb0406al.txt)

<cesAlign type=sent version=1.6>

<cesHeader version="2.3">

<fileDesc>

<titleStmt>

<h.title>잡지/おにぎり!</h.title>

<publicationStmt>

<distributor>국립국어원</distributor>

<pubAddress>서울시 강서구 방화3동 827번지</pubAddress>

<availability status=restricted>배포 불가</availability>

<PubDate>2003</PubDate>

</publicationStmt>

<sourceDesc>

<biblStruct>

<monogr>

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

```
<h.title>おにぎり!</h.title>
<h.author></h.author>
<imprint>
    <pubPlace>日本</pubPlace>
    <publisher>月刊留学生</publisher>
    <pubDate>2004/05</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">
    <link xtargets="1.1.p1.s2 ; 1.1.p1.s2">
    <link xtargets="1.1.p1.s3 ; 1.1.p1.s3">
</linkGrp>
<linkGrp domains="1.1.p2 ; 1.1.p2">
    <link xtargets="1.1.p2.s1 ; 1.1.p2.s1">
</linkGrp>
<linkGrp domains="1.1.p3 ; 1.1.p3">
    <link xtargets="1.1.p3.s1 ; 1.1.p3.s1">
</linkGrp>
<linkGrp domains="1.1.p4 ; 1.1.p4">
    <link xtargets="1.1.p4.s1 ; 1.1.p4.s1">
```

```

</linkGrp>
<linkGrp domains="1.1.p5 ; 1.1.p5">
    <link xtargets="1.1.p5.s1 ; 1.1.p5.s1">
</linkGrp>
<linkGrp domains="1.1.p6 ; 1.1.p6">
    <link xtargets="1.1.p6.s1 ; 1.1.p6.s1">
    <link xtargets="1.1.p6.s2 ; 1.1.p6.s2">
    <link xtargets="1.1.p6.s3 ; 1.1.p6.s3">
</linkGrp>
<linkGrp domains="1.1.p7 ; 1.1.p7">
    <link xtargets="1.1.p7.s1 ; 1.1.p7.s1">
</linkGrp>
<linkGrp domains="1.1.p8 ; 1.1.p8">
    <link xtargets="1.1.p8.s1 ; 1.1.p8.s1">
    <link xtargets="1.1.p8.s2 ; 1.1.p8.s2">
</linkGrp>
<linkGrp domains="1.1.p9 ; 1.1.p9">
    <link xtargets="1.1.p9.s1 ; 1.1.p9.s1">
</linkGrp>
<linkGrp domains="1.1.p10 ; 1.1.p10">
    <link xtargets="1.1.p10.s1 ; 1.1.p10.s1">
    <link xtargets="1.1.p10.s2 ; 1.1.p10.s2">
</linkGrp>
<linkGrp domains="1.1.p11 ; 1.1.p11">
    <link xtargets="1.1.p11.s1 ; 1.1.p11.s1">
</linkGrp>
<linkGrp domains="1.1.p12 ; 1.1.p12">
    <link xtargets="1.1.p12.s1 ; 1.1.p12.s1">
    <link xtargets="1.1.p12.s2 ; ">

```

```
</linkGrp>
<linkGrp domains="1.1.p13 ; 1.1.p13">
  <link xtargets="1.1.p13.s1 ; 1.1.p13.s1">
</linkGrp>
<linkGrp domains="1.1.p14 ; 1.1.p14">
  <link xtargets="1.1.p14.s1 ; 1.1.p14.s1">
  <link xtargets="1.1.p14.s2 ; 1.1.p14.s2">
  <link xtargets="1.1.p14.s3 ; 1.1.p14.s3">
</linkGrp>
<linkGrp domains="1.1.p15 ; 1.1.p15">
  <link xtargets="1.1.p15.s1 ; 1.1.p15.s1">
</linkGrp>
<linkGrp domains="1.1.p16 ; 1.1.p16">
  <link xtargets="1.1.p16.s1 ; 1.1.p16.s1">
  <link xtargets="1.1.p16.s2 ; 1.1.p16.s2">
  <link xtargets="1.1.p16.s3 ; 1.1.p16.s3">
</linkGrp>
<linkGrp domains="1.1.p17 ; 1.1.p17">
  <link xtargets="1.1.p17.s1 ; 1.1.p17.s1">
</linkGrp>
<linkGrp domains="1.1.p18 ; 1.1.p18">
  <link xtargets="1.1.p18.s1 ; 1.1.p18.s1">
  <link xtargets="1.1.p18.s2 ; ">
</linkGrp>
<linkGrp domains="1.1.p19 ; 1.1.p19">
  <link xtargets="1.1.p19.s1 ; 1.1.p19.s1">
</linkGrp>
<linkGrp domains="1.1.p20 ; 1.1.p20">
  <link xtargets="1.1.p20.s1 ; 1.1.p20.s1">
```

```
</linkGrp>
<linkGrp domains="1.1.p21 ; 1.1.p21">
  <link xtargets="1.1.p21.s1 ; 1.1.p21.s1">
</linkGrp>
<linkGrp domains="1.1.p22 ; 1.1.p22">
  <link xtargets="1.1.p22.s1 ; 1.1.p22.s1">
</linkGrp>
<linkGrp domains="1.1.p23 ; 1.1.p23">
  <link xtargets="1.1.p23.s1 ; 1.1.p23.s1">
  <link xtargets="1.1.p23.s2 ; 1.1.p23.s2">
  <link xtargets="1.1.p23.s3 ; 1.1.p23.s3">
  <link xtargets="1.1.p23.s4 ; 1.1.p23.s4">
  <link xtargets="1.1.p23.s5 ; 1.1.p23.s5">
  <link xtargets="1.1.p23.s6 ; 1.1.p23.s6">
</linkGrp>
<linkGrp domains="1.1.p24 ; 1.1.p24">
  <link xtargets="1.1.p24.s1 ; 1.1.p24.s1">
</linkGrp>
<linkGrp domains="1.1.p25 ; 1.1.p25">
  <link xtargets="1.1.p25.s1 ; 1.1.p25.s1">
</linkGrp>
<linkGrp domains="1.1.p26 ; 1.1.p26">
  <link xtargets="1.1.p26.s1 ; 1.1.p26.s1">
</linkGrp>
<linkGrp domains="1.1.p27 ; 1.1.p27">
  <link xtargets="1.1.p27.s1 ; 1.1.p27.s1">
  <link xtargets="1.1.p27.s2 ; 1.1.p27.s2">
</linkGrp>
<linkGrp domains="1.1.p28 ; 1.1.p28">
```

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

```
<link xtargets="1.1.p28.s1 ; 1.1.p28.s1">
  <link xtargets="1.1.p28.s2 ; 1.1.p28.s2">
</linkGrp>
<linkGrp domains="1.1.p29 ; 1.1.p29">
  <link xtargets="1.1.p29.s1 ; 1.1.p29.s1">
</linkGrp>
<linkGrp domains="1.1.p30 ; 1.1.p30">
  <link xtargets="1.1.p30.s1 ; 1.1.p30.s1">
</linkGrp>
<linkGrp domains="1.1.p31 ; 1.1.p31">
  <link xtargets="1.1.p31.s1 ; 1.1.p31.s1">
</linkGrp>
<linkGrp domains="1.1.p32 ; 1.1.p32">
  <link xtargets="1.1.p32.s1 ; 1.1.p32.s1">
  <link xtargets="1.1.p32.s2 ; 1.1.p32.s2">
  <link xtargets="1.1.p32.s3 s4 ; 1.1.p32.s3">
  <link xtargets="1.1.p32.s5 ; 1.1.p32.s4">
</linkGrp>
<linkGrp domains="1.1.p33 ; 1.1.p33">
  <link xtargets="1.1.p33.s1 ; 1.1.p33.s1">
  <link xtargets="1.1.p33.s2 ; 1.1.p33.s2">
  <link xtargets="1.1.p33.s3 ; 1.1.p33.s3">
</linkGrp>
</linkGrp>
</linkGrp>
</linkList>
```

제 4 장

북한 및 해외 한국어 말뭉치 개발

연구책임자 : 소 강 춘

여 백

제 4 장

북한 및 해외 한국어 말뭉치 개발

1. 연구의 필요성, 목표, 대상

1.1 연구의 필요성

지금까지 추진된 세종계획 2단계 6개년 사업은 국내의 국어국문학 연구자들에게는 물론 북한 및 해외에서 한국어를 전공하고 연구하는 모든 학자들이나 기관에 커다란 기여를 하고 있다.

특히 특수자료 분과의 연구물들은 국내에서도 새로운 연구 방향 설정을 위한 기초자료가 되지만 해외에서 한국어를 연구하고 있는 사람들에게는 한국어와 외국어의 관계를 파악하는 데 많은 장점을 가지고 있다.

남북한과 해외 한국어 말뭉치 구축사업은 북한, 중국, 구 소련의 카자흐스탄에 이르기까지 그 영역이 확대되고 있는 실정이다. 현지 연구들은 국어정보화에 대한 생각은 가지고 있지만, 체계적인 지원이 이루어지지 않기 때문에 추진하지 못하는 작업을 우리가 추진하고 그 결과를 제공 받음으로 연구에 많은 도움을 받고 있다.

따라서 앞으로도 양질의 데이터를 더 많이 구축하는 것이 우리의 할 일이라고 생각한다.

1.2 연구 목표

이 연구는 1998년부터 2007년까지 추진되는 21세기 세종계획(국어 기초자료 구축 분과 특수자료구축 소분과)의 3 단계 사업 중 1차년도에 해당하는 '북한 및 해외 한국어 말뭉치' 분야의 7차년도 연구과제이다.

금년도 사업은 북한과 중국에서 간행되어 현재 사용하고 있는 국어 또는 조선어 교과서 자료를 기초말뭉치로 구축하고, 카자흐스탄의 1994년도 고려일보를 구축 대상으로 하고 있다.

이번에 교과서 자료를 구축하게 된 것은 그동안 북한에서 교과서 자료는 대외 비로 간주하여 결코 외적으로 공개하지 않았기 때문에 자료 구입에 한계를 가지고 있었다. 그런데 이번에 북한의 2000년도 이후에 개정되어 사용되고 있는 교과서 전체를 구할 수 있어 이를 구축하게 된 것이다.

중국의 교과서를 구축한 것은 남한, 북한, 중국의 교과서 비교나 연구에 도움을 주기 위해서 해외 한국어 자료로 중국에서 사용되고 있는 '의무교육소학교 조선어문' 12권을 모두 구축하게 되었다.

주석말뭉치도 북한과 중국의 소학교 국어교과서에 대한 주석을 포함시켰다.

금년도 원시말뭉치 구축 목표는 100만 어절이고, 주석말뭉치는 35만 어절이다.

1.3 연구 범위 및 대상

금년도 연구 범위와 대상은 북한과 중국, 카자흐스탄 세 지역 자료이다. 먼저 원시말뭉치의 북한 자료는 매년 구축했던 신문 자료와 2000년도 이후에 발간되어 현재 교재로 사용하고 있는 북한의 소학교와 고등중학교 교과서를 대상으로 하고 있다. 아래 구축 결과에서 볼 수 있는 바와 같이 이번에는 국어, 역사, 사상 등에 치우친 자료를 구축 했다. 중국 자료는 북한과 남한의 대상 언어와 비교 연구를 하고자 하는 연구자들을 위해서 즉, 현재 각 국에서 사용되고 있는 가장 기본적인 어휘와 문장 등을 연구하기 위해서는 교과서의 비교가 필요하기 때문에, 중국 조선족 소학교에서 사용하고 있는 '의무교육소학교 조선어문' 12권을 원시말뭉치로

구축했다. 카자흐스탄 자료는 1993년과 1994년에 표기법이 변했기 때문에 작년에 이어서 1994년 고려일보 전체를 구축했다.

주석말뭉치는 전년도에 이어서 소설, 신문, 전문 잡지 내용을 추가했고, 금년도 부터 새롭게 교과서를 주석말뭉치에 포함시켰다.

2. 북한 말뭉치 구축 결과

1) 원시말뭉치 구축 목표 : 100만 어절

- 구축현황 :	북한 문화어	625,265 어절
	중국 조선어	137,170 어절
	카자흐스탄	267,611 어절
총 :		1,030,046 어절

2) 내용 :

2004년 북한 및 해외 한국어 말뭉치 개발에서는 원시 말뭉치 100만 어절 구축을 목표로 하고 있으며, 아래에 기 구축된 자료와 별도의 목록에서 균형을 고려하여 북한 자료의 균형성을 확보한다.

북한 이외의 자료는 2003년도부터 카자흐스탄의 알마아타에서 발행되고 있는 '고려일보'의 1993년도 분을 입력했고, 금년에는 1994년도 자료를 입력했다. 이는 1993년과 1994년에 이 지역에서 정서법 규정이 바뀌었기 때문이다. 즉, 1993년까지는 북한의 정서법을 따르다가 1994년부터는 우리의 정서법을 따르고 있기 때문이다.

중국의 소학교 6학년까지의 조선어 교과서 자료를 구축하였다. 이는 남한, 북한, 중국의 최근 교과서 자료를 동시에 구축하기 위한 것이다.

3) 목록 :

[카자흐스탄 말뭉치]

파일명	제목	발간 연도	어절수
▪ 7RAK94A0.HWP	고려일보	1994	267,611
			총 : 267,611어절

[중국 말뭉치]

파일명	제목	발간 연도	어절수
▪ 7OCAA901.HWP	의무교육소학교 조선어문 1권	1999	2,688
▪ 7OCAA902.HWP	의무교육소학교 조선어문 2권	1999	6,446
▪ 7OCAA903.HWP	의무교육소학교 조선어문 3권	1999	8,376
▪ 7OCAA904.HWP	의무교육소학교 조선어문 4권	1999	8,936
▪ 7OCAA905.HWP	의무교육소학교 조선어문 5권	1999	9,543
▪ 7OCAA906.HWP	의무교육소학교 조선어문 6권	1999	11,965
▪ 7OCAA907.HWP	의무교육소학교 조선어문 7권	1999	12,906
▪ 7OCAA908.HWP	의무교육소학교 조선어문 8권	1999	13,416
▪ 7OCAA909.HWP	의무교육소학교 조선어문 9권	1999	16,430
▪ 7OCAA910.HWP	의무교육소학교 조선어문 10권	1999	15,282
▪ 7OCAA911.HWP	의무교육소학교 조선어문 11권	1999	15,374
▪ 7OCAA912.HWP	의무교육소학교 조선어문 12권	1999	15,808
			총 : 137,170어절

[북한 말뭉치]

파일명	제목	발간 연도	어절수
• 7NA04V01.HWP	로동신문	2004	134,274
• 7NCAA201.HWP	국어'인민학교1'	2003	4,562
• 7NCAA202.HWP	국어'인민학교2'	2002	11,931
• 7NCAA203.HWP	국어'인민학교3'	2002	15,493
• 7NCAA204.HWP	국어'인민학교4'	2002	19,742
• 7NCAB201.HWP	국어'고등중학교1	2002	22,797
• 7NCAB202.HWP	국어'고등중학교2	2002	23,755
• 7NCAB203.HWP	국어'고등중학교3	2002	19,509
• 7NCAB205.HWP	국어문법'고등중학교2'	2002	11,965
• 7NCAB206.HWP	국어문법'고등중학교3'	2002	10,566
• 7NCBB201.HWP	조선력사'고등중학교2'	2002	16,074
• 7NCBB203.HWP	조선력사'고등중학교4'	2002	11,693
• 7NCBB205.HWP	조선력사'고등중학교6'	2002	8,455
• 7NCHA201.HWP	공산주의도덕'인민학교1'	2002	1,815
• 7NCHA202.HWP	공산주의도덕'인민학교2'	2002	4,435
• 7NCHB203.HWP	공산주의도덕'고등중학교3'	2002	13,011
• 7NCHA204.HWP	공산주의도덕'인민학교4'	2002	8,985
• 7NCHB205.HWP	공산주의도덕'고등중학교5'	2002	14,050
• 7NCHB201.HWP	공산주의도덕'고등중학교1'	2002	10,791
• 7NCJA201.HWP	경애하는 수령 김일성 대원수님 어린시절'인민학교1'	2001	2,554
• 7NCJA202.HWP	경애하는 수령 김일성 대원수님 어린시절'인민학교2'	2000	5,455
• 7NCJA203.HWP	김일성어린시절'인민학교3'	2000	7,353
• 7NCJA204.HWP	김일성어린시절'인민학교4'	2001	9,291
• 7NCJA205.HWP	항일의 녀성영웅 김정숙어머님	2002	9,448

	어린시절'소학교4'		
▪ 7NCJA206.HWP	항일의 녀성영웅 김정숙어머님의 혁명력사'고등중학교4'	2003	27,234
▪ 7NCJA207.HWP	김정일어린시절'인민학교1'	2001	3,382
▪ 7NCJA208.HWP	김정일어린시절'인민학교2'	2001	5,327
▪ 7NCJA209.HWP	김정일어린시절'인민학교3'	2000	8,345
▪ 7NCJA210.HWP	김정일어린시절'인민학교4'	1999	10,662
▪ 7NCJB994.HWP	위대한 수령 김일성대원수님 혁명력사'고등중학교4'	1999	37,133
▪ 7NCJB995.HWP	김정일혁명력사'고등중학교5'	1999	13,700
▪ 7NCJB204.HWP	김일성혁명활동'고등중학교1'	2001	23,981
▪ 7NCJB205.HWP	위대한 수령 김일성대원수님 혁명활동'고등중학교3'	2001	14,269
▪ 7NCJB996.HWP	위대한 령도자 김정일원수님 혁명력사'고등중학교4'	1999	10,628
▪ 7NCJB207.HWP	김일성혁명력사'고등중학교5'	1998	15,222
▪ 7NCJB208.HWP	위대한 령도자 김정일원수님 혁명력사'고등중학교6'	1999	31,202
▪ 7NCJB209.HWP	김정일혁명활동'고등중학교1'	2001	26,176
총 : 625,265어절			

3. 연구 추진 방법

남북한 및 중국, 해외 한국어의 이질성을 극복하고 어문 생활의 민족적 통합을 이룩하기 위해서 북한 언어(북한의 문화어) 자료와 카자흐스탄의 고려인 대표 기관 지인 고려일보, 중국 조선족 소학교에서 사용하고 있는 조선어 교과서의 대용량 데이터 베이스를 구축한다.

3.1 설계

3.1. 북한어

- 1990년 이후의 언어자료를 주축으로 하여 구축한다.
- 2000년부터 새롭게 적용되는 맞춤법을 적용한 자료를 구축하기 시작한다.
- 금년도 대상 자료는 로동신문을 제외하고는 모두 현재 사용하고 있는 북한의 교과서 자료로 한다.
- 북한어의 이질화 현실을 폭넓게 파악하고, 통일에 대비한 언어 통합 추진에 활용할 수 있도록 구성한다.
- 2003년에 수정된 북한의 띄어쓰기 규범을 반영한 자료로 로동신문을 입력한다.

3.1.2 중국 조선어

- 세종계획 1단계에서 구축된 중국 조선어 자료에 부족한 부분을 보충한다.
- 남북한과 균형을 유지하기 위해서 현재 중국 조선족 소학교에서 사용하고 있는 조선어문 교과서를 금년에 입력한다.
- 초중과 고중에서 사용하고 있는 조선어문 교과서는 다음 해에 입력한다.
- 앞으로도 자료의 균형을 위해 필요하다면 적절한 자료를 입력한다.

3.1.3 카자흐스탄의 한국어

- 전년도에 구축한 1993년 52주 전체 고려일보의 한글 원문에 이어, 금년에는 1994년 고려일보를 주축으로 구축한다. 이는 이 시점에서 북한 표기법에서 남한 표기법으로 변화가 이루어졌기 때문이다.
- 기타 레닌기치, 고려신문(사할린) 자료 등은 2차 목표로 잡는다.

- 자료의 한정 때문에 가능하면 종합지적인 성격이 강한 문헌 자료를 선정한다.

3.2 구축 방법

- 표준화 지침에 의거하여 부호와 문서 형식을 통일한다.
- 말뭉치의 효율적 관리와 이용에 필요한 헤더 정보와 마크업을 첨부한다.
- 품질 수준의 유지에 유의하여, 최종 오류율이 0.1% 이내가 되도록 한다.

3.3 표준화 구축 지침

연구 내용에 대한 언급에 앞서 고려대학교 민족문화연구원 자연언어 DB실의 “21세기 세종 계획 말뭉치 구축 지침”(1998. 5.) 중 이 연구에서 사용된 내용을 제시한다.

1) 기본 원칙

- ① 원문 유지 : 문헌의 경우 원문을 그대로 유지한다.
- ② 문서 정보 표기 방식 : SGML/TEI, 특히 TEI 라इट를 수용하여 문서 정보를 표기한다.
- ③ 헤더 : 헤더는 자세히 표기한다.
- ④ 본문 마크업 : 본문에는 최소한의 정보 마크업을 붙인다.

2) 파일 이름

- ① 도스 파일 명칭을 준수한다. 파일 이름은 8자리를 모두 사용하며, 3자리 확장자를 붙인다.
- ② 책의 경우, 한 권을 한 파일로 만드는 것을 원칙으로 한다.
- ③ 파일 이름 작성

- 첫 번째 자리

7 : 세종계획 7차년도 표시

- 두 번째 자리

N : 북한 자료

O : 중국 자료

S : 카자흐스탄 자료

- 세 번째 자리

A : 신문

B : 잡지

C : 책-교과서

D : 책-사전

E : 책-소설

F : 책-시

G : 책-기타 상상적 텍스트(수필 등)

H : 책-정보(인문, 사회, 자연, 예술), 기타 비상상적 텍스트

- 네 번째/다섯 번째 자리

▶신문 : 연도를 두 자리로 표시

▶잡지 : 연도를 두 자리로 표시

▶문학작품 : 자유롭게 두 자리 표시

- 여섯 번째 자리 : 부가정보 표시(선택적)

▶신문 : 종류 표시

V : 로동신문

W : 통일신보

X : 민주조선

▶잡지 : 종류 표시

A : 시사월간지

B : 시사주간지

C : 대중/연예지

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

D : 여성지
E : 경제지
F : 문화/예술지
G : 과학/기술지
H : 취미/실용지
I : 초등용 학생잡지
J : 중고등용 학생잡지
Z : 기타

- 일련번호 : 6-8번째 자리

3) 입력 환경

- ① 한글 3.x 이상에서 작업
- ② 전체를 바탕 스타일로 하고, 그 속성은 다음과 같다.

- 글자모양

-명조
-10 포인트
-보통 모양
-글자간격 0%
-장평 100%

- 문단모양

-왼쪽, 오른쪽, 문단간격, 낱말간격 모두 0
-들여쓰기 2
-줄간격 150%
-정렬방식 왼쪽

- 편집용지

-A4
-용지방향 좁게
-위/아래/왼쪽/오른쪽/머리말/꼬리말 모두 15.00mm
제본 0mm

- 페이지 : 표시하지 않음
- 머리말/꼬리말 : 표시하지 않음

4) 본문의 자료 표준화

① 입력 내용

- 본문을 입력하는 것을 원칙으로 한다.
- 앞표지 부분, 목차 등과 부록, 색인, 뒷표지 등의 내용은 특별한 경우를 제외하고는 입력하지 않는다.

② 텍스트 외형 정보

- 현대어 텍스트는 인쇄된 텍스트의 외형을 중요시하지 않는다.

③ 맞춤법

- 가능한 한 원문의 띄어쓰기를 유지한다.
- 단, 띄어쓰기에 상당한 오류가 있을 경우 교정을 하고 헤더에 반영한다.
- 교정 시 현행 북한과 연변의 규정을 준수한다.

④ 문장 부호 및 특수 기호

- 원문 그대로 유지하는 것을 원칙으로 한다.
- 가능하면 자판의 부호나 기호를 사용한다.
- 자판에 없는 경우는 일차적으로 전각기호를 다음에 반각기호를 사용한다.
- 가운데점 : 전각기호(3404)
- 말줄임표 : 전각기호(3406) 등

⑤ 한자

- 특별한 허용 규정이 없는 한, 원문의 한자를 모두 유지한다.
- 가능한 한 독음대로 해당 한자를 입력한다.
- 표준 한지 이외의 한자는 <훈글>의 확장한자로 입력한다. 사용자 정의문자는 본 사업 수행을 위해 향후에 공인되는 것 이외에는 사용하지 않는다.
- 표준한자/확장한자에 없는 한자는 다음과 같이 외부 전구 중의 하나로써 표시한다.
- 외부 전거는 가능한 한 UCS 코드 또는 국내 간행 사전을 우선적으로 사

용한다.

-외부 전거를 사용했을 경우에는 그 사실과, 전거의 종류를 헤더에 기록한다. 여기서는 漢語大字典:縮印本(湖北辭書出版社, 1992)에 따라 입력한다.

⑥ 영문자 및 외국문자

-원문의 외국문자는 그대로 유지한다.

-일차적으로는 자판에서 입력하되, 자판 입력이 불가능한 경우에는 전각 기호를 사용한다. 전각 기호가 없을 경우에는 반각 기호를 이용한다.

-글자 겹침은 사용하지 않는다.

⑦ 기타

-각 입력기(흔글)에서 제공하는 틀(수식, 표, 그림) 등은 사용하지 않는다.

-원문에 도표/그림이 있을 경우 생략하고, 그 사실을 명시적으로 밝혀 준다. 이때 <gap> 태그를 사용하고, reason 속성으로 그 이유를 밝힌다.

-각주는 원칙적으로 생략하나, 보존할 필요가 있을 경우 <note> 태그를 사용하여 표기한다.

-북한어, 연변어의 경우, 대화를 표시하는 각괄호는 전각 기호(《(3416), 》(3417))를 사용한다.

5) 텍스트의 분류

네 자리 코드를 사용하고, 그 앞에 말뭉치 종류를 로마자로 밝힌다.

N : 북한 자료

3.4 파일 전체 구조 및 SGML/TEI 태깅 일반 원리

3.4.1 문서의 구조

- 문서 구조의 예

<teiHeader>

헤더 내용

```

</teiHeader>
<text>
본문 내용
</text>

```

3.4.2 헤더

- 헤더 구조의 예

```

<!DOCTYPE tei.2 SYSTEM "c:\sgml\dtd\tei2.dtd" [
  <!ENTITY % TEL.corpus "INCLUDE">
  <!ENTITY % TEL.extensions.ent SYSTEM "sejong2.ent">
  <!ENTITY % TEL.extensions.dtd SYSTEM "sejong2.dtd">
]>
<tei.2>
<teiHeader>
<fileDesc>
<titleStmt>
<title>천리마, 전자파일</title>
<author>천리마사</author>
<sponsor>대한민국 문화관광부</sponsor>
<respStmt><resp>문헌 입력, 표준화, 헤더 붙임</resp>
<name>전주대학교</name>
</respStmt>
</titleStmt>
<extent>30704 어절</extent>
<publicationStmt>
<distributor>국립국어원</distributor>
<idno>NB96A030.hwp</idno>
<availability><p>2000-04 배포 불가</p></availability>
</publicationStmt>

```

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

```
<notesStmt>
<note><p></p></note>
</notesStmt>
<sourcedesc>
<bibl><author>천리마사</author>
<title>천리마</title>
<pubPlace>평양</pubPlace>
<publisher>천리마사</publisher>
<date>1996</date>
</bibl>
</sourceDesc>
</fileDesc>
<encodingDesc>
<projectDesc><p>21세기 세종계획 3차 년도 말뭉치 구축</p>
</projectDesc>
<samplingDecl><p>본문 전체 입력</p>
</samplingDecl>
</encodingDesc>
<profileDesc>
<creation><date>1996</date></creation>
<langUsage>
<language id=NK usage=99>북한어, 문화어</language>
</langUsage>
<textClass>
<catRef scheme='SJ21' target=N1260> 잡지: 사회, 일반</catRef>
</textClass>
</profileDesc>
<revisionDesc>
<change><date>(입력, 교정한 날짜)2000/06</date>
```

```

<respStmt><resp>입력자</resp><name>리현희</name></respStmt>
<item>본문 전체 입력, 헤더 붙임</item>
</change>
<change><date>2000/06/12</date>
<respStmt><resp>제 1교 교정자</resp><name>한명숙</name></respStmt>
<item>제 1교</item>
</change>
<change><date>2000/06/17</date>
<respStmt><resp>제 2교 교정자</resp><name>이성자</name></respStmt>
<item>제 2교 및 헤더 수정</item>
</change>
<change><date>2000/06/23</date>
<respStmt><resp>제 3교 교정자</resp><name>김금자</name></respStmt>
<item>제 3교</item>
</change>
<change><date>2000/10</date>
<respStmt><resp>최종 교정자</resp><name>소강춘</name></respStmt>
<item>최종교정자</item>
</change>
</revisionDesc>
</teiHeader>

```

3.4.3 본문 태그

- 본문은 <body>를 반드시 포함하고 있어야 한다.

1) 일반적 형태

```

<text>
  <front>
    앞표지, 목차 등의 머리말 부분
  </front>

```


21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

```
<body>
    본문 내용
</body>
<back>
    부록, 색인, 뒷표지 등의 꼬리말 부분
</back>
</text>
```

2) 합성적 형태

```
<text>
    <front>
        앞표지, 목차 등의 머리말 부분
    </front>
    <group>
        <text>
            <front>
                책의 머리말 부분
            </front>
            <body>
                <div>
                    본문 내용
                </div>
            </body>
        </text>
    <back>
        부록, 색인, 뒷표지 등의 꼬리말 부분
    </back>
</text>
```

3) 반드시 포함되어야 할 태그

<head> : 제목, 부제목, 표 제목 등 모든 종류의 제목

<p> : 산문에서의 문단

<l>, <lg> : 운문에서의 시행과 시행군(연 등)

4) 회화에 필요한 태그

<sp> 대사 부분, 대사자의 이름과 지문을 포함할 수 있다. 대사 자체는 <p>로 표시한다.

<speaker> 대사자의 이름 부분, 반드시 <sp>내에 있어야 한다.

<stage> 무대 지시 등 지문. <sp> 내 혹은 밖에 있을 수 있다.

5) 판독 불가, 생략 부분 등에 대한 태그

<gap reason='판독 불가'>

<gap reason='그림 부분'>

6) 년도에 등에 대한 태그

<date value='1996-3'>1996. 3</date>

3.5. 말뭉치 지도

2004년도까지 구축된 북한 및 해외한국어 말뭉치의 구성내용 및 비율은 다음과 같다.

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

구분 \	년도	80년대	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	총합계	비율(%)
신	문 32.1	0	0	0	0	0	0	0	0	116209	431992	123497	296323	356487	763413	0	134,274	2,222,195	30.4
잡 지 18.7	종합	0	0	0	0	0	0	0	152477	0	14498	25664	80741	39957	0	0	0	313,337	4.3
	문학	0	0	0	0	0	0	0	84879	55501	0	44443	207536	131419	0	0	0	523,778	7.2
	예술	0	0	0	0	0	0	0	76291	35383	0	0	154655	0	0	0	0	266,329	3.6
	인문	0	0	0	0	0	0	0	0	0	0	0	0	66144	0	0	0	66,144	0.9
	사회	0	0	0	0	0	0	0	0	0	0	0	0	47746	0	0	0	47,746	0.7
	자연	0	0	0	0	0	0	0	0	0	0	0	8936	0	0	0	0	8,936	0.1
책	총류	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	사상	67467	0	21469	136806	76601	0	76648	77366	182542	0	0	0	0	0	0	0	638,899	8.7
	교육자료	50546	0	0	0	0	15306	0	0	0	15222	103325	21153	84980	225530	31796	0	547,858	7.6
	체험기술적	0	0	16890	56277	68524	0	35363	0	0	0	0	0	0	0	0	0	177,054	2.4
	인문	0	0	0	45331	124275	109023	15461	34260	0	0	0	0	0	0	0	0	328,350	4.5
	사회	0	44268	126530	0	67394	47592	46851	6558	0	0	0	0	0	0	0	0	339,193	4.6
	자연	0	0	0	55813	0	15403	39066	37412	0	0	94957	0	0	0	0	0	242,651	3.3
	예술/취미/생활	0	0	0	33822	0	0	0	0	0	0	0	0	0	0	0	0	33,822	0.4
	장편 소설	0	0	0	0	0	0	0	0	0	0	0	277821	0	0	0	0	277,821	3.8
	중·단편 소설	33882	0	27018	61038	138158	37301	194203	0	0	0	38967	33882	0	0	0	0	564,449	7.8
상상 13.7	동화	0	0	0	0	0	0	24684	0	0	0	10165	2409	0	0	0	0	37,258	0.5
중 국		0	0	0	0	0	0	0	0	0	0	137170	0	0	0	0	0	137,170	1.9
카자흐스탄		0	0	0	0	258734	267611	0	0	0	0	0	0	0	0	0	0	526,345	7.3
계		151,895	44,268	191,907	389,087	733,686	492,236	432,276	469,243	389,635	461,712	578,188	1,083,456	726,733	988,943	31,796	134,274	7,299,335	100

(음영 표시 자료가 금년도 구축분임)

4. 결론 및 제언

금년도 과제의 가장 특기할 점은 북한의 교과서 전체를 구하고, 그 자료를 입력할 수 있었다는 점이다.

어느 나라나 공교육 기관에서 사용하는 교과서는 많은 연구를 통해 가장 기본적인 언어현상을 반영한 자료이다. 이런 점에서 본다면 금년도에 북한과 중국의 교과서를 입력했다는 것은 국어의 기초적인 연구는 물론이고, 교육 내용, 교육 방법, 교육 정책 등 국어 관련 사항과 해당 국가의 교육 전반에 걸친 연구 자료로 이 자료가 활용될 수 있다는 면에서 중요한 의미를 갖는다고 할 수 있다.

그리고 금년에 구축한 자료들은 카자흐스탄의 어문 규범이 바뀐 전후의 신문자료를 구축했고, 북한도 2000년과 2003년도에 개정된 어문규범이 반영된 교과서와 신문자료를 구축했다는 면에서 두 나라의 어문규범의 실제 활용 양상을 파악할 수 있게 했다.

앞으로는 지속적으로 후속 연도의 자료를 구축하여 충분한 현실 언어 자료를 구축하는 것이 중요할 것이다.

여 백

제 5 장

역사자료 말뭉치 개발

연구책임자 : 임 용 기

여 백

제 5 장

역사자료 말뭉치 개발

1. 연구 목적

본 연구의 목적은, 국어의 발달 변천사를 밝히고, 국어 교육과 어문 생활의 발전을 위한 국어사 자료 말뭉치의 구축을 수행하는 것이다.

올해는 3단계(2004-2007년) 사업 수행 중 첫 번째 해로 원시말뭉치 70만 어절과 주석말뭉치 5만 어절을 구축하는 것을 목표로 한다.

2. 원시 말뭉치 개발

2004년도의 역사자료 말뭉치 구축에서는 1) 문세영의 『조선어사전』을 대상으로 40만 어절의 원시 말뭉치를 구축하고, 2) 15세기 자료 중 법화경, 능엄경을 대상으로 30만 어절을 구축하는 것을 목표로 한다.

전체 입력 어절수는 원시 말뭉치 70만 어절로 자세한 어절수와 입력 파일은 아래와 같다.

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

파일명	제목	언어	장르	출판연도	어절수	저작권
PAZZ0001	조선어사전	한국어	기타(사전)	1942년	123,985	없음
PAZZ0002	조선어사전	한국어	기타(사전)	1942년	204,788	없음
PAZZ0003	조선어사전	한국어	기타(사전)	1942년	75,311	없음
P5CD0001	능엄경 권1	한국어	종교서류	1461년	31,338	없음
P5CD0002	능엄경 권2	한국어	종교서류	1461년	33,740	없음
P5CD0003	능엄경 권3	한국어	종교서류	1461년	31,938	없음
P5CD0004	능엄경 권4	한국어	종교서류	1461년	36,317	없음
P5CD0005	능엄경 권5	한국어	종교서류	1461년	25,277	없음
P5CD0006	능엄경 권6	한국어	종교서류	1461년	31,489	없음
P5CD0007	능엄경 권7	한국어	종교서류	1461년	27,258	없음
P5CD0008	능엄경 권8	한국어	종교서류	1461년	38,695	없음
P5CD0009	법화경 권3	한국어	종교서류	1448년	48,134	
합계					708,270	

3. 역사자료 구축 지침

본 연구 개발에서의 역사자료 구축 지침 두 가지로 나뉜다. 하나는 원시 말뭉치 구축 지침이고 또 하나는 주석 말뭉치 구축 지침이다. 여기서는 원시 말뭉치 구축 기준에 대한 지침을 기술하기로 한다.

3.1 목적

이 글은 21세기 세종 계획 말뭉치 구축 분야 중 역사자료 말뭉치 구축에서 이용할 말뭉치 자료의 작성 지침이다. 향후 구축될 역사자료 말뭉치는 이 지침에 의거, 표준화한다.

3.1.1 적용 대상

이 지침은 21세기 세종 계획 말뭉치 중 역사 자료 말뭉치에 적용한다.

3.1.2 지침 작성 방법

이 지침은 국립국어원의 역사자료 구축 지침과 고려대학교 민족문화연구소에서 마련한 21세기 세종 계획 말뭉치 구축 지침에 의거한다.

※기본 원칙

- 1) 원본의 유지 : 역사 문헌의 경우는 원문의 띄어쓰기가 이루어지지 않은 경우가 많고, 현대국어 자료와는 달리 연철·중철을 보이는 경우가 많다. 띄어쓰기의 경우는 현대 맞춤법의 원칙을 준용하여 띄어쓰기를 하고, 연철 및 분철의 경우는 그대로 반영하는 것을 원칙으로 한다.
- 2) 문서 정보 표기 방식 : SGML/TEI, 특히 TEI 라इट를 수용하여 문서 정보를 표기한다.
- 3) 헤더 : 헤더는 자세히 표기한다.
- 4) 본문 마크업 : 본문에는 최소한의 정보 마크업을 붙인다.
- 5) 표본 : 헤더의 경우 표본 파일을 제공하여 그곳에 채워 넣는 것을 원칙으로 하며, 본문도 텍스트의 종류에 따른 마크업 예를 제공한다.

3.2 파일 이름

3.2.1 일반원칙

- 1) 도스 파일 명칭을 준수한다. 파일 이름은 8자리를 모두 사용하며, 3자리의 확장자를 붙인다. 따라서 기본 파일의 형식은 다음과 같다. xxxxxxxx.hwp
- 2) 작업 중간 단계의 가공 파일은 일정한 확장자를 붙여 정리한다. (*.h01, *.h02, *.h03 등)

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

※ 파일 이름 작성 원칙

1) 파일의 첫 자에는 자료 코드(P)를 표시한다. 2번째 자리 이후부터는 아래에
규정한 방식에 따른 코드를 부여한다.

2) 역사 자료(국어사/고전문학)의 분류체계

- 두 번째 자리에는 시기를 표시한다.

<분류코드>

- 1 : 11세기 및 그 이전 시기의 자료
- 2 : 12세기 자료
- 3 : 13세기 자료
- 4 : 14세기 자료
- 5 : 15세기 자료
- 6 : 16세기 자료
- 7 : 17세기 자료
- 8 : 18세기 자료
- 9 : 19세기 자료
- 0 : 연대 불명 및 연대 혼합 자료
- A : 20세기 자료

- 세 번째 자리에는 문헌의 성격에 따른 구분을 표기한다.

<분류코드>

- A : 한문자료
- B : 원국문본
- C : 언해/번역자료
- D : 음역(音譯)자료
- E : 이두(吏讀)자료
- F : 구결(口訣)자료
- G : 향찰(鄉札)자료

...

Z: 기타

- 네 번째 자리에는 텍스트 유형에 따른 구분을 표기한다.

<분류코드>

A : 시가류

B : 소설류

C : 산문기록류 : 일기 등

D : 종교서류 : 불경, 유교서, 도가서, 성서류 등

E : 기술서류 : 의학서(醫學書), 병서(兵書), 농서(農書) 등

F : 역학서류 : 한어, 몽어, 만주어(청어), 왜어 등

G : 교민서류 : 훈음, 학습서, 교화서 등

H : 역사서류

I : 운서(韻書)류

J : 사서(辭書)류

K : 언간 및 문서류

Z : 기타

- 5~8자리에는 일련번호를 붙인다(일련번호는 역사자료 전체를 하나로 한다).

- 파일 이름의 예

고전 자료

고전자료-시기-문헌유형-종류-일련번호

P9BB0001.HWP (19세기 원국문본 고전소설)

3.3 입력 환경

- 1) 한글 3.x 이상에서 입력한다.
- 2) 탭 등 포맷 조판과 관련된 기능을 사용하지 않는다.
- 3) 전체를 바탕글 스타일로 하며, 바탕글은 다음 속성을 가져야 한다.

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

- 글자모양
 - 바탕체
 - 10 포인트
 - 보통 모양
 - 글자간격 0%
 - 장평 100%
- 문단모양
 - 왼쪽여백, 오른쪽여백, 문단간격, 낱말간격 모두 0
 - 들여쓰기 2
 - 줄간격 150%
 - 정렬방식 왼쪽
- 편집용지
 - A4
 - 용지방향 좁게
 - 위/아래/왼쪽/오른쪽/머리말/꼬리말 모두 15.00mm, 제본 0mm
- 페이지
 - 쪽번호는 매기지 않는다.
- 머리말/꼬리말
 - 머리말/꼬리말은 달지 않는다.

4) 전년도까지는 일반적인 입력 글씨체가 명조였으나, 올해 작업에서 명조체 일 경우 글씨가 깨지는 부분이 있어 부득이하게 바탕체를 사용하였다. 예를 들면 ‘ㅂ 、 ㅇ’의 경우 명조체로 입력할 경우 ‘ㅂ 、 ㅇ’과 같이 나타났다.

3.4 본문의 자료 표준화

3.4.1 입력 내용

1) 문헌 자료에 기재되어 있는 것은 가능한 한 모두 입력하는 것을 원칙으로

한다. : 외표지, 내표지, 간기(刊記) 후기 등

- 2) 본문 이외의 부분에는 적절한 태그(<front>, <back>)를 부착하여 본문(<body>)과 구분한다.

3.4.1.1 텍스트 외형 정보

- 1) 역사자료는 그 성격상 텍스트의 외형이 무시될 수 없다. 다만 TEI포맷을 그대로 반영하는 것은 본 연구의 목적에서 벗어나는 것이므로 최소한의 외형 정보만을 준다.

① 페이지 바꿈<pb>, 문단<p>, </p>, 제목<head>, </head> 등의 태그를 사용한다.

② 마침표 이외에 원문에 없는 문장부호, 즉 쉼표나 느낌표, 따옴표 등의 문장부호는 사용하지 않는다.

③ 기타 필요한 경우 적절한 태그를 사용하여 텍스트의 외형을 반영한다.

- 2) 글자 크기, 밑줄, 드러냄표 등은 무시한다.

- 3) 원문에서 장식 등의 이유로 어절의 내부 음절을 띄운 경우 모두 붙인다.

【예】 춘 향 전 → 춘향전

- 4) 대화와 지문은 문단나누기를 시행하고, 이야기 단위별로 문단을 구분한다. 그리고 문장이 끊어지지 않았다 하더라도 이야기가 바뀌었으면 문단을 구분한다.

3.4.1.2 표기

- 1) 특별한 이유가 없는 한 원문의 표기를 그대로 유지한다.

- 2) 탈자, 오자의 경우에도 수정하지 않는다.

3.5 띄어쓰기

3.5.1 일반원칙

- 1) 고전자료의 띄어쓰기는 기본적으로 현행 맞춤법 규정을 준용한다.
- 2) 관형사와 명사는 띄어쓰기를 원칙으로 하되, 관용어로 쓰이는 경우는 예외로 한다.

【예】 ‘잇썸’, ‘이썸’, ‘잇디’ 등

- 3) 어미와 조사, 파생접사는 붙여쓴다.
- 4) 합성어는 의미단위별로 띄어쓰는 것을 원칙으로 하되, 한자성어로서 굳어진 경우는 붙여쓴다.

【예】 성화년간, 공후거족

- 5) 연철은 붙여쓴다.

【예】 하느니(하는 이)

- 6) 한시구절은 5자나 7자라 하더라도 모두 붙여쓴다.

3.5.2 특수한 어절의 띄어쓰기

- 1) 어휘화된 ‘어간+접사’ 결합체의 처리

어휘화된 ‘어간+접사’ 결합체란 ‘이런드로, 그런드로, 갓으로, 보야호로, 여라문, 모든, 스스로’ 등과 같이 어원적으로는 ‘명사 + 조사’, 혹은 ‘용언 + 어미’ 등의 구성으로 분석할 수 있지만, 현실적으로 굳어진 의미를 지니는 단위를 가리킨다. 이들은 띄어쓰기를 하지 않는다.

- 2) 관형형 어미가 개재된 합성어

‘아촌아들, 올흔넌, 가온디’과 같이 ‘용언의 관형형 + 명사(혹은 의존명사)’의 형태로 이루어진 합성어를 통사적 구성으로 파악할 것인가 아니면 한 단어를 이룬 것으로 분석할 것인가 하는 문제이다. 본 연구에서는 이들을 어휘화한 것으로 다루어 한 단어로 처리하며 따라서 띄어쓰기를 하지 않는다.

- 3) 문법화된 명사 + 조사 결합체

2)의 경우와는 달리 어휘형태소와 문법형태소가 결합해서 또다른 문법형태

소를 이루게 되는 경우가 있다. ‘그에, 손디, 거괴’ 등이 대표적인 예인데, 본 연구에서는 이들 역시 하나의 단위로 다루어 띄어쓰기를 하지 않는다.

4) 관형형 어미 + 불완전명사 ‘ㅣ’로 구성된 어절

어절 단위로의 분석에서 가장 큰 어려움을 주는 것이 형태소 경계가 개재되었음에도 불구하고 연철된 단위의 처리이다. ‘ㄹ느니, 모드니, 사로리’ 등이 그 예인데, 본 연구에서는 이들은 띄어쓰기를 하지 않는다.

5) 어미화된 의존명사

어미화한 의존명사로는 ‘-ㄴ돌, -ㄴ던, ㄴ돈’ 등의 ‘ㄷ’와 ‘-ㄴ씨, -ㄴ쫄’ 등의 ‘스’를 들 수 있는 데, 이들은 통시적으로는 통사적 구성으로 파악되지만, 후기 중세국어 단계에서는 이미 하나의 단위로 굳어져서 단어 및 형태소 경계를 인정하기 어렵다. 본 연구에서는 이들을 분석하지 아니하고 하나의 어미로 다룬다.

6) 어원이 불명확하거나 하나로 굳어져서 분석하기 어려운 것들

‘다들-, 혼궂ㄹ티, 다몬’ 등이 그 예로, ‘다들-’의 경우 두 개의 어기가 결합한 합성어일 개연성이 있고 ‘다몬’의 경우는 용언의 관형형일 가능성이 있지만 둘 다 어원을 밝히기 어렵고, ‘혼궂ㄹ티’의 경우 ‘혼 + 궂 + 티 + 이’로 분석될 수 있을 것이지만 여기서는 하나의 부사로 다룬다. 따라서 이들은 한 어절로 다룬다.

3.6 문장 부호 및 특수 기호

1) 원문 그대로 유지하는 것을 원칙으로 한다.

2) 자판에 있는 부호 및 기호는 자판의 것을 사용한다.

① 자판에서 입력 가능한 기호를 전각 기호나 반각 기호를 사용하여 입력하지 않도록 한다.

② 겹낫표는 자판에서 입력한다.

③ 분수 표시는 전각 기호를 사용하지 않는다.

【예】 1/2, 3/4 (½ 등을 사용하지 않는다.)

3) 자판에 없는 부호 및 기호는 일차적으로 전각 기호를 선택하여 입력한다. 전

각 기호가 없을 경우, 한글의 반각 기호로 입력한다.

- ① 가운뎃점은 전각 기호(3404)를 사용한다. 한글 2.0 이상에서 가능한 'Ctrl'과 'F10'키를 동시에 눌러 문답상자를 부른 후 마침표를 쳐서 입력하는 방식을 취하지 않도록 한다.

【예】 국가·민족

- ② 말줄임표는 전각 기호(3406)을 사용한다.

【예】 그러나...

- ③ 영문자, 한글 자모, 괄호문자는 자판에 있는 기호로 간주하여 전각 기호를 사용하지 않는다.

【예】 ㄱ ㄴ ㄷ ㄹ, (1) (2) (3)

- ④ 원문자는 전각 기호를 그대로 사용하되 원문자 숫자가 20 이상일 경우 전각 기호가 없으므로 두 숫자의 원문자를 나란히 붙여 사용한다.

【예】 ① ② ②①

- 4) 부호 또는 기호가 연이어 사용되되 어절이 구분되는 경우 그 사이를 띄운다.

【예】 ...했다." "그리고...

3.7 한자

- 1) 특별한 허용 규정이 없는 한, 원문의 한자를 모두 유지한다.
- 2) 가능한 한 해당 한자를 그대로 입력하되, 약자와 같은 것은 정자로 바꿔 입력한다.
- 3) 표준 한자 이외의 한자는 <한글>의 확장한자로 입력한다. 사용자 정의문자는 본 사업 수행을 위해 향후에 공인되는 것 이외에는 사용하지 않는다.
- 4) 표준한자/확장한자에 없는 한자는 다음과 같이 외부 典據 중의 하나로써 표시한다.
- 5) 외부 전거는 가능한 한 UCS 코드 또는 국내 간행 사전을 우선적으로 사용한다.
- 6) 외부 전거를 사용했을 경우에는 그 사실과, 전거의 종류를 헤더에 기록한다.

외부 전거 표시의 형식 및 본보기는 다음과 같다.

<형식>

- ① 외부 전거 시작 표지로 &를, 마침 표지로 ;를 쓴다
- ② 위의 표지 사이에 전거에 따른 출처 표시를 넣는다.
- ③ 외부 전거 표시 내부에는 빈칸이 들어가지 않는다.
- ④ 외부 전거 표지는 그 전체가 한 글자에 해당하며, 앞뒤의 입력 상태에 별도의 스페이스 등을 추가하지 않는다.

【예】

孟子謂蚘&U+9F03;曰: “子之辭靈丘而請士師, 似也, 爲其可以言也; … ”
淳于&U+9AE0;은 <孟子>에 등장하는 인물로서...

<외부 전거 종류와 표현 형식>

- UCS 코드(Unicode) CJK 영역 한자 목록(별도의 참조서 제공 예정)
&U+67E4; (67E4는 해당 한자의 16진수 코드값)
- CNS 코드(Unicode) 대만에서 제정한 Chinese National Standard 1~7수준 한자
&C3-76A3; (3은 CNS 문자 수준, 76A3은 16진수 코드값. Big5 코드는 'C0-'으로 표시)
- 東亞出版社 刊, 東亞 漢韓大辭典(李家源 외 감수, 1982) *본문 2220면 판.
&XD0873.05; (D=출처, 0873=해당 한자 수록면, 05=해당 면에서의 수록 순차)
- 民衆書林 刊, 漢韓大字典(전면개정 · 증보판, 1997). *본문 2394면 판.
&XE0792.09; (E=출처, 0792=해당 한자 수록면, 09=해당 면에서의 수록 순차)
- 諸橋轍次, 大漢和辭典(修訂版. 大修館書店, 1986).
&XM08269; (M=출처, 08269=모로하시 사전 수록 한자의 일련번호)
- 漢語大詞典(漢語大詞典出版社, 1990).
&XH08.0483.02; (X=출처, 08=권차, 0483=수록면, 02=해당 면에서의 수록 순

차)

3.7.1 영문자 및 외국문자

- 1) 원문에 외국문자가 섞인 경우 그대로 유지한다.
- 2) 일차적으로는 자판에서 입력하되, 자판 입력이 불가능한 경우에는 전각 기호를 사용한다. 전각 기호가 없을 경우에는 반각 기호를 이용한다.

3.8 기타

- 1) 각 입력기(한글)에서 제공하는 틀(수식, 표, 그림) 등은 사용하지 않는다.
- 2) 원문에 도표/그림이 있을 경우 생략하고 그 사실을 명시적으로 밝혀 준다.
이때 <gap> 태그를 사용하고 reason 속성으로 그 이유를 밝힌다.
- 3) 각주는 원칙적으로 생략하나, 보존할 필요가 있을 경우 <note> 태그를 사용하여 표기한다.
- 4) 협주는 <note> 태그를 사용하여 구분한다.

4. 기존 말뭉치와의 차이점

- (1) 조선어사전의 경우 띄어쓰기가 현대 국어와 다른데 이것을 이번 작업에서는 현대 국어와 달리 띄어쓰기가 되어 있는 어절은 '#' 표시를 하고 현대 국어에 맞게 띄어쓰기를 하였다.

예: 가는것→가는# 것

또한 현대 국어에서는 붙여 써야 함에도 불구하고 원전에 띄어쓰기가 되어 있는 경우는 원전을 따라 띄어쓰기를 하였다.

(2) 15세기 자료는 원전에 있는 방점과 한자음 모두를 입력하는 것을 원칙으로 하였고 한문원본 역시 입력하였다.

이럴 경우 한자음에 대한 연구와 방점 연구가 용이해 지며 한문원본까지 입력하였기 때문에 문헌에 대한 이해도를 높일 수 있다.

5. 실제 말뭉치 작업의 예

5.1. 조선어사전의 예

```
<!DOCTYPE tei.2 SYSTEM "c:\wsgml\wdtd\wtei2.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>修正增補朝鮮語辭典, 전자파일</title>
```

```
<author>문세영</author>
```

```
<sponsor>대한민국 문화관광부</sponsor>
```

```
<respStmt><resp>문헌입력, 표준화, 헤더붙임</resp>
```

```
<name>연세대학교, 임용기</name>
```

```
</respStmt>
```

```
</titleStmt>
```

```
<extent>75751어절</extent>
```

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

```
<publicationStmt>
  <istributor>국립국어원</istributor>
  <idno>xxx.hwp</idno>
  <availability><p>배포 불가</p></availability>
</publicationStmt>
<notesStmt>
  <note><p>이 텍스트는 朝鮮語辭典刊行會에서 소화17년에 간행된
<修正增補朝鮮語辭典>을 입력한 것이다. </p></note>
</notesStmt>
<sourceDesc>
  <bibl><author>문세영</author>
    <title>修正增補朝鮮語辭典</title>
    <pubPlace>朝鮮語辭典刊行會</pubPlace>
    <publisher>大山治永</publisher>
    <date>소화17</date>
  </bibl>
</sourceDesc>
</fileDesc>
<encodingDesc>
  <projectDesc><p>21세기 세종계획 3단계 1차년도 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>2003</date></creation>
  <langUsage>
    <language id=K0 usage=99>한국어, 고어</language>
  </langUsage>
```

```

<textClass>
    <catRef scheme='SJ21' target='p9bc'>사전</catRef>
</textClass>
</profileDesc>
<revisionDesc>
    <change><date>2003/9 </date>
        <respStmt><resp>프로젝트
책임자</resp><name>임용기</name></respStmt>
        <item>세종21프로젝트 헤더붙임</item>
    </change>
</revisionDesc>
</teiHeader>
<text>
<body>

<pb n='992'><p>【식사】 (式辭) 名 식장(式場)에서 그 식에 대하여 인사로 하는
말 [-하다 自]</p>
<p>【식:사】 (食事) 名 음식을 먹는 것 먹음재 [-하다 自]</p>
<p>【식사】 (飾詐) 名 거짓 꾸미는 것 [-하다 他]</p>
<p>【식사】 (飾辭) 名 거짓 꾸미어 하는 말</p>
<p>【식:산】 (殖産) 名 재산을 늘리는 것 생산물을 늘이는 것 [-하다 他]</p>
<p>【식-상】 (食床) 名 밥상</p>
<p>【식상】 (食傷) 名 먹는 음식이 소화가 되지 않고 복통 토사 (腹痛 吐瀉)가
나는 병</p>
<p>【식-색】 (食色) 名 음식과 여색</p>
<p>【식서】 (飾緒) 名 피륙의 가장자리를 울이 풀리지 않도록 짚는 부분</p>
<p>【식-성】 (食性) 名 음식에 대하여 좋아하고 싫어하는 성질</p>
<p>【식소-사번】 (食少事煩) 名 먹을 것은 적도 일만 많은 것 [-하다 形]</p>
<p>【식술】 (食率) 名 집안에 딸린 식구</p>
<p>【식:송】 (息訟) 名 화해하고 소송을 그치는 것 [-하다 他]</p>
<p>【식송-망정】 (息訟網亭) 句 술심거정사</p>
<p>【식-:수】 (食樹) 名 「식복」 (食福)</p>

```

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

- <p>【식수】(植樹)名 「식목」(植木)과 같음 [-하다 自]</p>
<p>【식-수유】(食菜庾)名 植 「수유」(菜庾)와 같음</p>
<p>【식-시】(食時)名 아침과 저녁에 식사하는 시각 끼니때</p>
<p>【식식-하다】刑 「씩씩하다」의 옛말</p>
<p>【식-신】(食神)名 음식을 맡은 신명 (술가의 말)</p>
<p>【식신-방】(食神方)名 식신의 방위 사람의 생년 ·월 ·일 ·시로 작정되는 것이요 이 방위를 길방(吉方)이라 함 (술가의 말)</p>
<p>【식심】(食甚)名 无 일식 또는 월식 때에 해 또는 달이 제일 많이 가려지는 것</p>
<p>【시어다】刑 「회다」의 사투리</p>
<p>【식언】(食言)名 남과 약속한 것을 지키지 아니하는 것 [-하다 他]</p>
<p>【식역증】(食依症)名 뱃속에 적열(積熱)이 있어서 음식을 먹어도 몸은 수척하여 지는 병</p>
<p>【식열】(食熱)名 어린애의 과식으로 나는 신열</p>
<p>【식염】{-념}(食鹽)名 소금</p>
<p>【식욕】(食慾)名 음식을 탐하는 욕심</p>
<p>【식용】(食用)名 음식에 쓰는 것 [-하다 他]</p>
<p>【식용-식물】(食用植物)名 사람의 식료로 쓸# 수 있는 식물</p>
<p>【식원:복】(食遠服)名 음식을 먹은# 지 한참 있다가 약을 먹는 것 [-하다 他]</p>
<p>【식육】(瘰肉)名 굶은살</p>
<p>【식육-류】(食肉類)名 動 오로지 육식만하고 사는 짐승</p>
<p>【식은-땀】名 사람이 죽으려고 할 때에 흘리는 땀 진땀 油汗 柔汗 冷汗</p>
<p>【식은-죽-먹기】名 하기 쉬운 일을 가리키는 말</p>
<p>【식-음】(食飲)名 먹는# 것과 마시는 것</p>

</body>

</text>

</tei.2>

5.2. 15세기 문헌의 예

!DOCTYPE tei.2 SYSTEM "c:\Wsgml\Wtd\Wtei2.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>楞嚴經諺解, 전자 파일</title>
        <author>?</author>
        <sponsor>대한민국 문화관광부</sponsor>
        <respStmt><resp>국립국어원 말뭉치 입수, 표준화, 헤더 불임</resp>
          <name>국립국어원</name>
        </respStmt>
      </titleStmt>
      <editionStmt>
        <edition><date>1994/11/10</date>전산입력</edition>
      </editionStmt>
      <extent>31338 어절</extent>
      <publicationStmt>
        <distributor>국립국어원</distributor>
        <idno>P5CA9001.hwp(Knegea00.hwp)</idno>
        <availability><p>배포 불가</p></availability>
      </publicationStmt>
      <notesStmt>
        <note><p>이 텍스트는 국립국어원 작업 내용 유지</p></note>
      </notesStmt>
      <sourceDesc>
        <bibl><author>?</author>
          <title>楞嚴經諺解</title>
          <pubPlace>서울</pubPlace>
          <publisher>홍문각</publisher>

```


21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

```
<date>1982</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>1461</date></creation>
<langUsage>
  <language id=KO usage=99>한국어, 고어</language>
</langUsage>
<textClass>
  <catRef scheme='SJ21' target='P5CA'>역사자료: 15세기, 언해/번역자료, 시
가류</catRef>
</textClass>
</profileDesc>
<revisionDesc>
  <change>
    <date>1994/11/10</date>
    <respStmt>
      <resp>입력자</resp><name>송기중</name>
    </respStmt>
  </change>
  <change>
    <date>1994/12/05</date>
    <respStmt>
```

```

        <resp>교정자</resp><name>송기중</name>
    </respStmt>
    <item>교정</item>
</change>
<change>
    <date>2004/11</date>
    <respStmt>
        <resp>프로젝트 책임자</resp><name>임용기</name>
    </respStmt>
    <item>파일 변환, 세종 21 프로젝트 헤더 붙임, 마킹</item>
</change>
</revisionDesc>
</teiHeader>

<text>
<body>
<div1 n='10' type='권'>
<pb n='楞解1,1a'><head>首楞嚴經要解序</head>
<p>前住福州上生禪院嗣祖沙門 及南 撰</p>
<p>像季己還애 道術이 既裂호야 明心之士ㅣ 妄認緣塵호야 爲物의 所轉호며
義學之徒ㅣ 虛驕多聞호야 不全道力호야 奇才茂器ㅣ 皆流爲蒸砂迷客이며
說食飢夫ㅣ 로다 首楞嚴王이 懸知其然호샤 誕敷秘典호샤 力救倒妄호샤
淘汰啓迪호산 諄諄之慈ㅣ 靡所不至어시닐 而末世初機ㅣ 罕能究盡호도다
溫陵環師ㅣ 一生을 掩關호고 <pb n='楞解1,2b'>與世異好호야 獨陪黃卷聖賢호야
冥搜博訪호야 藉其筌筏호야 以探如來藏호야 游徧知海호야
而造乎一切畢竟之地호야 思與同志와 共코져 호야 因爲是解호니 昔에 月盖比丘ㅣ
嘗問藥王如來의 法供養義호대 藥王이 告言호샤되 諸佛所說難信深經은
清淨無染호야 能令衆生으로 成最正覺호야 離衆魔事케 호느니 若於是經에
方便으로 解說호되 依義호고 不依語호며 依智호고 不依識호며 依了義호고
不依不了義호며 依法호고 不依人호야 <pb n='楞解1,2a'>直使無明生死로
畢竟滅盡호되 而無滅盡相호시 是名最上法之供養이라 호신대 月盖ㅣ 蒙敎호스와
通達妙道호야 得無礙辯호야 卽於藥王所轉法輪에 隨順分布호야 化百萬億人호야
於無上覺애 立不退轉호니라 環師ㅣ 既達妙道호야 仰睇月盖호야 於釋迦如來入

```

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

所轉法輪에 最初華嚴과 最後法華와 及此楞嚴無上寶印에 皆以方便健相分別로
迴脫語言호며 離心意識호야 唯義所在오 曾不依人호야 覺嘗見病이 <pb
n='楞解1,2b'>廓無纖翳호야 於清淨經에 能不染汙호야 我願以此로 流布無窮호리라
호니 其所化人이 何啻億萬이리오 直如一燈이 然百千燈호야 使冥者 | 皆明호되
而明이 終不盡호니라 古로 述序引호야 翼替流通於塵墨劫호야 作法供養호노라
建炎改元中秋吉日善集堂序 像:상季·평·룻 :오매 <note>如성來림 滅·멸·도·동·호·신
後:후·에 正·정·法·법 像:쌍·法·법·이 各·각·各·각 一·일·千·천·年·년·이·오 末·말·法·법·이
一·일·萬·만·年·년·이·니 像:쌍·은·새·죽·홀·씨·니 부터 :겨·신·적·과·새·죽·홀·씨·라
季·평·논 末·말·이·라</note> 道:똥·術·숯·이 호·마·씩·야·다·여<note>術·숯·은
法·법·이·라</note> ㅁ·슴·불·깊 :사·르·미 <pb n='楞解1,3a'>:거·츠·리 緣·원·塵·면·을
아·라 <note>
</body>
</text>
</tei.2>

6. 구축된 말뭉치 현황

지금까지 구축된 말뭉치를 시기별, 분야별로 도식화 하면 다음과 같다. 전반적으로 말뭉치의 균형성이 떨어지는 것 같으나 15세기 문헌의 경우 우리가 접할 수 있는 문헌이 상대적으로 제한되어 있어 구축된 말뭉치는 한 쪽으로 치우칠 수밖에 없다. 또한 17, 18세기 문헌의 경우 필사본이 많은 관계로 다양한 장르의 입력에 많은 어려움이 있다. 전체적으로 20세기 쪽에 많이 입력된 경향이 있으나 이것 역시 문헌의 중요도를 고려하면 이해할 수 있는 부분이다. 신소설이 450,000 어절이 입력되어 있는데 이것은 대부분의 신소설을 다 입력한 것이고 산문기록류에 100만에 가까운 어절이 입력되어 있는데 이 부분 역시도 최초의 사전인 문세영의 『조선어사전』을 입력해서 그런 것이다.

하지만 여전히 19, 20세기 문헌에 치우친 경향이 있는 것은 사실이다. 이 부분은 좀더 고려해서 다음 입력 문헌을 선정해야 할 것이다.

제5장 역사자료 말뭉치 개발

		15세기	16세기	17세기	18세기	19세기	20세기	미상	합계
원 국 문 본	시가류		3,919	16,192	103,534	17,618		600,801	742,064
	소설류			16,699		63,219		2,390,075	2,469,993
	산문기록 류		19,794	57,424	163,493	337,765	942,417	19,088	1,539,981
	종교서류						587,778		587,778
	기술서류					26,186			26,186
	교민서류					32,711			32,711
언 해 / 번 역 자 료	시가류	177,169		104,763					281,932
	소설류				409,931	12,056		74,086	496,073
	산문기록 류				8231				8231
	종교서류	318,413 373,748	81,277 3,840	51,246 327,399	87,792	241,114		7,491 146,245	787,333 592,528
	기술서류	37,230	17,669	75,663	38,200	12,671		5,476	186,909
	역학서류	3,029	27,293	52,124	119,363	10,279			212,088
	교민서류	30,364	80,494 6,262	115,107	320,646	74,623			621,234 6,262
	역사서류				13,789				13,789
기 타	교민서류		23,568	13,789					37,357
	사서류					123,645			123,645
국어연구원 구축분		566,205	234,220	503,024	1,101,486	951,887	424,114	3,097,017	6,877,953
세종21계획 구축분		373,748	29,896	384,823	163,493	677,764	2,347,573	335,140	4,312,437

여 백

제 6 장

전문용어 말뭉치 개발

연구책임자 : 이 원 표

여 백

제 6 장

전문용어 말뭉치 개발

1. 연구의 필요성, 목표 및 방법

1.1 연구의 필요성

21세기가 지식정보화 사회로 접어들면서 컴퓨터와 인터넷 전산망을 통한 ‘지식’과 ‘정보’가 하루가 다르게 쏟아져 나오고 있다. 따라서 사용자에게는 그러한 도구를 통해서 손쉽게 빠르게 ‘정보’를 접할 수 있는 기회가 기하급수적으로 늘어나고 있다. 즉, ‘정보’의 인플레이션이 일어나는 것이다. 그러나 그에 대한 부정적인 측면으로 ‘정보’의 표현과 유통의 다양화로 인해 외국어 특히 영어에 대한 수용의 경우, 용어 사용의 개인화 현상이 일어나면서 상당한 혼란이 일고 있는 실정이다. 게다가 하나의 동일한 전문용어가 학회나 사용자에 따라 다르게 번역되어 쓰이는 것이 학문 발전에 저해 요인이 되고 있다. 이런 문제를 해결하기 위해 현재 각 세계 IT선진국에서는 자국의 전문용어의 정비는 물론 새로이 생성되는 전문용어를 좀더 효율적으로 관리할 수 있는 시스템의 구축과 개발 공정의 체제 구축에 전력을 기울이고 있다.

그래서 전문용어 말뭉치 개발 과제는 이러한 세계적인 연구 동향에 발맞추어 전문용어에 대한 정리 작업인 KORTERM “전문용어의 정비” 분과의 전문용어 표준화 사업에 기반 자료를 제공하는 것을 기본 골격으로 한다. 해당 전문학술학회에

협조를 구해 그 분야의 가장 대표적인 텍스트를 선정하여 그 내용을 전문용어 말뭉치의 자료로 구축한다. 또 이러한 전문용어 말뭉치의 빈도를 이용해 구축한 전문용어 목록은 한국어 사전의 전문용어 표제어 선정 기준의 유용한 기초 자료가 된다. 그리고 한·영, 한·일 등 대역어 사전에서도 전문용어 표제어 선정에 통일성 있는 도움을 줄 수 있다. 더 나아가 병렬 말뭉치 구축이나 기계번역에서도 전문용어 선택에 대한 어려움을 상당 부분 덜어줄 수 있을 것이다.

물론 연구 인력의 부족, 한국어식 전문용어 사용의 경시, 인터넷상에서의 용어의 난립과 사전의 보급에 의한 비표준 전문용어의 확산 등이 여전히 해결해야 할 문제로 남지만, 이러한 현실적인 취약점을 점진적으로 보완해 가면서 날로 빠르게 성장, 발전하고 있는 정보지식사회에 발맞출 수 있는 전문용어를 구축할 해야 할 것이다.

1.2 연구의 목표

전문용어 말뭉치 개발은 각 분야별로 전문용어를 집성하여 전문용어의 사용 실태 조사 및 표준화를 위한 대규모 통합 정보 자료를 구축하고, 전문용어의 표준화를 위한 전산학적·언어학적 연구의 토대를 제공하는 것을 목표로 한다. 이를 통해 표준화된 용어를 보급하고, 새로 생성되는 전문용어를 효율적으로 관리할 수 있는 전문용어 관리 시스템을 확립하며, 전문용어를 자국어화할 수 있는 기반을 마련하게 될 것이다.

전문용어 말뭉치 개발에서 구축되는 원시 말뭉치는 이후 KORTERM의 '전문용어 정비' 연구에 활용되는 것이므로 원시 말뭉치 구축에서 생기는 여러 문제점들은 KORTERM과 협의하여 원활하게 풀어간다.

1.3 연구 방법

○ 연구 방법

전문용어 말뭉치는 2단계까지 ‘전문용어 정비’ 분과에서 이루어 낸 결과물을 기초로 하여 추후 ‘전문용어 정비’ 분과에서 용어 추출, 용어의 사용빈도 파악, 해당 분야 언어모델링 등 다양한 언어공학적 처리를 할 수 있도록 말뭉치를 구축한다. 또한 해당 분야의 용어학적 대표성을 가지면서 균형성 있는 말뭉치가 될 수 있도록 말뭉치 구성을 설계한다.

○ 구축 내용

분량 : 원시 말뭉치 100만 어절

내용 : 전문용어의 형태와 목록을 추출할 수 있도록, 균형성과 정보성이 고려된 전문분야 도서와 자료들을 선정하여 전문분야 말뭉치를 구축한다.

2. 말뭉치 구축 결과

2.1 말뭉치 구축 현황

○ 원시 말뭉치 구축 목표 : 전산학 및 전자·전기 분야 100만 어절

- 구축 현황 : 1,000,067 어절 구축 완료

2.2 말뭉치 구축 내용

이전까지 세종 특수자료 말뭉치 구축에서 살펴보면 한 권의 대상 텍스트를 입력했을 때 그 결과물은 하나의 파일이었으나 KORTERM의 ‘전문용어 정비’ 분과에서 구축한 원시 말뭉치를 살펴보면 한 권의 텍스트 내의 한 장(chapter)이 하나의 파일을 구성한다. 따라서 2004년도 전문용어 말뭉치 구축에서는 KORTERM과의 협력 문제가 있기 때문에 KORTERM의 원시 말뭉치 구축 방법대로 하나의 장

(chapter)을 하나의 파일로 구성하는 것을 지침으로 삼고 작업하여 131개의 파일 1,000,067 어절을 구축하였다.

그러나 KORTERM의 지침대로 말뭉치를 구축하면서 기존 세종 말뭉치와의 통일성 문제가 제기되었다. 따라서 2004년도 전문용어 말뭉치 개발은 KORTERM과 협력관계에 있으나 입력지침 등에 대한 협의가 완전하지 않아서 KORTERM에 제공하는 말뭉치는 기존에 KORTERM에서 입력하던 방식대로 작업하여 KORTERM에 제공하고, 최종보고서를 작성하면서 결과물을 제출할 때에는 전문용어 말뭉치가 21세기 세종계획 특수말뭉치 분과에 속해 있는 만큼 결과물의 통일성을 고려하여 기존의 “세종 기초말뭉치 통합 표준화”에 맞추어 “책의 경우, 한 권을 한 파일로 만드는 것을 원칙으로 한다”는 규정을 따르기로 한다.

따라서 중간보고 때 제출한 파일의 수는 131개였으나 이들을 각각의 책으로 수합한 결과 최종보고의 시점에서는 전체 13권의 도서, 13 파일을 결과물로 제출하게 되었다.

KORTERM의 원시 말뭉치 구축 지침과 세종의 말뭉치 구축 지침이 서로 다르기 때문에 문제가 되는 것은 보고서의 뒤에서 다루기로 한다.

원시 말뭉치 구축 도서의 목록은 아래와 같다.

○ 원시 말뭉치 구축 도서

- C언어 프로그래밍, p222
- 컴퓨터 과학 총론, p497
- 컴퓨터 공학, p510
- C로 쓴 자료구조론, p586
- 한국어판 응용운동체제개념, p901
- 알고리즘, p610
- 마이크로 전자회로, p1369
- 신호와 시스템, p579
- 이산신호처리(1-7장 입력, ‘찾아보기’ 입력)
- 전자기학(‘찾아보기’ 입력)

- 데이터베이스시스템('찾아보기' 입력)
- 컴퓨터시스템구조('찾아보기' 입력)
- 컴퓨터시스템구조론('찾아보기' 입력)

작업 상황이 말뭉치 구축 목표인 100만 어절에 가까워짐에 따라 “이산신호처리”는 1장에서 7장까지 입력하였다(‘찾아보기’ 포함). 그리고 KORTERM에서 구성한 구축 자료 목록 중에서 목표량을 달성함으로써 입력이 되지 않은 도서 4권(“데이터베이스시스템”, “전자기학”, “컴퓨터시스템구조”, “컴퓨터시스템구조론”)은 전문용어 연구에 상당한 도움이 될 수 있는 ‘찾아보기’를 각각 입력하였다. 따라서 이 네 개의 파일에는 ‘찾아보기’만 입력되어 있다.

2.3 말뭉치 목록

앞서 밝힌 것과 같이 말뭉치 목록은 KORTERM에 제공한 목록과 세종 결과물로 제출할 목록 두 가지가 있다. 그 목록은 아래와 같다.

○ KORTERM에 제공한 말뭉치 목록

파일명	자료명	어절
박슬아_전산1_1장	C언어 프로그래밍	6,075
박슬아_전산1_2장	C언어 프로그래밍	4,483
박슬아_전산1_3장	C언어 프로그래밍	2,616
박슬아_전산1_4장	C언어 프로그래밍	5,580
박슬아_전산1_5장	C언어 프로그래밍	6,990
박슬아_전산1_6장	C언어 프로그래밍	4,765
박슬아_전산1_7장	C언어 프로그래밍	4,083
박슬아_전산1_8장	C언어 프로그래밍	5,023
박슬아_전산1_찾아보기	C언어 프로그래밍	798
박슬아_전산2_chapter0	컴퓨터 과학 총론	2,964
박슬아_전산2_chapter1	컴퓨터 과학 총론	9,617
박슬아_전산2_chapter2	컴퓨터 과학 총론	5,564

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

박슬아_전산2_chapter3	컴퓨터 과학 총론	9,036
박슬아_전산2_chapter4	컴퓨터 과학 총론	7,365
박슬아_전산2_chapter5	컴퓨터 과학 총론	10,431
박슬아_전산2_chapter6	컴퓨터 과학 총론	5,956
박슬아_전산2_chapter7	컴퓨터 과학 총론	5,159
박슬아_전산2_chapter8	컴퓨터 과학 총론	4,537
박슬아_전산2_chapter9	컴퓨터 과학 총론	5,299
박슬아_전산2_chapter10	컴퓨터 과학 총론	8,598
박슬아_전산2_chapter11	컴퓨터 과학 총론	7,788
박슬아_전산2_INDEX	컴퓨터 과학 총론	913
유형옥_전산3_1장	컴퓨터공학	7,041
유형옥_전산3_2장	컴퓨터공학	11,233
유형옥_전산3_3장	컴퓨터공학	7,742
유형옥_전산3_4장	컴퓨터공학	6,6433
유형옥_전산3_5장	컴퓨터공학	6,186
유형옥_전산3_6장	컴퓨터공학	7,578
유형옥_전산3_7장	컴퓨터공학	6,964
유형옥_전산3_8장	컴퓨터공학	8,185
유형옥_전산3_9장	컴퓨터공학	9,234
유형옥_전산3_10장	컴퓨터공학	7,532
유형옥_전산3_11장	컴퓨터공학	7,887
유형옥_전산3_12장	컴퓨터공학	4,020
유형옥_전산3_13장	컴퓨터공학	975
유형옥_전산4_1장	C로 쓴 자료구조론	9,723
유형옥_전산4_2장	C로 쓴 자료구조론	7,989
유형옥_전산4_3장	C로 쓴 자료구조론	4,926
유형옥_전산4_4장	C로 쓴 자료구조론	8,354
유형옥_전산4_5장	C로 쓴 자료구조론	10,505
유형옥_전산4_6장	C로 쓴 자료구조론	10,139
유형옥_전산4_7장	C로 쓴 자료구조론	12,925
유형옥_전산4_8장	C로 쓴 자료구조론	6,792
유형옥_전산4_9장	C로 쓴 자료구조론	6,880
유형옥_전산4_10장	C로 쓴 자료구조론	16,440

제6장 전문용어 말뭉치 개발

유형옥_전산4_찾아보기	C로 쓴 자료구조론	1,493
박슬아_전산5_1장	응용운영체제개념	4,432
박슬아_전산5_2장	응용운영체제개념	7,065
박슬아_전산5_3장	응용운영체제개념	10,916
박슬아_전산5_4장	응용운영체제개념	5,906
박슬아_전산5_5장	응용운영체제개념	3,504
박슬아_전산5_6장	응용운영체제개념	8,432
박슬아_전산5_7장	응용운영체제개념	8,178
박슬아_전산5_8장	응용운영체제개념	5,431
박슬아_전산5_9장	응용운영체제개념	8,619
박슬아_전산5_10장	응용운영체제개념	9,115
박슬아_전산5_11장	응용운영체제개념	10,529
박슬아_전산5_12장	응용운영체제개념	7,846
박슬아_전산5_13장	응용운영체제개념	6,890
박슬아_전산5_14장	응용운영체제개념	8,053
박슬아_전산5_15장	응용운영체제개념	2,327
박슬아_전산5_16장	응용운영체제개념	5,027
박슬아_전산5_17장	응용운영체제개념	6,918
박슬아_전산5_18장	응용운영체제개념	6,564
박슬아_전산5_19장	응용운영체제개념	8,020
박슬아_전산5_20장	응용운영체제개념	14,316
박슬아_전산5_21장	응용운영체제개념	14,270
박슬아_전산5_22장	응용운영체제개념	13,756
박슬아_전산5_23장	응용운영체제개념	14,909
박슬아_전산5_24장	응용운영체제개념	20,399
박슬아_전산5_부록A	응용운영체제개념	3,623
박슬아_전산5_색인	응용운영체제개념	7,265
이진화_전산7_1장	알고리즘	15,310
이진화_전산7_2장	알고리즘	6,451
이진화_전산7_3장	알고리즘	5,522
이진화_전산7_4장	알고리즘	14,796
이진화_전산7_5장	알고리즘	9,108
이진화_전산7_6장	알고리즘	7,241
이진화_전산7_7장	알고리즘	9,842
이진화_전산7_8장	알고리즘	6,721
이진화_전산7_9장	알고리즘	8,278
이진화_전산7_10장	알고리즘	14,536
이진화_전산7_11장	알고리즘	10,818
이진화_전산7_12장	알고리즘	18,478
이진화_전산7_13장	알고리즘	7,378

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

이진화_전산7_찾아보기	알고리즘	598
김동휘_전자전기1_1장	마이크로전자회로	10,476
김동휘_전자전기1_2장	마이크로전자회로	9,951
김동휘_전자전기1_3장	마이크로전자회로	15,219
김동휘_전자전기1_4장	마이크로전자회로	16,087
김동휘_전자전기1_5장	마이크로전자회로	20,363
김동휘_전자전기1_6장	마이크로전자회로	14,059
김동휘_전자전기1_7장	마이크로전자회로	12,285
김동휘_전자전기1_8장	마이크로전자회로	13,153
김동휘_전자전기1_9장	마이크로전자회로	9,048
김동휘_전자전기1_10장	마이크로전자회로	12,438
전자전기1_11장	마이크로전자회로	14,633
전자전기1_12장	마이크로전자회로	13,105
전자전기1_13장	마이크로전자회로	15,409
전자전기1_14장	마이크로전자회로	11,739
전자전기1_부록A	마이크로전자회로	2,534
전자전기1_부록B	마이크로전자회로	876
전자전기1_부록C	마이크로전자회로	2,778
전자전기1_부록D	마이크로전자회로	908
전자전기1_부록E	마이크로전자회로	3,958
전자전기1_부록F	마이크로전자회로	51
전자전기1_찾아보기	마이크로전자회로	2,579
이진화_전자전기2_1장	신호와 시스템	5,003
이진화_전자전기2_2장	신호와 시스템	8,865
이진화_전자전기2_3장	신호와 시스템	5,454
이진화_전자전기2_4장	신호와 시스템	7,172
이진화_전자전기2_5장	신호와 시스템	6,864
이진화_전자전기2_6장	신호와 시스템	5,699
이진화_전자전기2_7장	신호와 시스템	5,145
이진화_전자전기2_8장	신호와 시스템	3,961
이진화_전자전기2_9장	신호와 시스템	5,608
이진화_전자전기2_10장	신호와 시스템	5,609
이진화_전자전기2_찾아보기	신호와 시스템	885
전자전기4_1장	이산신호처리	1,856
전자전기4_2장	이산신호처리	11,069
전자전기4_3장	이산신호처리	8,084
전자전기4_4장	이산신호처리	8,379
전자전기4_5장	이산신호처리	11,552
전자전기4_6장	이산신호처리	18,097
전자전기4_7장	이산신호처리	4,473

전자전기4_찾아보기	이산신호처리	487
데이터베이스시스템_찾아보기(영문)	데이터베이스시스템	562
데이터베이스시스템_찾아보기(한글)	데이터베이스시스템	1,169
전자기학_찾아보기	전자기학	912
컴퓨터시스템구조_찾아보기	컴퓨터시스템구조	2,334
컴퓨터시스템구조론_찾아보기	컴퓨터시스템구조론	776

○ 세종 결과물 말뭉치 목록

파일명	자료명	어절
7CHTC001.hwp	C언어 프로그래밍	40,413
7CHTC002.hwp	컴퓨터 과학 총론	83,227
7CHTC003.hwp	컴퓨터 공학	91,210
7CHTC004.hwp	C로 쓴 자료구조론	96,166
7CHTC005.hwp	한국어판 응용운동체제개념	222,310
7CHTC006.hwp	알고리즘	135,077
7CHTE007.hwp	마이크로 전자회로	201,649
7CHTE008.hwp	신호와 시스템	60,265
7CHTE009.hwp	이산신호처리	63,997
7CHTE0010.hwp	전자기학	912
7CHTC0011.hwp	데이터베이스시스템	1,731
7CHTC0012.hwp	컴퓨터시스템구조	2,334
7CHTC0013.hwp	컴퓨터시스템구조론	776
계		1,000,067

3. 연구 추진 방법

3.1 말뭉치 구축 방법

○ 구축 방법

- KORTERM과의 협력관계가 있기 때문에 기존의 '전문용어정비' 분과에서 말뭉치를 활용할 수 있도록 KORTERM의 입력방식을 따른다.
- 말뭉치의 효율적 관리와 이용에 필요한 헤더 정보와 마크업을 첨부한다.
- 단 마크업은 KORTERM 입력방식을 따르고, 헤더 정보와 문서 형식은 세종 말뭉치 전체에 대한 일관성 때문에 '세종 말뭉치 표준화 지침'(1998)에 의거하여 통일한다.
- 품질 수준의 유지에 유의하여, 최종 오류율이 0.1% 이내가 되도록 한다.

3.2 표준화 구축 지침

○ 파일 형식

- KORTERM에서 구축해온 말뭉치의 파일 형식은 한 책의 한 장(chapter)을 확장자가 '*.txt'인 한 파일로 만들었다. 그러나 올해부터 전문 용어 말뭉치가 국어 특수자료 구축 분과의 결과물로 통합되어 제출되게 되므로, 전문용어 말뭉치 개발의 결과물은 세종계획에서 구축하는 말뭉치의 기본 틀인 TEI 헤더 방식과 1책 1파일 원칙을 지킨다.

○ 파일 이름

- 구축 연차, 구축 기관, 매체와 장르, 자료의 성격을 반영하여 다음과 같이 입

력한다.

- 첫 번째 자리: 구축 연차 7
- 두 번째 자리: 구축 기관 C(연세대학교)
- 세 번째 자리: 매체와 장르 H(책)
- 네 번째 자리: 자료의 성격 T(for Terminology Research)
- 다섯 번째 자리: 용어의 분야
C(전산학 분야)
E(전자전기학 분야)
- 여섯 번째 자리 이후: 일련 번호

예) 연세대에서 구축한 책-정보 컴퓨터 관련 전문용어 말뭉치의 첫 파일이라면 다음과 같이 기입한다.

: 7CHTC001.hwp

○ 텍스트 분류

첫째 자리

- 문어/구어의 분류: 1(문어)

둘째 자리

- 매체에 따른 분류: 3(책)

셋째 자리/넷째 자리

- 내용에 따른 분류: 7(자연)
2(공학)

○ 입력 환경

- 말뭉치를 KORTERM에서 구축해오던 대로 작업하여 제공하기 위해 텍스트 파일(*.txt)로 입력하였으나 세종 말뭉치에 맞게 '훈글' 파일로 변환한다.

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

- 전체를 바탕 스타일로 하고, 그 속성은 다음과 같다.

글자모양

- 명조
- 10 포인트
- 보통 모양
- 글자간격 0%
- 장평 100%

문단모양

- 왼쪽, 오른쪽, 문단간격, 낱말간격 모두 0
- 줄간격 150%
- 정렬방식 왼쪽

편집용지

- A4
- 용지방향 좁게
- 위/아래/왼쪽/오른쪽/머리말/꼬리말 모두 15.00mm, 제본

0mm

페이지

- 표시하지 않음

머리말/꼬리말

- 표시하지 않음

○ 본문의 자료 표준화

- 본문의 자료는 KORTERM의 방식을 따른다. 그러나 전문 용어 말뭉치가 국어 특수자료 구축 분과의 결과물로 통합되어 제출되게 되므로 앞으로 KORTERM과 협의하여 '전문 용어 말뭉치 개발'의 본문 자료를 표준화하여 지침을 마련한다.

3.2 본문 입력 방법

3.2.1 헤더

○ 헤더(header) 구조의 예

```
<!DOCTYPE tei.2 SYSTEM "c:\sgml\dtd\tei2.dtd" [
    <!ENTITY % TEL.corpus "INCLUDE">
    <!ENTITY % TEL.extensions.ent SYSTEM "sejong1.ent">
    <!ENTITY % TEL.extensions.dtd SYSTEM "sejong1.dtd">
]>
<tei.2>
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>컴퓨터 공학</title>
        <author>권영빈, 김준년 공역</author>
        <sponsor>대한민국 문화관광부</sponsor>
        <respStmt><resp>전자/표준화</resp>
          <name>연세대학교 언어정보연구원</name>
        </respStmt>
      </titleStmt>
      <extent>91,210 어절</extent>
      <publicationStmt>
        <distributor>국립국어원</distributor>
        <idno>7CHTC003.hwp</idno>
        <availability><p>배포 불가</p></availability>
      </publicationStmt>
```

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

```
<sourceDesc>
  <bibl><author>권영빈, 김준년 공역</author>
    <title>컴퓨터 공학</title>
    <pubPlace>한국</pubPlace>
    <publisher>브레인코리아</publisher>
    <pubDate>2003.3.12</pubDate>
  </bibl>
</sourceDesc>
</fileDesc>
<encodingDesc>
  <projectDesc><p>21세기 세종계획 3단계 1차년도 말뭉치 구축</p></projectDesc>
  <samplingDecl><p>선정 자료를 Mark-Up 톨로 입력</p></samplingDecl>
  <editorialDecl><p>KORTERM의 입력 방식을 따름</p></editorialDecl>
</encodingDesc>
<profileDesc>
  <creation><date>2004</date></creation>
  <langUsage>
    <language id=KO usage=99>한국어</language>
  </langUsage>
  <textClass>
    <catRef scheme='SJ21' target='1372'>책: 자연, 공학</catRef>
  </textClass>
</profileDesc>
<revisionDesc>
  <change><date>2004</date>
    <respStmt>
      <resp>프로젝트책임자</resp><name>이원표</name>
```

```

        </respStmt>
        <item>선정된 자료를 전산 입력</item>
    </change>
    <change><date>2004.7.29</date>
        <respStmt><resp>연구보조원</resp><name>유형욱</name></respStmt>
        <item>자료 입력, 마크업</item>
    </change>
    <change><date>2004.8.8</date>
        <respStmt><resp>연구보조원</resp><name>유형욱</name></respStmt>
        <item>본문 내용 수정, 마크업, 태그 오류 수정</item>
    </change>
    <change><date>2004.11.28</date>
        <respStmt><resp>연구보조원</resp><name>최정도</name></respStmt>
        <item>마크업, 태그 오류 수정, 세종 21 프로젝트 헤더 붙임</item>
    </change>
</revisionDesc>
</teiHeader>

```

3.2.2 문서의 구조 및 본문

○ 2004년도 전문용어 말뭉치 개발에서는 KORTERM에서 제공한 '수식입력 Tool'을 사용한다. 말뭉치 입력은 TAG 지침에 맞게 일일이 직접 입력하는 것이 주를 이루었으나 입력시의 오류를 최소로 줄이기 위해 아래와 같은 Tool을 사용하여 말뭉치를 입력한다. 아래에서는 KORTERM에서 제공한 '수식입력 Tool'을 사용해

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

KORTERM의 체계에 맞게 말뭉치를 입력하는 방법(지침)을 소개한다.

○ 아래 그림은 본 프로그램의 실행 화면을 나타낸다. 그림에서 1번창은 코퍼스 구축자들이 코퍼스를 구축하면서 서지정보를 입력하는 창으로써 1번창에 모든 정보가 채워야 작업된 코퍼스가 저장된다.

The screenshot shows the KORTERM program interface. The top part is a metadata entry form with fields for Title, Author, Publisher, etc. The bottom part is a text editor showing a sample text with LaTeX formulas. Two callout boxes point to specific parts of the interface: '수식 입력의 예' (Example of formula input) points to a LaTeX formula in the text editor, and '문장 입력 예' (Example of sentence input) points to a sentence in the text editor.

타이틀 :	저자 :	출판기관 :	자료구축대상 :	기준명 :	이월수 :
수학의 기	이원 오	대한민국	수학발달지국	국립국립	1000
비표자 :	비표명 :	출판일자 :	출판처 :	출판사 :	출판일 :
국립국립	국립국립	2000.11.1	국립국립	국립국립	2000.11.1
프로젝트명 :	구축방법 :	편집자명 :	엑스텐션명 :	엑스텐션수 :	
수학1000년	국립국립	수학1000년	2000.10.29	2000.10.31	

1

2

수식 입력의 예

문장 입력 예

- 1번 창에서 설정된 정보들은 정해진 마크업 언어로 저장된다.

- 2번창은 텍스트를 편집할 수 있는 기능을 제공한다. 마크업 코퍼스를 구축하는 편집자는 여기에서 텍스트를 입력하고 편집을 수행한다. 2번창에서 작업자의 편의를 제공하기 위해 단축 메뉴를 제공한다. 제공 메뉴는 프로그램 상에 [메뉴] ->

[문장구성기호]를 선택하면 제공되는 리스트를 볼 수 있다. 메뉴리스트는 아래 그림과 같은 리스트를 제공한다.

가) 본문: <text></text>

텍스트들에 대해서 제목이 출현할 때 태그를 추가하여 사용

나) 절/단원제목: <head></head>

텍스트들에 대해서 제목의 정보를 입력할 때 태그를 추가하여 사용

다) 단원 : <body></body>

텍스트들에 대해서 제목의 정보 다음부터 나오는 정보들을 입력할 때 태그를 추가하여 사용

라) 문단구분 : <p></p>

텍스트들에 대해서 Body 정보를 입력시 절단위를 입력할 때 태그를 추가하여 사용

마) 문장구분 : <s></s>

텍스트들에 대해서 문단 정보를 입력시 문장을 입력할 때 태그를 추가하여 사용

바) 표 : <table></table>

텍스트들에 대해서 문장 정보를 입력시 표를 입력할 때 태그를 추가하여 사용

사) 그림 : <fig></fig>

텍스트들에 대해서 문장 정보를 입력시 그림을 입력할 때 태그를 추가하여 사용

아) 굵은글씨 :

텍스트들에 대해서 문장 정보를 입력시 굵은 글씨를 입력할 때 태그를 추가하여 사용

자) 이탤릭체 : <i></i>

텍스트들에 대해서 문장 정보를 입력시 이탤릭체 글씨를 입력할 때 태그를 추가하여 사용

차) 라텍기호 : <TEX></TEX>

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

텍스트들에 대해서 문장 정보를 입력시 수식을 입력할 때 태그를 추가하여 사용

- 2번창에서 파란 박스로 표시된 부분이 문장입력을 위한 예를 나타낸 것이고 빨간 박스로 표시된 부분이 수식입력을 위한 예를 나타낸다.

문장구성기호 헤드태커달기	
문문:	<text> </text>
절/단원제목:	<head> </head>
단원	<body> </body>
문단구분:	
문장구분:	<s> </s>
표 :	<table> </table>
그림:	<fig> </fig>
굵은글씨:	
이탤릭체:	<i> </i>
라텍기호:	<TEX> </TEX>

- 처음의 창에 뜨는 서지 정보는 다음과 같이 입력한다(이 서지 정보는 KORTERM에서 말뭉치에 부착하는 헤더이다).

- ㄱ. 타이틀- 입력하는 책의 한글 제목
- ㄴ. 저자- 원서가 아닌 이상에는 번역자의 이름을 쓴다. 번역자가 다수이면 그 다수를 다 기입한다.
- ㄷ. 출원기관- 대한민국
- ㄹ. 자료구축내용- 우리가 구축하는 자료의 내용은 '전산학'분야와 '전자전기학' 분야 두 부문이므로 이 둘 중 하나를 맞게 기입한다. ex) 전산학말뭉치구축 혹은

전자전기학말뭉치구축

- ㄱ. 기관명- 우선 'temp'로 기입한다.
- ㄴ. 어절수- 100만
- ㄷ. 배포자- temp
- ㄹ. 파일명- 본인 이름과 전산 혹은 (전자전기)의 순위와 몇 장인지로 기입한다.
예) 박길동_전산2_3장 => 입력 자료 순위 참고
- ㅈ. 원전저자- 한국책이면 한국사람, 외국책이면 외국저자를 쓰면 된다.
- ㅊ. 원전제목- 마찬가지로 외국도서이면 원래의 제목을 적는다.
- ㅋ. 출판지- 당연히 '한국'(외국에서 수입한 책, 원서가 아닐 경우).
- ㅌ. 출판사- 책 뒤에 보면 자세히 나와 있다. 정확한 명칭을 쓴다.
- ㅍ. 출판일- 위와 마찬가지로
- ㅎ. 프로젝트명- 전문용어말뭉치개발
- ㄱ. 구축방법- 'Mark-up 입력'이라고 쓴다.
- ㄴ. 편집자설명- 작업하는 책의 제목과 몇 장 입력이라고 씀. 예) 전자기학2장
입력
- ㄷ. 텍스트생성일- 해당 파일의 작업 시작 날짜.
- ㄹ. 텍스트수정일- 우선은 위와 같은 날짜로 기입. 그리고 수정할 때마다 그날
을 다시 기입.

- 하나의 파일을 이루는 기본 단위는 한 권의 책이 아니라, 한 책의 한 장 (Chapter)이다. 예를 들어 하나의 책이 7장으로 되어 있다면, 하나의 파일에 한 장 (Chapter)을 입력한다.

- 단순히 책의 내용을 입력하는 것이 아니라 TAG 지침에 맞게 각 내용에 TAG를 달아야 한다. 이때 착오가 없어야 한다. TAG를 달 때의 대원칙은 "TAG를 열면 꼭 달아야한다"라는 것을 명심한다.

ex) <text> 표준국어문법론 </text>
 열면 닫는다

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

- 본문 입력 방법

가) 본문: <text></text>

텍스트들에 대해서 제목(소제목)이 출현할 때 태그를 추가하여 사용.

ex) text의 전체 짜임은 'head'와 'body'로 이루어져 있다.

```
<text>
<head>
</head>

<body>
</body>
</text>
```

ex) 주의: 제목을 쓸 때에는 항상 그 Chapter의 제목을 반복해서 써야 한다.

- 2장 1절(2.1) 입력시

```
<head>
2. 유클리드의 원론
<head>
2.1 유클리드 원론의 중요성과 형식적인 특성
</head>
</head>
```

- 2장 2절(2.2) 입력시

```
<head>
2. 유클리드의 원론
<head>
2.2 유클리드 원론의 중요성과 형식적인 특성
```

</head>

</head>

나) 절/단원제목: <head></head>

텍스트들에 대해서 제목의 정보를 입력할 때 태그를 추가하여 사용

다) 단원 : <body></body>

텍스트들에 대해서 제목의 정보 다음부터 나오는 정보들을 입력할 때 태그를 추가하여 사용

라) 문단구분 : <p></p>

텍스트들에 대해서 Body 정보를 입력 시 절단위를 입력할 때 태그를 추가하여 사용

마) 문장구분 : <s></s> (← 모든 본문의 문장은 이 TAG를 사용해야 한다.)

텍스트들에 대해서 문단 정보를 입력 시 문장을 입력할 때 태그를 추가하여 사용

바) 표 : <table></table>

텍스트들에 대해서 문장 정보를 입력 시 표를 입력할 때 태그를 추가하여 사용

사) 그림 : <fig></fig>

텍스트들에 대해서 문장 정보를 입력 시 그림을 입력할 때 태그를 추가하여 사용

아) 굵은글씨 :

텍스트들에 대해서 문장 정보를 입력 시 굵은 글씨를 입력할 때 태그를 추가하여 사용

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

ex) 원론을 편집한 유클리드는 에우독소스의 많은 정리를 수집했다.

자) 이탤릭체 : <i></i>

텍스트들에 대해서 문장 정보를 입력 시 이탤릭체 글씨를 입력할 때 태그를 추가하여 사용

차) 라텍(수식)기호 : <TEX></TEX>

텍스트들에 대해서 문장 정보를 입력 시 수식과 자판으로 입력할 수 없는 특수문자·기호 등을 입력할 때 태그를 추가하여 사용

수식 입력은 LaTeX 매뉴얼에 따라 입력

※ 참고 : <http://www.maths.tcd.ie/~dwilkins/LaTeXPrimer/>

카) 연습

ex)

```

┌─── <text> (← 열기1)
|
|   ┌─── <head> (← 열기2)
|   |   썸 (← Chapter의 제목: 이 Chapter의 제목은 소단원마다 기입)
|   └─── </head> (← 닫기2)
|
|   ┌─── <body> (← 열기3)
|   |
|   |   ┌─── <p> (← 열기4)
|   |   |   ┌─── <s> (← 열기5)
|   |   |   |   자연수는 ... 존재하는 양식들인 '단일성' 양식, '이중성' 양식,
|   |   |   |   '삼중성' 양식 등에 대한 인식으로부터 발생한다.
|   |   |   └─── </s> (← 닫기5)
|   |   |   .....
|   |   └─── <s> (← 열기x)
|   |   |   이런 형태의 ... 연구를 <b>수론</b>(number theory)이라 부른다.
|   |   └─── </s> (← 닫기x)
|   └─── </p> (← 닫기4)
|
|   └─── </body> (← 닫기3)
|
└─── </text> (← 닫기1)

```

파일 저장

: 파일을 저장할 때는 txt 파일로 저장.

ex) 파일명.txt

이 확장자를 꼭 입력해주어야 한다.

: 파일명은 “이름_전산(혹은 전자전기)_장”의 방식으로.

ex) "홍길동_전산1_3장.txt", "박길동_전자전기1_3장.txt"

- 기타

수식과 표는 제목만 그 수식과 표의 이름만 기입한다.

오류를 최대한 줄인 말뭉치를 구축하기 위해, 신중히 입력한다.

4. KORTERM과 세종 말뭉치의 차이점

○ 전문용어 원시 말뭉치의 Mark-up

전문용어 원시 말뭉치를 구축함에 있어 기존의 KORTERM에서 구축해오던 것이 세종 특수말뭉치로 흡수가 되면서 KORTERM의 Mark-up과 연세대(세종 특수말뭉치)의 것이 서로 달라 문제가 된다. 여기서는 이들이 어떻게 차이가 나는지에 대해서만 밝히기로 한다.

이들의 차이를 보이기 위해 아래와 같이(<표1>) 『응용운영체제개념』의 일부분을 발췌하여, KORTERM과 세종 특수말뭉치 각각의 Mark-up(<표3>과 <표4>)될 예를 보이고, 서로의 Mark-up이 어떻게 다른지 보이기로 한다.

<표1> 원전

CHAPTER 20 UNIX 시스템 (The UNIX System)

운영 체제 개념들을 순수하게 이론적인 측면에서 살펴볼 수 있지만, 때로는 이들이 실제 어떻게 구현되는가를 살펴보는 것이 유용하다. 이 장에서는 이 책에 제시된 다양한 개념의 예로써 UNIX의 한 버전인, 4.3BSD 운영 체제에 대한 깊이 있는 검토를 제시한다. 완전한 실제적인 시스템을 살펴봄으로써 이 책에 제시된 다양한 개념들이 어떻게 서로간에 혹은, 실제와 연관되어 있는지를 알 수 있다. 우리는 먼저 UNIX의 간략한 역사를 살펴보고, 시스템 사용자 그리고 프로그래머 인터페이스를 제시한다. 그리고 나서, 사용자 프로그래머 인터페이스를 지원하기 위해 UNIX 커널에서 사용되는 내부 자료 구조와 알고리즘에 대해 논의한다.

20.1 역사(history)

UNIX의 첫 번째 버전은, 그 당시 사용되지 않고 있던 PDP-7을 사용하기 위해, 벨 연구소의 연구 그룹의 Ken Thompson에 의해 1969년에 개발되었다. 여기에 곧 Dennis Ritchie가 가담하였다. Thompson, Ritchie와 그 연구 그룹의 다른 연구원들은 UNIX의 초기 버전을 개발하였다.

.
.
 .

○ <text>, <div> tag

우선 두 Mark-up이 <text>를 가장 상위의 단위로 설정하는 것은 같다. 그러나 KORTERM은 한 권의 책의 하나의 단락(장, 章)의 소단락을 <text>로 설정(예, '1장'의 '1.1절', '1.2절' 등이 각각 하나의 <text>)하여, 수많은 <text>가 모여 하나의 파일(이때의 파일은 하나의 '단락 혹은 장')이 되는데 반해, 세종 특수말뭉치는 한

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

권 전체를 <text>로 설정하고 한 권의 책 전체가 하나의 파일을 이룬다.

개념이 서로 조금 다르기는 하지만 KORTERM의 <text>에 해당하는 것이 세종 특수말뭉치의 <div>이다. 수많은 <div>가 모여서 <text>를 이룬다. 이렇게 차이 나는 <text>와 <div>는 파일을 열어서 차례대로 읽고 수정하면 되므로 전체 Mark-up을 통일하는데 크게 문제가 되지 않는다.

Mark-up의 전체 체계를 굵직하게 다시 한 번 확인해 보면 다음 표와 같다.

<표2>

KORTERM	세종 특수말뭉치(연세대학교)
<text>	<text>
<head>	<body>
</head>	<div>
<body>	<head></head>
	</div>
</body>	<div>
</text>	<head></head>
<text>	
<head>	
</head>	</div>
<body>	<div>
	<head></head>
</body>	</div>
</text>	<div>
<text>	<head></head>
<head>	
</head>	
<body>	</div>
	.
	.
	.
</body>	</body>
</text>	</text>
.	
.	
.	

위 <표2>와 같이 KORTERM의 Mart-up은 하나의 장에서 소단락(절)이 <text>로 반복해서 나타나고, 세종 특수말뭉치의 Mart-up은 전체 <text>는 하나인데 그 안에 소단락(절)이 <div>로 반복해서 나타난다.

○ <p>, <s> tag

KORTERM에서 단락과 문장 기호인 <p>와 <s>는 <body> 안에 포함된다. 세종 특수말뭉치에서는 <p>와 <s>가 <div> 안에 포함된다. <표3>과 <표4>를 비교해 보면 그 차이를 한 눈에 알아볼 수 있다. KORTERM의 것은 <p>와 <s>를 위·아래(세로)로 Marking하는데 비해, 세종 특수말뭉치는 좌우(가로)로 Marking한다. 세종 특수말뭉치는 입력할 때 원전의 구성을 그대로 충실히 반영하는데 비해 KORTERM은 문장단위로 모두 분할하여 입력한다는 차이가 있다. 두 Mark-up이 서로 다르기는 하지만 이 부분도 일괄처리를 통해 변환하는 것은 문제가 되지 않는다.

○ <head> tag

지금 구축되고 있는 KORTERM 방식의 Mark-up을 세종 특수 말뭉치의 Mark-up으로 변환하려고 할 때 문제가 되는 것이 <head>이다. <표2, 3, 4>에서 확인되듯이 KORTERM의 <head>는 실제 본문이 입력되는 <body> 밖에 위치하는데, 세종 특수 말뭉치의 <head>는 본문이 입력되는 <div> 안에 위치한다. 그리고 <표3>과 <표4>에서 알 수 있듯이 세종 특수 말뭉치의 <head>는 현재의 <head>를 단 한 번만 실현하는데 비해, KORTERM의 <head>는 현재 입력되는 소단락 <head>의 상위의 소단락 <head>들을 계속해서 실현해 간다. 실제 이 <head> 부분이 일괄적으로 변환하고자 할 때 문제가 된다.

○ 기타

아래와 같이 파일 이름과 확장자가 다르다.

예) KORTERM : "유형욱_전산3_2장.txt"

세종 말뭉치 : "7CHTC003.hwp"

원전의 구성 특성 상 KORTERM의 Mark-up 체계로는 문단으로 구분하기 힘든 부분의 <head>를 궁여지책으로 문장 안에 포함시켜서 입력된 부분이 있는데, 이 부분의 <head>를 실현하기 위해서는 꼼꼼하게 원전과 다시 비교해야 하는 작업이 병행되어야 하는 어려움이 따른다.

<표3> - KORTERM

```
<text>
<head>
CHAPTER 20 UNIX 시스템 (The UNIX System)
</head>
<body>
<p>
<s>
운영 체제 개념들을 순수하게 이론적인 측면에서 살펴볼 수 있지만, 때로는 이
이 실제 어떻게 구현되는 가를 살펴보는 것이 유용하다.
</s>
<s>
이 장에서는 이 책에 제시된 다양한 개념의 예로써 UNIX의 한 버전인, 4.3BSD
운영 체제에 대한 깊이 있는 검토를 제시한다.
</s>
<s>
완전한 실제적인 시스템을 살펴봄으로써 이 책에 제시된 다양한 개념들이 어떻
게 서로간에 혹은, 실제와 연관되어 있는지를 알 수 있다.
</s>
<s>
우리는 먼저 UNIX의 간략한 역사를 살펴보고, 시스템 사용자 그리고 프로그래머
인터페이스를 제시한다.
</s>
<s>
그리고 나서, 사용자 프로그래머 인터페이스를 지원하기 위해 UNIX 커널에서 사
용되는 내부 자료 구조와 알고리즘에 대해 논의한다.
</s>
</p>
</body>
</text>
```

```

<text>
<head>
CHAPTER 20 UNIX 시스템 (The UNIX System)
<head>
20.1 역사(history)
</head>
</head>
<body>
<p>
<s>
UNIX의 첫 번째 버전은, 그 당시 사용되지 않고 있던 PDP-7을 사용하기 위해,
벨 연구소의 연구 그룹의 Ken Thompson에 의해 1969년에 개발되었다.
</s>
<s>
여기에 곧 Dennis Ritchie가 가담하였다.
</s>
<s>
Thompson, Ritchie와 그 연구 그룹의 다른 연구원들은 UNIX의 초기 버전을 개
발하였다.
</s>
</p>

```

<표4> - 연세대(세종)

```

<text>
<body>
.
.
.
<div>
<head>CHAPTER 20 UNIX 시스템 (The UNIX System)</head>
<p><s>운영 체제 개념들을 순수하게 이론적인 측면에서 살펴볼 수 있지만, 때로는 이들이 실제 어떻게 구현되는가를 살펴보는 것이 유용하다.</s><s>이 장에서는 이 책에 제시된 다양한 개념의 예로써 UNIX의 한 버전인, 4.3BSD 운영 체제에 대한 깊이 있는 검토를 제시한다.</s><s>완전한 실제적인 시스템을 살펴봄으로써 이 책에 제시된 다양한 개념들이 어떻게 서로간에 혹은, 실제와 연관되어 있는지를 알 수 있다.</s><s>우리는 먼저 UNIX의 간략한 역사를 살펴보고, 시스템 사용자 그리고 프로그래머 인터페이스를 제시한다.</s><s>그리고 나서, 사용자 프로그래머 인터페이스를 지원하기 위해 UNIX 커널에서 사용되는 내부 자료 구조와 알고리즘에 대해 논의한다.</s></p>
<head>20.1 역사(history)</head>
<p><s>UNIX의 첫 번째 버전은, 그 당시 사용되지 않고 있던 PDP-7을 사용하기 위해, 벨 연구소의 연구 그룹의 Ken Thompson에 의해 1969년에 개발되었다.</s><s>여기에 곧 Dennis Ritchie가 가담하였다.</s><s>Thompson, Ritchie와 그 연구 그룹의 다른 연구원들은 UNIX의 초기 버전을 개발하였다.</s></p>
.
.
.

```

5. 문제점 및 향후 계획

○ 문제점

- 2004년도 세종 사업에서 KORTERM의 '전문용어 정비' 분과에서 원시 말뭉치 구축이 세종 특수 말뭉치에 흡수되면서 몇 가지 문제점이 나타났다. 기존의

KORTERM에서 구축해온 원시 말뭉치의 구성이 세종의 원시 말뭉치 구성과 파일의 단위, 파일의 이름, Mark-up 구성 및 방법, header 등에서 차이를 보인다. 대표적으로 Mark-up의 경우 전문용어 말뭉치 구축의 대상이 대체로 이학·공학 계열이어서 이 분야의 원서에 특수문자(기호)와 수식이 많이 나타나는 바, 특수문자(기호)와 수식을 입력할 수 있는 LaTeX를 사용하여 입력했다. 이는 세종 말뭉치의 통일성을 고려했을 때 문제가 되는 부분이다. 그리고 전문용어 말뭉치 구축 대상 분야와 저작권 해결 등이 문제가 된다.

- 우선적으로는 ‘전문용어 말뭉치 개발’ 사업이 KORTERM과 협력관계에 있기 때문에 KORTERM에서 자료를 활용할 수 있도록 KORTERM의 구축 지침을 따라서 입력하고 말뭉치를 KORTERM에 제공했다. 그리고 헤더입력과 파일의 구성, 파일의 이름 등은 세종 말뭉치의 일관성을 고려해서 국립국어원과 협의 하에 기존의 세종 말뭉치 표준화를 따랐다.

○ 향후 계획

- 향후 전문용어 말뭉치를 구축하기 위한 표준화 작업이 KORTERM과 국립국어원, 언어정보연구원 사이에 있어야 할 것이다.

- 올해 2004년 3단계 1차년도 전문용어 원시 말뭉치의 본문 구성과 Mark-Up이 기존의 세종 말뭉치의 구성과 다르기 때문에 이후 사업에서 점차적으로 세종 말뭉치에 맞게 수정·보완하기로 한다.

여 백

제 2 부

형태소 분석 말뭉치의 구축

여 백

제 1 장

현대 국어 구어 전사 말뭉치 개발

연구책임자 : 김 향 희

여 백

제 1 장

현대 국어 구어 전사 말뭉치 개발

1. 연구의 필요성 및 목표

1.1. 연구의 필요성

광범위한 원시 말뭉치의 구축은 국가 언어 자원 확보에 근간을 이루는 것이지만, 구축된 구어 전사 말뭉치가 여러 분야에서 실질적으로 활용되기 위해서는 기계 가독형이 되어야 하는데, 이를 원활히 하기 위해서는 형태 분석 말뭉치의 개발이 필수적이다. 구어 자료는, 문어 자료와는 달리 발화 장면, 발화 목적, 화자의 연령 및 성별, 개인적인 언어 습관 등에 따라 문어 자료에서는 잘 드러나지 않는 다양한 변이형이 나타나며, 발화의 구성 또한 복잡하게 실현된다. 그러므로, 이러한 구어 전사 말뭉치의 형태 분석을 위해서는 다양한 구어적 변이형을 조사하고 분류하는 기초 연구가 필요하며, 수작업을 거쳐 정제된 다량의 형태 분석 말뭉치가 구축되어 있어야 한다.

광범위한 영역에서 세심히 구축된 구어 자료와 이에 대한 기초 가공 작업으로서의 형태 분석 말뭉치 개발은 대규모 국가 언어 자료의 기반이며, 동시에 자연 언어 처리 및 언어 공학 분야에서의 활용도를 높일 수 있는 자료로서 제공될 수 있을 것이다.

1.2. 연구 목표

2004년도 구어 형태소 분석 말뭉치 구축에서는 균형성 및 다양성을 갖춘 말뭉치 개발을 위하여 기구축 말뭉치에서 보완이 필요한 장르에 초점을 맞춰서 18만 어절의 형태소 분석 말뭉치를 개발한다.

2. 구어 전사 형태소 분석 말뭉치 개발

2.1. 구축 도구

2.1.1. 1차 주석

언어정보연구원에서 자체 개발한 형태 주석 편집기 SORITAG를 통해 반자동 방식으로 1차 주석을 진행하였다.

○작업 공정

- “어절분석사전” 생성
 - 이미 주석된 말뭉치를 토대로 사전 생성
 - 필요에 따라 주석 사전의 내용 추가 및 수정이 가능함
- 원시 말뭉치 파일 불러오기/태깅 작업
 - “어절분석사전”을 불러오기 해서 “어절일치분석”으로 자동 분석
 - 미분석 어절은 “어절추정분석” 과정을 통해 보충: 원어절을 표제로 제시된 분석 후보 어절 중 해당되는 것을 클릭하면 주석 편집창에 결과가 반영됨

<<SORITAG를 이용한 “어절일치분석” 처리 결과>>

파일 형태 주석 화용 주석 도구 도움말			
일련번호	원어열	분석어열	참고
005754	<u who=P2>		
005755	<s n=00953>		
005756	교리.	교리/NNG+./SP	
005757	</s>		
005758	<s n=00954>		
005759	교리라고	교리/NNG+(O)/VCP+라고/EC	
005760	매기발	매기/NNG+하/XSV+e/ETM	
005761	수	수/NNB+는/JX	
005762	없고	없/V+고/EF+./SP	
005763	</s>		
005764	</u>		
005765	<u who=P1>		
005766	<s n=00955>		
005767	미	미/IC+./SF	
005768	</s>		
005769	</u>		
005770	<u who=P2>		
005771	<s n=00956>		
005772	기독교	기독교/NNG	
005773	문화적인	문화적/NNG+미/VCP+L/ETM	
005774	게	게/NNB+I/JKS	
005775	보수적이라고.	보수적/NNG+미/VCP+라고/EC+./SF	
005776	</s>		
005777	</u>		
005778	<u who=P1>		
005779	<s n=00957>		

미질분석사전
<분석> <개>
1 일치 거/NNB+I/JKS
2 일치 거/NNB+I/JKS
3 일치 거/NNB+I/JKB
4 일치 거/NP+I/JKC
5 일치 거/NNB+미/JKS
6 일치 것/NNB+미/JKS
7 일치 게/NP
8 일치 게/EC
9 일치 게/MAG
10 주절 /VV+게/EC
11 주절 게/MAJ
12 주절 /VA+게/EC
13 주절 /VX+게/EC
14 주절 게/VA
15 주절 거/NP+I/JKS
16 주절 것/NP+미/JKS
17 주절 거/NP+I/JKG
18 주절 /VCP+게/EC
19 주절 거/NNG+I/JKC
20 주절 /VV+게/NNB
21 주절 게/NNG
22 주절 /VV+게/EF
23 주절 게/NNP1
24 주절 /NP+게/JKB
25 주절 게경우/NNG+는/JX+./SP
26 주절 게미/IC+./SF

○ 개선 효과

- SORITAG는 기존의 형태 분석 말뭉치를 이용한 주석 편집 도구로서, 기존에 생성된 형태 분석 말뭉치를 데이터베이스로 하여 사전을 생성하고 이에 근거하여 주석 작업을 진행하게 되므로 주석 작업이 진행될수록 분석률과 분석의 정확도가 높아진다.
- 지능형 형태소 분석기를 이용했을 시 발생했던 짝태그 인식 오류 등의 문제가 발생하지 않아 후처리 작업을 대폭 줄일 수 있다.
- 기존에 생성된 형태 분석 말뭉치를 데이터베이스로 하여 사전을 생성하므로 여러 가지 이형태가 많이 나타나는 구어 주석 작업의 효율성을 높일 수 있다.
- 이용자 중심의 인터페이스 환경을 구현하고자 작업자의 편의를 최대한 고려하여 설계하였으며, 지속적으로 개선 작업 진행 중이다.

2.1.2. 2차 주석

동시에 여러 개의 파일 처리가 가능한 텍스트 에디터를 통해 수작업으로 2차 주석 작업을 진행하였다. 이 과정에서 1차 주석에서의 오분석 어절을 수정하였으며, 전체적인 통일성을 기하고자 노력하였다.

2.2. 말뭉치 목록

기 구축분 구어 형태 분석 말뭉치 504,496어절의 목록은 다음과 같다.

<기 구축 구어 형태 분석 말뭉치 목록>

파일명_주석	제목	전체 어절수	저작권 (원시)	비율
5CT_0006.hwp	TV방송_도올강의_왓이란	5,781	×	
5CT_0007.hwp	TV방송_성태용강의	3,662	×	
1인담화/공적	방송	9,443		1.87%
5CT_0046.hwp	강의_교양 국어	3,783	○	
5CT_0047.hwp	강의_대화의 기법	4,696	○	
5CT_0033.hwp	강의_목사	4,246	○	
5CT_0048.hwp	강의_언어와 사회	6,860	○	
6CT_0002.hwp	강의_연구방법론	12,549	○	
1인담화/공적	일반/강의	32,134		6.37%
5CT_0049.hwp	발표-대화의기법#1	1,245	○	
5CT_0050.hwp	발표-대화의기법#2	990	○	
5CT_0051.hwp	발표-대화의기법#3	1,242	○	
5CT_0052.hwp	발표-대화의기법#4	2,240	○	
5CT_0053.hwp	발표-대화의기법#5	1,154	○	
5CT_0054.hwp	발표-스터디	2,749	○	
5CT_0055.hwp	발표_워크샵#1	3,361	○	
5CT_0056.hwp	발표_워크샵#2	4,612	○	
5CT_0057.hwp	발표_워크샵#3	3,146	○	
5CT_0058.hwp	발표_워크샵#4	4,397	○	

제1장 현대국어 구어 전사 말뭉치 개발

6CT_0003.hwp	발표_대학원수업	5,839	○	
1인담화/공적	일반/발표	30,975		6.14%
5CT_0034.hwp	독백_연애담과신혼여행기	3,132	○	
5CT_0039.hwp	독백_잡담#1	314	○	
5CT_0040.hwp	독백_잡담#2	1,179	○	
5CT_0041.hwp	독백_잡담#3	1,963	○	
5CT_0036.hwp	독백_동화들려주기#1	5,173	○	
5CT_0037.hwp	독백_동화들려주기#2	1,467	○	
5CT_0038.hwp	독백_동화들려주기#3	579	○	
6CT_0004.hwp	독백_동화(백설공주)	610	○	
6CT_0005.hwp	독백_사랑이야기#1	1,027	○	
6CT_0006.hwp	독백_사랑이야기#2	2,037	○	
6CT_0007.hwp	독백_사랑이야기#3	466	○	
6CT_0008.hwp	독백_선생님	339	○	
6CT_0009.hwp	독백_아르바이트경험	11,494	○	
6CT_0010.hwp	독백_어린시절의우정	2,435	○	
6CT_0011.hwp	독백_어린시절의추억	2,120	○	
6CT_0012.hwp	독백_여행이야기	2,219	○	
6CT_0013.hwp	독백_연애담과신혼여행기	3,132	○	
6CT_0014.hwp	독백_영화이야기#1	1,189	○	
6CT_0015.hwp	독백_영화이야기#2	1,677	○	
6CT_0016.hwp	독백_영화이야기#3	1,600	○	
6CT_0017.hwp	독백_영화이야기#4	2,174	○	
6CT_0018.hwp	독백_영화이야기#5	2,539	○	
6CT_0019.hwp	독백_영화이야기#6	1,431	○	
6CT_0020.hwp	독백_외국인학생	2,102	○	
6CT_0021.hwp	독백_장래희망의변화	1,613	○	
6CT_0022.hwp	독백_추천영화베스트7	2,070	○	
6CT_0023.hwp	독백_친구	1,356	○	
6CT_0024.hwp	독백_학창시절	2,129	○	
1인담화/사적	독백	59,566		11.81%
5CT_0001.hwp	TV방송_100분토론(2001/08/30)	11,129	×	
5CT_0002.hwp	TV방송_세상사는이야기(1992/7/11)	5,117	×	
5CT_0003.hwp	TV방송_심야토론(2001/8/25)	13,735	×	
5CT_0004.hwp	TV방송_sbs공방(2001/06/29)	9,945	×	
5CT_0005.hwp	TV방송_sbs공방(2001/08/31)	10,717	×	
5CT_0008.hwp	라디오방송_그건이렇습니다(97/6/18)	4,047	×	
5CT_0009.hwp	라디오방송_서울서울서울(97/6/18)	6,077	×	
4CT_0001.hwp	라디오방송_여성시대(19960401, 3,4부)	5,063	×	
5CT_0010.hwp	라디오방송_이지현의라디오상담 (01/09/13)#1	4,040	×	
5CT_0011.hwp	라디오방송_이지현의라디오상담 (01/09/13)#2	3,578	×	

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

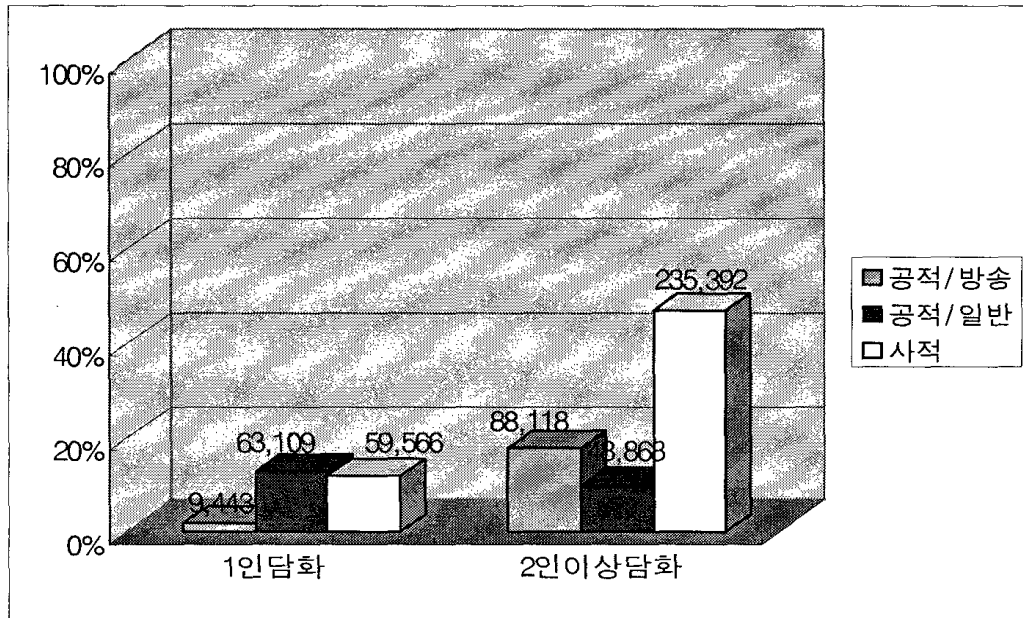
5CT_0012.hwp	라디오방송_SBS건강상담실(97/6/18)	3,015	×	
6CT_0001.hwp	방송_mbc100분토론(010913)	11,655	×	
2인이상담화/공적	방송	88,118		17.47%
4CT_0008.hwp	토론_경영학과토론(델타조)	5,293	×	
4CT_0009.hwp	토론_경영학과토론(잡동사니조)	2,885	×	
4CT_0002.hwp	토론_경영학과토론1(금융실명제)	5,491	×	
4CT_0003.hwp	토론_경영학과토론2(논문주제선정)	6,384	×	
5CT_0042.hwp	토론_언어생활#1	3,368	○	
5CT_0043.hwp	토론_언어생활#2	4,360	○	
5CT_0044.hwp	토론_언어와사회#1	2,161	○	
5CT_0045.hwp	토론_언어와사회#2	3,116	○	
6CT_0025.hwp	토론_논문주제선정(누네피네조)	736	×	
6CT_0026.hwp	토론_논문주제선정(미녀와야수조)	3,740	×	
2인이상담화/공적	일반/토론	37,534		7.44%
4CT_0005.hwp	상담_대인관계상담(19920515)	11,334	×	
2인이상담화/공적	일반/상담	11,334		2.25%
5CT_0013.hwp	일상대화_교통수단,하루생활	3,730	○	
5CT_0014.hwp	일상대화_날씨에대해	4,109	○	
5CT_0015.hwp	일상대화_물품구입	6,063	○	
5CT_0016.hwp	일상대화_미팅	14,234	○	
5CT_0017.hwp	일상대화_삼십대	14,937	○	
4CT_0004.hwp	일상대화_서클 ycc 잡담	7,813	×	
5CT_0018.hwp	일상대화_식사중 대학생3인	4,213	○	
5CT_0019.hwp	일상대화_식사중 회사원 3인	3,310	○	
5CT_0020.hwp	일상대화_식생활에 대해	4,720	○	
5CT_0021.hwp	일상대화_아버지의학교생활	4,701	○	
5CT_0022.hwp	일상대화_운전에 대해	4,608	○	
4CT_0006.hwp	일상대화_잡담#1	2,733	×	
4CT_0007.hwp	일상대화_잡담#2	3,686	×	
5CT_0023.hwp	일상대화_취미에대해	1,742	○	
6CT_0027.hwp	일상대화_가족과사랑에대해	9,140	○	
6CT_0028.hwp	일상대화_강의시작전7인#1	1,012	○	
6CT_0029.hwp	일상대화_강의시작전7인#2	3,728	○	
6CT_0030.hwp	일상대화_관광명소에대해	7,766	○	
6CT_0031.hwp	일상대화_교육에대해#2	4,128	○	
6CT_0032.hwp	일상대화_대학생2인잡담#1	1,151	○	
6CT_0033.hwp	일상대화_대학생2인잡담#2	969	○	
6CT_0034.hwp	일상대화_대학생3인잡담	2,196	○	
6CT_0035.hwp	일상대화_대학생4인잡담	5,840	○	
6CT_0036.hwp	일상대화_도서관에서	2,186	○	
6CT_0037.hwp	일상대화_동아리	689	○	
6CT_0038.hwp	일상대화_머리에대해서#2	1,498	○	
6CT_0039.hwp	일상대화_방학에 대해	3,898	○	

제1장 현대국어 구어 전사 말뭉치 개발

6CT_0040.hwp	일상대화_버스에서친구들과	2,134	○	
6CT_0041.hwp	일상대화_수강신청과목	2,752	○	
6CT_0042.hwp	일상대화_식사중대학생2인#1	2,391	○	
6CT_0043.hwp	일상대화_식사중대학생2인#2	1,916	○	
6CT_0044.hwp	일상대화_여대생10인잡담	3,804	○	
6CT_0045.hwp	일상대화_여대생2인잡담	4,218	×	
6CT_0046.hwp	일상대화_잡담#3	775	×	
6CT_0047.hwp	일상대화_재수강과목에대해	2,038	○	
6CT_0048.hwp	일상대화_후배들과대화	533	○	
6CT_0049.hwp	일상대화_후배와의대화	2,060	○	
2인이상대화/사적	일상대화	147,421		29.22%
5CT_0025.hwp	전화대화_홈쇼핑#1	6,269	×	
5CT_0026.hwp	전화대화_홈쇼핑#2	4,019	×	
5CT_0027.hwp	전화대화_홈쇼핑#3	6,005	×	
5CT_0028.hwp	전화대화_홈쇼핑#4	4,799	×	
5CT_0029.hwp	전화대화_20대 2인	335	○	
5CT_0030.hwp	전화대화_대학생2인	1,629	○	
5CT_0031.hwp	전화대화_여대생 2인	1,346	○	
5CT_0032.hwp	전화대화_이십대 남자 2인	407	○	
6CT_0050.hwp	전화대화_자료수집	345	○	
2인이상대화/사적	전화대화	25,154		4.98%
5CT_0024.hwp	주점대화_대학생 10인 이상	4,542	×	
6CT_0051.hwp	주점대화_대학생3인	13,349	○	
2인이상대화/사적	주점대화	17,891		3.55%
6CT_0052.hwp	주제대화_감기이야기	3,522	○	
6CT_0053.hwp	주제대화_건강이야기	5,879	○	
6CT_0054.hwp	주제대화_생일결혼이야기	1,396	○	
6CT_0055.hwp	주제대화_연애에피소드	8,576	○	
6CT_0056.hwp	주제대화_직장생활	2,959	○	
6CT_0057.hwp	주제대화_학원강사와고등학생	7,529	○	
6CT_0058.hwp	주제대화_향수와영화	9,475	○	
6CT_0059.hwp	주제대화_황사	1,045	○	
2인이상대화/사적	주제대화	40,381		8.00%
5CT_0035.hwp	기타_회의와 스터디#1	4,545	○	
2인이상대화/사적	기타	4,545		0.90%
합계		504,496		100%

위의 표에서 볼 때 강의, 발표 등 공적 독백 자료가 전체 14.38%, 사적 독백 자료가 11.81%, 토론, 상담 등 공적 대화 자료가 27.16%, 나머지 46.65%가 사적 대화 자료임을 알 수 있다. 장르 면에서는 공적 자료가, 화자수와 관련해서는 독백 자료가 지속적으로 보완되어야 함을 알 수 있다.

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과



이에 2004년 형태 분석 말뭉치 개발에서는 이러한 자료들을 중점적으로 선별하여 주석 작업을 진행하였다. 아래의 표에서 보면 사적 독백 자료가 전체 22.32%로서 예년의 비율에 비해 상당히 개선되었음을 알 수 있다. 다만 2인 이상의 담화에서 올해 공적 자료 부분은 9.02%로 그 비율이 저조한 편인데, 이는 향후 구어 원시 말뭉치 개발에서 이러한 장르가 집중적으로 구축된 후 주석이 이루어질 수 있을 것이라 생각된다.

<2004년 구어 형태 분석 말뭉치 목록>

파일명_주석	파일명_원시	제목	전체 어절수	저작권	비율
7CT_0001	6CM00001.HWP	강의_가정교과	4,124	○	
7CT_0002	6CM00003.HWP	강의_국어정보학	4,314	○	
7CT_0003	6CM00009.HWP	강의_저작권	7,528	○	
7CT_0004	6CM00010.HWP	강의_한국미술사	6,688	○	
1인담화/공적		강의	22,654		12.37%
7CT_0005	6CM00044.HWP	발표_학습과기억	4,013	○	

제1장 현대국어 구어 전사 말뭉치 개발

7CT_0006	6CM00045.HWP	발표_한시	2,206	○	
1인담화/공적		발표	6,219		3.39%
7CT_0007	6CM00020.HWP	독백_교환학생	4,485	○	
7CT_0008	6CM00022.HWP	독백_대학교	6,077	○	
7CT_0009	6CM00024.HWP	독백_사관학교	4,163	○	
7CT_0010	6CM00025.HWP	독백_살면서중요한것	2,413	○	
7CT_0011	6CM00029.HWP	독백_연애	1,498	○	
7CT_0012	6CM00030.HWP	독백_영화이야기	836	○	
7CT_0013	6CM00034.HWP	독백_일본여행기	3,100	○	
7CT_0014	6CM00036.HWP	독백_작사랑	1,072	○	
7CT_0015	6CM00037.HWP	독백_작사랑,전공,군대	3,966	○	
7CT_0016	6CM00038.HWP	독백_취업#1	3,791	○	
7CT_0017	6CM00040.HWP	독백_친구	3,122	○	
7CT_0018	6CM00042.HWP	독백_학교생활과여행기	6,351	○	
1인담화/사적		독백	40,874		22.32%
7CT_0019	6CM00017.HWP	구매대화_화장품	1,598	○	
2인이상담화/공적		구매대화	1,598		0.87%
7CT_0020	6CM00046.HWP	수업대화_과외지도	4,353	○	
7CT_0021	6CM00047.HWP	수업대화_한국어와컴퓨터	6,857	○	
2인이상담화/공적		수업대화	11,210		6.12%
7CT_0022	6CM00107.HWP	토론_언어와사회토론	3,768	○	
2인이상담화/공적		토론	3,768		2.06%
7CT_0023	6CM00048.HWP	일상대화_개인담#1	8,105	○	
7CT_0024	6CM00051.HWP	일상대화_대학생놀이문화	8,839	○	
7CT_0025	6CM00054.HWP	일상대화_미팅	5,652	○	
7CT_0026	6CM00056.HWP	일상대화_수강과목	1,624	○	
7CT_0027	6CM00057.HWP	일상대화_식사	2,579	○	
7CT_0028	6CM00058.HWP	일상대화_식사잡담	1,056	○	
7CT_0029	6CM00062.HWP	일상대화_인터넷사이트	3,019	○	
7CT_0030	6CM00064.HWP	일상대화_정치와경제	9,804	○	
7CT_0031	6CM00067.HWP	일상대화_질병과건강	2,878	○	
7CT_0032	6CM00071.HWP	일상대화_칠레	2,509	○	
7CT_0033	6CM00074.HWP	일상대화_휴식시간	2,429	○	
2인이상담화/사적		일상대화	48,494		26.48%
7CT_0034	6CM00075.HWP	전화대화_#1	4,318	○	
7CT_0035	6CM00076.HWP	전화대화_#2	4,868	○	
7CT_0036	6CM00077.HWP	전화대화_#3	1,011	○	
2인이상담화/사적		전화대화	10,197		5.57%
7CT_0037	6CM00079.HWP	주제대화_교육	2,761	○	
7CT_0038	6CM00082.HWP	주제대화_대통령선거	1,835	○	
7CT_0039	6CM00083.HWP	주제대화_대학생동아리문화	6,148	○	
7CT_0040	6CM00088.HWP	주제대화_병역	3,277	○	
7CT_0041	6CM00090.HWP	주제대화_연예가	3,283	○	

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

7CT_0042	6CM00092.HWP	주제대화_영화#2	4,345	○	
7CT_0043	6CM00098.HWP	주제대화_이야기만들기	4,967	○	
7CT_0044	6CM00099.HWP	주제대화_일상#1	3,478	○	
7CT_0045	6CM00104.HWP	주제대화_학생운동	8,048	○	
2인이상담화/사적		주제대화	38,142		20.82%
합계			183,156		100%

3. 현대 국어 구어 형태소 분석 말뭉치 구축 실례

```

<!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>강의_가정교과, 전자전사자료</title>
        <author>김인경</author>
        <sponsor>대한민국 문화관광부</sponsor>
        <respStmt><resp>전자/표준화, 형태소 정보 부착</resp>
          <name>연세대학교 언어정보연구원</name>
        </respStmt>
      </titleStmt>
      <extent>4124 어절</extent>
      <publicationStmt>
        <distributor>국립국어원</distributor>
        <idno>7CT_0001.hwp, 원시말뭉치:6CM00001.hwp</idno>
        <availability><p>배포 불가</p></availability>
      </publicationStmt>

```

```

<sourceDesc>
  <p>원전 있음, 강의를 녹음하여 전사함</p>
</sourceDesc>
</fileDesc>
<encodingDesc>
  <projectDesc><p>21세기 세종계획 3단계 1차년도 형태소 분석 말뭉치 구축</p></projectDesc>
  <samplingDecl><p>21세기 세종계획 원시 말뭉치 파일을 형태소 정보 부착 파일로 변환</p></samplingDecl>
  <editorialDecl><p>21세기 세종계획 구어 형태소 분석 말뭉치 구축 지침에 따른 형태소 정보 부착</p></editorialDecl>
</encodingDesc>
<profileDesc>
  <creation><date>2004</date></creation>
  <langUsage>
    <language id=K0 usage=99>한국어, 표준어</language>
  </langUsage>
  <particDesc>
    <person id=P1 sex=F age=40s><p>김인경, 교사, 서울</p></person>
    <person id=P2 sex=? age=10s><p>학생들</p></person>
    <person id=P3 sex=? age=10s><p>학생1</p></person>
    <person id=P4 sex=? age=10s><p>학생2</p></person>
    <person id=P5 sex=? age=10s><p>이주영</p></person>
    <person id=P6 sex=? age=10s><p>학생3</p></person>
  </particDesc>
  <settingDesc>
    <p>중학교 가정 수업 녹음임</p>
  </settingDesc>
  <textClass>
    <catRef scheme='SJ21' target='M2801'>구어 녹화 전사</catRef>
  </textClass>
</profileDesc>

```


21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

```
<revisionDesc>
  <change><date>2004</date>
    <respStmt>
      <resp>프로젝트책임자</resp><name>김향희</name>
    </respStmt>
  </change>
  <change><date>2004-8</date>
    <respStmt><resp>보조원</resp><name>정은숙</name></respStmt>
    <item>21세기 세종계획 구어 형태소 분석 말뭉치 구축 지침에 따른
형태소 정보 부착</item>
  </change>
  <change><date>2003-9</date>
    <respStmt><resp>연구보조원</resp><name>이지혜</name></respStmt>
    <item>형태소 정보 수정, 헤더 붙임, 마킹</item>
  </change>
</revisionDesc>
</teiHeader>
<text>
7CT_0001000001 <u who=P1>
7CT_0001000002 <s n=00001>
7CT_0001000003 자~ 자/IC DM
7CT_0001000004 삼십일 삼십/NR+일/NR
7CT_0001000005 페이지부터 페이지/NNG+부터/JKB
7CT_0001000006 옷을 옷/NNG+을/JK0
7CT_0001000007 어떻게 어떻게/MAG
7CT_0001000008 사야 사/VV+ㅏ야/EC
7CT_0001000009 할까. 하/VX+ㄴ까/EF+. /SF
7CT_0001000010 </s>
7CT_0001000011 <s n=00002>
7CT_0001000012 옷의 옷/NNG+의/JKG
7CT_0001000013 구매 구매/NNG
7CT_0001000014 계획과? 계획/NNG+과/JKB+?/SF
```

7CT_0001000015 </s>
 7CT_0001000016 <s n=00003>
 7CT_0001000017 아~ 아/IC DM
 7CT_0001000018 구매 구매/NNG
 7CT_0001000019 계획을 계획/NNG+을/JKO
 7CT_0001000020 할려면, 하/VV+ㄹ려면/EC+, /SP
 7CT_0001000021 </s>
 7CT_0001000022 <s n=00004>
 7CT_0001000023 먼저 먼저/MAG
 7CT_0001000024 나에게 나/NP+에게/JKB
 7CT_0001000025 필요한 필요/NNG+하/XSA+ㄴ/ETM
 7CT_0001000026 옷이? 옷/NNG+이/JKS+?/SF
 7CT_0001000027 </s>
 7CT_0001000028 <s n=00005>
 7CT_0001000029 어~ 어/IC DM
 7CT_0001000030 어떠한 어떠한/VA+ㄴ/ETM
 7CT_0001000031 것인가를 것/NNB+이/VCP+ㄴ가/EF+를/JKO
 7CT_0001000032 갖다가 갖/VV+다가/EC
 7CT_0001000033 계획을 계획/NNG+을/JKO
 7CT_0001000034 해야 하/VV+ㄷ야/EC
 7CT_0001000035 돼요. 되/VV+ㄷ요/EF+. /SF
 7CT_0001000036 </s>
 7CT_0001000037 <s n=00006>
 7CT_0001000038 그런데, 그런데/MAJ+, /SP
 7CT_0001000039 </s>
 7CT_0001000040 <s n=00007>
 7CT_0001000041 그 그/MM
 7CT_0001000042 옷을 옷/NNG+을/JKO
 7CT_0001000043 갖다 갖/VV+다/EC
 7CT_0001000044 사기 사/VV+기/ETN
 7CT_0001000045 전에 전/NNG+에/JKB
 7CT_0001000046 해야 하/VV+ㄷ야/EC

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

7CT_0001000047 될 되/VV+ㄹ/ETM
 7CT_0001000048 권 꺼/NNB+ㄴ/JX
 7CT_0001000049 뭐냐면, 뭐/NP+(이)/VCP+냐면/EC+, /SP
 7CT_0001000050 </s>
 7CT_0001000051 <s n=00008>
 7CT_0001000052 과연 과연/MAG
 7CT_0001000053 내게 내/NP+게/JKB
 7CT_0001000054 어떠어떠한 어떠/XR+어떠/XR+하/XSA+ㄴ/ETM
 7CT_0001000055 옷들이 옷/NNG+들/XSN+이/JKS
 7CT_0001000056 있는가 있/VA+는가/EF
 7CT_0001000057 한번 한번/MAG
 7CT_0001000058 봐야 보/VV+ㅏ야/EC
 7CT_0001000059 되겠조, 되/VV+겠/EP+조/EF+, /SP
 7CT_0001000060 </s>
 7CT_0001000061 <s n=00009>
 7CT_0001000062 내가 내/NP+가/JKS
 7CT_0001000063 봄이 봄/NNG+이/JKS
 7CT_0001000064 됐다. 되/VV+ㅓㅑ/EP+다/EF+, /SF
 7CT_0001000065 </s>
 7CT_0001000066 <s n=00010>
 7CT_0001000067 그럼 그럼/MAJ
 7CT_0001000068 봄이 봄/NNG+이/JKS
 7CT_0001000069 됐으면, 되/VV+ㅓㅑ/EP+으면/EC+, /SP
 7CT_0001000070 </s>
 7CT_0001000071 <s n=00011>
 7CT_0001000072 그냥 그냥/MAG
 7CT_0001000073 나가서 나가/VV+ㅓ서/EC
 7CT_0001000074 충동구매 충동/NNG+구매/NNG
 7CT_0001000075 모~ 모/IC DM
 7CT_0001000076 내가 내/NP+가/JKS
 7CT_0001000077 원하는 원하/VV+는/ETM
 7CT_0001000078 것 것/NNB

7CT_0001000079 아무거나 아무/MM+거/NNB+나/JX
 7CT_0001000080 사는 사/VV+는/ETM
 7CT_0001000081 게 거/NNB+ㅣ/JKC
 7CT_0001000082 아니라? 아니/VCN+라/EC+?/SF
 7CT_0001000083 </s>
 7CT_0001000084 <s n=00012>
 7CT_0001000085 과연 과연/MAG
 7CT_0001000086 나에게, 나/NP+에게/JKB+, /SP
 7CT_0001000087 </s>
 7CT_0001000088 <s n=00013>
 7CT_0001000089 에~ 에/IC DM
 7CT_0001000090 어떠한 어떠한/VA+ㄴ/ETM
 7CT_0001000091 스커트나 스커트/NNG+나/JC
 7CT_0001000092 어떠한 어떠한/VA+ㄴ/ETM
 7CT_0001000093 바지가 바지/NNG+가/JKS
 7CT_0001000094 지금 지금/MAG
 7CT_0001000095 현재 현재/NNG
 7CT_0001000096 있는가::, 있/VA+는가/EF+, /SP
 7CT_0001000097 </s>
 7CT_0001000098 <s n=00014>
 7CT_0001000099 그러구? 그러구/MAJ+?/SF
 7CT_0001000100 </s>
 7CT_0001000101 <s n=00015>
 7CT_0001000102 무엇이 무엇/NP+이/JKS
 7CT_0001000103 가장 가장/MAG
 7CT_0001000104 필요한가::, 필요/NNG+하/XSV+ㄴ가/EF+, /SP
 7CT_0001000105 </s>

(중략)

...

여 백

제 2 장

한영 병렬 말뭉치 개발

연구책임자 : 이 석 재

여 백

제 2 장

한영 병렬 말뭉치 개발

1. 연구의 필요성 및 목표

1.1 연구의 필요성

정보화와 국제화의 시대에 더욱 빠르고 효율이 높은 정보 교환을 위해서는 여러 언어로 된 정보를 빠르게 처리하는 다국어 언어 처리 시스템의 개발이 필수적이다. 이러한 다국어 언어처리 시스템에서 가장 기본이 되는 것은 바로 병렬말뭉치(parallel corpus)인데, 활용성, 응용성을 높이고 말뭉치 자료를 보다 정제화하기 위해서는 형태론적 정보를 부착한 주석 말뭉치의 개발이 필수적이다.

형태소 분석 병렬 말뭉치의 형태 주석 정보는 언어 처리 시스템 개발뿐 아니라, 한영 대조 연구 등의 언어학 분야, 이중 언어 사전 편찬 등의 응용 분야, 한국어 교육 및 영어 교육 등의 교육 분야 등의 분야에서도 응용될 수 있을 것이다.

1.2 연구 목표

본 과제의 2004년도 사업에서는 형태론적 주석이 달린 19만 규모의 형태소 분석 한영 병렬 말뭉치 개발을 목표로 한다. 2001년도에 작성한 세부 지침 및 2002년도에 작성한 수정 지침을 토대로 하여 주석해 나가되, 형태소 분석 과정에서 기존

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

의 형태소 분석 세부 지침을 수정, 보완하는 작업을 병행한다.

2. 한영 병렬 형태소 분석 말뭉치 구축

2.1 말뭉치 목록

구분	주석 파일명	원시 파일명	제목	언어	장르	생산 연도	어절수	저작권
영한	7RT_01ek	rh0106ek	생각연습	한글	책-인문	1994	34021	O
영한	7RT_01ee	rh0106ee	Practical Thinking	영어	책-인문	1971	50018	X
영한	7RT_03ek	rh0319ek	토익 점수 마구 올려주는 토익 토마토 2	한글	책-교육	2002	2663	O
영한	7RT_03ee	rh0319ee	English Care TOEIC Series 2 Reading	영어	책-교육	2002	3319	O
한영	7RT_02kk	rh0317kk	21세기의 한글	한글	책-인문	1996	11586	O
한영	7RT_02ke	rh0317ke	Han-gul in the 21st Century	영어	책-인문	1996	17400	O
한영	7RT_04kk	rh0320kk	영어 표현 top 10	한글	책-교육	2003	17874	O
한영	7RT_04ke	rh0320ke	TOP 10 English Expressions	영어	책-교육	2003	24208	O
한영	7RT_05kk	rh0321kk	영어 회화 60일 초보탈출기_ek	한글	책-교육	2003	9948	O
한영	7RT_05ke	rh0321ke	영어 회화 60일 초보탈출기_ee	영어	책-교육	2003	14196	O
한영	7RT_06kk	rj0323kk	2002년 외교통상부 장관 신년사(2002.01.02)	한글	연설문	2002	774	O
한영	7RT_06ke	rj0323ke	New Year Message by Minister of Foreign Affairs and Trade Han Seung-soo	영어	연설문	2002	1300	O
한영	7RT_07kk	rj0329kk	제4차 ASEM 정상회의 개최식 대통령 연설문(2002.09.23)	한글	연설문	2002	443	O
한영	7RT_07ke	rj0329ke	Remarks by President Kim Dae-jung of the Republic of Korea at the Opening Ceremony of ASEM IV	영어	연설문	2002	827	O
한영	7RT_08kk	rj0332kk	[노벨상 제정 100주년 기념전] 개막식 대통령 축사(2002.08.23)	한글	연설문	2002	557	O
한영	7RT_08ke	rj0332ke	Congratulatory Remarks by President Kim at the Opening Ceremony in Seoul of The Centennial Exhibition of the Nobel Prize	영어	연설문	2002	990	O
한영 병렬 형태소 주석 말뭉치 총합							190124	

<기호 설명> O:저작권 동의서 받음 X:저작권 동의서 못 받음

올해 형태소 분석 말뭉치 구축을 위해 선정된 텍스트는 비상상 텍스트에 치중되어 있는데, 이는 기존에 구축된 원시 말뭉치 중에서 저작권 동의를 받은 텍스트를 중심으로 구성하였기 때문이다. 올해 한영 병렬 형태소 분석 말뭉치의 저작권 해결 비율은 73.7%(한국어 100%, 영어 55%)이다.

2.2 말뭉치 구성

	(총 190,124 어절)
언어별	한글 텍스트 77,866 어절 (41.0%) 영어 텍스트 112,258 어절 (59.0%)
원본별	영한 대역 : 100,103 어절 (52.6%) 한영 대역 : 90,021 어절 (47.4%)
저작권 해결 비율	한글 텍스트 100% 영어 텍스트 55%

2.3. 말뭉치 구축 과정

1) 원시 말뭉치 가공 작업

원시 말뭉치에 부착되어 있는 CES 방식의 헤더 정보를 제거하고, 말뭉치 자료의 오류를 검토한다.

2) 형태소 분석 작업

주석 편집 도구 'SORITAG061.exe'를 이용하여 원시 말뭉치를 일괄 분석한다. 일괄 분석은 2003년까지 구축된 한영 병렬 형태소 분석 말뭉치 52만 어절로 생성한 어절분석사전을 기반으로 한 반자동 분석이다.

3) 분석 표지 부착 말뭉치 가공 작업

어절 분석 표지가 부착된 말뭉치에는 병렬 문서의 정렬 정보가 함께 들어 있으므로, 각 문단, 문장의 id 정보가 손실되지 않도록 주의한다.

4) 태깅 결과 후처리 및 분석 규칙 작성을 위한 자료 수집

'SORITAG061'를 이용한 형태소 분석 결과물은 텍스트 에디터에서 3차에 걸친 수작업 과정을 통하여 후처리한다. 이 과정에서 오분석을 수정하여 올바른 주석을 부착하는데, 작업자간의 주석 회의를 통해 분석 결과의 일관성을 높이고 필요에 따라 기존의 지침을 수정, 보완한다. 또한 주석의 정확성을 높이기 위해서 자주 나타나는 오분석의 유형을 정리하고, 이는 반드시 작업자간에 교차 검수한다.

5) 형태소 분석 한영 병렬 말뭉치의 문서 표준안을 따르는 형태소 분석 말뭉치의 생성

형태소 분석 한영 병렬 말뭉치는 XML 문서 표준안을 따른다. 따라서 2002년도 최종 보고서에서 제시한 XML방식의 마크업 체계를 따라 한영 병렬 형태소 분석 말뭉치를 생성한다.

6) 구축된 말뭉치의 적합성 및 유효성 검사

형태소 분석 한영 병렬 말뭉치가 XML 형식의 마크업 체계를 준수하고 있는지 점검하기 위해서 2002년도에 개발한 validator를 이용하여 적합성 및 유효성 검사를 실시한다.

2.4. 말뭉치 구축 도구

'한영 병렬 형태소 분석 말뭉치 구축'은 2003년과 마찬가지로 언어정보연구원 내에서 자체 개발한 'SORITAG061.exe'를 이용한다. 이는 '기분석 형태 분석 말뭉치'를 이용한 주석 편집 도구로서, 한영 병렬 형태소 분석 말뭉치의 경우 2003년까지 구축했던 말뭉치 52만 어절의 주석 말뭉치를 데이터로 이용해서 올해 형태소 분석

말뭉치를 구축한다.

소리태그는 같은 지침에 따라 분석된 말뭉치를 바탕으로 하여 분석하기 때문에 비교적 일관된 결과물을 얻을 수 있다. 단, 부호, 기호가 포함된 문자열에서 단순한 오류가 많아서 이를 수정하는 데에 적잖은 노력이 필요하였다. 이 점이 수정된다면 분석기에서 좀더 정제된 결과물을 얻을 수 있을 것이다.

다음은 기존의 주석 말뭉치를 사전 데이터로 불러온 후, 분석할 원시 말뭉치를 불러와서 형태 주석을 하는 과정이다.

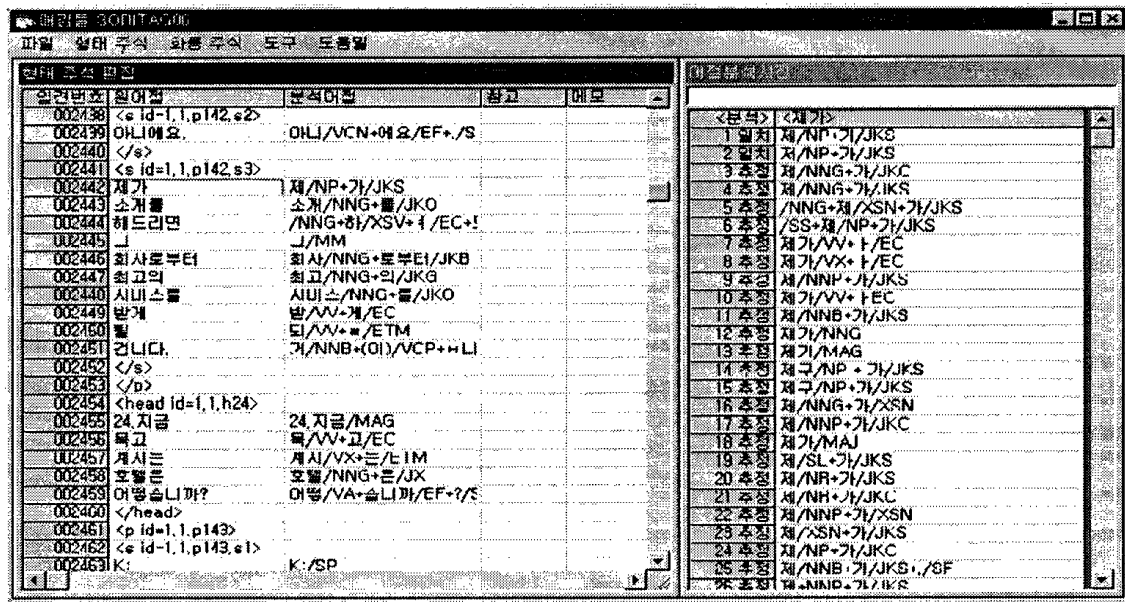


그림 1. <SORITAG061.exe_어절 추정 분석 결과 보기>

3. 형태소 분석 지침 보완 및 논의 사항

3.1 말뭉치 구축 지침

한영 병렬 한국어 문서 형태소 분석 말뭉치는, <21세기 세종계획>의 국어 기초 자료 구축 분과에서 제안된 ‘세종 분석 표지’(2000 국어 기초자료 구축 연구보고서)

를 분석 표지로 하고, 지침은 2002년도에 ‘한영 병렬 한국어 문서 형태소 분석 말뭉치 구축 지침(2001-1-한 ver 1.0)’과 이를 수정, 보완한 ‘한영 병렬 한국어 문서 형태소 분석 말뭉치 구축 지침(2001-1-한 ver 2.0)’을 이용한다. 한영 병렬 영어 문서 형태소 분석 말뭉치는, ‘The Penn Treebank POS tagset’를 분석 표지로 하고, 지침은 ‘Part-of-Speech Tagging Guidelines for the Penn Treebank Project(3rd Revision, 2nd printing): Beatrice Santorini, June 1990’을 기초로 하여 작성된 ‘한영 병렬 영어 문서 형태소 분석 말뭉치 구축 지침(2001-1-영 ver 2.0)’이용한다. 한편, 주석 과정 중에 작업자의 논의를 통하여 필요에 따라 지침을 수정, 보완하여 ‘ver 3.0’을 작성하였다.

3.2 말뭉치 구축 지침의 수정 및 보완

올해 주석 작업 과정에서 추가된 한국어 형태소 분석 지침은 다음과 같다.

- 복원하지 않는 단어 추가: 적잖다 (4.2.4.줄어든 형태의 표기)

- 복원하지 않는 어절 분석 추가 (4.3.3.줄어든 형태의 표기)

예) 하려는(‘하려고 하는’의 줄어든 꼴) 하/VV+려는/ETM

- 형용사 파생 접미사 ‘답다, 스럽다, 립다’는 분석하지 않지만, 임의적 쓰임의 ‘답다’는 분석한다. (5.15.접미사)

예) 대학원생/NNG+답/XSA+다/EF

아름답/VA+다/EF

참답/VA+다/EF

- 홀로는 쓰이지 못하고 새로운 단어나 구를 만드는 데 쓰이는 단어는 일반명사로 분석한다. (어근으로 분석하지 않는다.) (6.1.1.명사)

예) 주요 논의 주요/NNG 논의/NNG

주요한

주요/NNG+하/XSV+ㄴ/ETM

예) 강력, 건전, 관련, 단순, 다양, 당면, 당해, 유력, 유용, 이상, 정확, 주요, 중요, 충분, 특이, 특별, 편리

- 일반명사가 부사적 쓰임을 가지면 부사로 분석한다.-목록 추가(6.1.1.명사)

예) 그대로, 당시, 먼저, 보통, 사실, 사실상, 여태, 오랫동안, 이제, 잠시, 잠깐, 지금, 최대한, 최소한, 한데, 한동안, 한참, 혼자

- ‘간’의 경우, 의존명사와 접미사로 모두 쓰일 수 있다. (6.1.1.명사)

예) 종교간 종교/NNG+간/XSN
한국과 일본 간 한국/NNP+과/JC 일본/NNP 간/NNB

- '동안'이 포함된 단어는 다음과 같이 분석한다. (6.1.1.명사)

예) 오랫동안/MAG 기다렸다	오랫동안/NNG+의/JKG
한동안/MAG 또는 NNG	
한참/NNG+동안/NNB	잠시/NNG+동안/NNB

- ‘번째’는 분리하지 않고 하나의 의존명사로 분석한다. (6.1.1.명사)

예) 네 번째 네/MM 번째/NNB
 네 번째 네/MM+번째/NNB

- ‘내, 네(니), 제’의 경우에는 대명사와 관형사를 구분하여 분석하고, 나머지 대명사는 관형적 쓰임을 가질 경우에도 대명사로 분석한다. (6.1.2.대명사)

예) 우리/NP 집
 누이/NP 집

- ‘이/그/저+아이’의 준꼴인 ‘애, 개, 재’는 일반명사로 분석한다. (6.1.2.대명사)

예) 잔 재/NNG+ㄴ/JX

- 한 단어처럼 쓰인 '대명사+ㄴ가/은가'의 경우, '누군가'는 대명사로, '언젠가'

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

는 부사로 분석하고, 그 외 어절은 다음과 같이 분석한다. (대명사적 쓰임이 아니면 모두 분석한다.) (6.1.2.대명사)

- 예) 무엇인가가 나왔다. 무엇/NP+이/VCP+ㄴ가/EC+가/JKS
 어디엔가에 있고 어디/NP+에/JKB+(이)/VCP+ㄴ가/EC+에/JKB
예) 누군가 찾아왔다. 누군가/NP
 그 사람이 누군가에 따라 누구/NP+(이)/VCP+ㄴ가/EC+에/JKB

- 부사에 '하다, 되다'가 결합하여 동사, 형용사가 파생된 경우에는 '부사+접미사'로 분석한다. (6.3.2.부사)

예) 가득하여 가득/MAG+하/XSV+여/EC

예) 가득하다, 거둡하다, 그릇되다, 그만하다, 마주하다

- 접속사 추가: 아울러, 하기야 (6.3.2.부사)

- 홈페이지 주소나 이메일 주소는 전체를 하나의 외국어로 분석한다. (6.7.1.기호)

예) paralleltagger@yonsei.ac.kr/SL
 [www://clid.yonsei.ac.kr](http://clid.yonsei.ac.kr)/SL

다음은 올해 주석 작업 과정에서 추가된 영어 형태소 분석 지침이다.

- 사람을 부를 때 쓰이는 'Miss', 'Sir'은 NN으로 분석한다.

예) Miss/NN, are we on schedule for 4:00?

- 간투사처럼 쓰이는 so는 접속사로 태깅하지 않고 UH로 분석한다.

예) A: So/UH, what will we do, visit your parents or mine, this weekend?

예) Sure, Great

- at last의 last는 명사로 분석한다.

예) at/IN last/NN

예) last/JJ week/NN

- come on의 줄어든 꼴인 co'mon은 동사로 분석한다.

예) Hey, Co'mon/VB.

- furthermore는 RB로 태깅한다.

예) Furthermore/RB, he varied the place of circles.

- even though는 다음과 같이 분석한다.

예) even/RB though/IN

- on the contrary는 다음과 같이 분석한다.

예) on/IN the/DT contrary/NN

3.3 말뭉치 지도

2001년부터 올해까지 4년 동안 구축된 총 711,227어절의 한영 병렬 형태소 분석 말뭉치의 장르별, 언어별, 원본별 구성 내용을 살펴보면 다음과 같다.

원본별	2단계(2001-2003)		3단계(2004)		언어별 합계		합계	비율
언어	영어	한국어	영어	한국어	영어	한국어		
영한 대역	145,489	132,798	53,337	36,684	198,826	169,482	368,308	51.8%
한영 대역	127,865	114,951	58,921	41,182	186,786	156,133	342,919	48.2%
합계	273,354	247,749	112,258	77,866	385,612	325,615	711,227	100.0%
비율					54.2%	45.8%		

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

매체별	내용별	2단계(2001-2003)				3단계(2004)				장르별 어절수	비율
		원텍스트 (SI)		번역텍스트 (TI)		원텍스트 (SI)		번역텍스트 (TI)			
		영어	한국어	영어	한국어	영어	한국어	영어	한국어		
잡지	총류	16,737			13,542					30,279	4.3%
책	상상적 텍스트	21,805			15,409					37,214	5.2%
	인문					50,018	11,586	17,400	34,021	113,025	16.0%
	사회	25,081			19,967					45,048	6.3%
	생활	35,169			22,815					57,984	8.2%
	교육 자료	44,496	80,408	71,329	59,762	3,319	27,822	38,404	2,663	328,203	46.2%
정부 문서	사회	2,201	34,543	56,536	1,303		1,774	3,117		99,474	14.0%
	언어별 어절수	145,489	114,951	127,865	132,798	53,337	41,182	58,921	36,684	711,227	100.0%
	합계	521,103				190,124				711,227	

지금까지 구축된 한영 병렬 형태소 분석 말뭉치는 영한 대역 텍스트와 한영 대역 텍스트, 한국어 텍스트와 영어 텍스트를 비교적 고루 갖추고 있다. 텍스트의 내용 장르별로 살펴보면 인문, 사회, 상상적 텍스트, 생활, 교육 자료 등 다양한 분야를 포함하고 있다. 이 중 영어 교과서를 중심으로 한 교육 자료의 비율이 높은데, 교육 자료는 양질의 번역 텍스트를 가지고 있기 때문이다. 또한 기구축된 원시 말뭉치 중에서 형태소 분석 말뭉치로 가공할 텍스트를 선정할 때 저작권 해결 문제를 우선적으로 고려하지 않을 수 없으므로, 결과적으로 기술적 텍스트, 제품 매뉴얼, 실용적 텍스트의 비율이 낮아졌다.

지금까지의 한·영 병렬 말뭉치 구성 비율과 자연언어처리 및 언어 공학 분야로의 응용 가능성을 고려하여 앞으로 다양성과 균형성을 갖출 수 있도록 다양한 텍스트의 비율을 높여 나가야 할 것이며, 이를 위해서는 저작권 해결된 원시 말뭉치의 구축이 선행되어 한다. 올해에는 더욱 다양한 장르의 텍스트로 구성된 원시 말뭉치가 구축되었고 꾸준히 저작권 해결에 노력하고 있으므로 후년의 형태소 분석 말뭉치는 더 다양한 장르별로 구성할 수 있을 것이다.

4. 한영 병렬 형태소 분석 말뭉치의 예

- 한국어문서: 7RT_06kk.hwp

```

<?xml version="1.0" encoding="EUC-KR"?>
<!--<!DOCTYPE cesCLID SYSTEM "http://clid.yonsei.ac.kr/xml/dtd/cesCLID.dtd"-->
<ces>
  <validation>
    <sejongcorpus>parallel</sejongcorpus>
    <type>korean</type>
    <version>1.0</version>
  </validation>
  <cesHeader version="2.3">
    <fileDesc>
      <titleStmt>
        <h.title>2002년 외교통상부 장관 신년사, 형태소 분석 전자파일
      </h.title>
      <respStmt>
        <respType>헤더붙임, 형태 주석</respType>
        <respName>연세대학교 언어정보연구원</respName>
      </respStmt>
    </titleStmt>
    <extent>
      <wordcount>774</wordcount>
    </extent>
    <publicationStmt>
      <distributor>국립국어원</distributor>
      <pubAddress>서울시 강서구 방화3동 827번지</pubAddress>
      <availability status="restricted">배포 불가</availability>
      <idno>7RT_06kk.hwp, 원시말뭉치 : rj0323kk.hwp</idno>
      <PubDate>2004</PubDate>
    </publicationStmt>
    <sourceDesc>

```

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

```
<biblStruct>
  <monogr>
    <h.title>2002년 외교통상부 장관 신년사</h.title>
    <h.author></h.author>
    <imprint>
      <pubPlace></pubPlace>
      <publisher></publisher>
      <pubDate></pubDate>
    </imprint>
  </monogr>
</biblStruct>
</sourceDesc>
</fileDesc>
<encodingDesc>
  <projectDesc>21세기 세종계획 3단계 1차년도 한영 병렬 말뭉치 구축
</projectDesc>
  <editorialDecl>21세기 세종계획 말뭉치 문헌 입력 지침 및 한영 병렬 한국어
형태소 주석 지침에 따름</editorialDecl>
  <samplingDecl></samplingDecl>
  <refsDecl>본문, 구획, 문단, 문장 단위로 일련번호를 붙여 병렬텍스트와 정렬
시킨 후 형태 주석 정보를 어절의 하위 형태 단위에 태깅
</refsDecl>
</encodingDesc>
<profileDesc>
  <creation><date>2004/08/01</date></creation>
  <langUsage>
    <language id="kr" iso639="kr"></language>
  </langUsage>
  <wsdUsage>
    <writingSystem id="EUC-KR(KSC5601)"></writingSystem>
  </wsdUsage>
  <textClass>
    <catRef scheme='SJ21' target="R1564">정부문서: 사회, 외교</catRef>
  </textClass>
```

```

    <translations>
      <translation          trans.loc="7RT_06ke.hwp"          lang="en"
wsd="iso8859-1"></translation>
      <translator></translator>
    </translations>
  </profileDesc>
  <revisionDesc>
    <change>
      <changeDate>2004/11/01</changeDate>
      <respName>봉미경, 박유현</respName>
      <respName>이석재</respName>
      <h.item>본문 마크업, 형태 주석</h.item>
    </change>
  </revisionDesc>
</cesHeader>
<chunkList>
<chunk>
<div id="1.1">
<head id="1.1.h1">
<tok>
<orth>2002</orth>
<seg><mor>2002</mor><pos>SN</pos></seg>
</tok>
<tok>
<orth>-</orth>
<seg><mor>-</mor><pos>SS</pos></seg>
</tok>
<tok>
<orth>01</orth>
<seg><mor>01</mor><pos>SN</pos></seg>
</tok>
<tok>
<orth>-</orth>
<seg><mor>-</mor><pos>SS</pos></seg>

```

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

</tok>
<tok>
<orth>02</orth>
<seg><mor>02</mor><pos>SN</pos></seg>
</tok>
</head>
<head id="1.1.h2">
<tok>
<orth>제목</orth>
<seg><mor>제 목</mor><pos>NNG</pos></seg>
</tok>
<tok>
<orth>2002년</orth>
<seg><mor>2002</mor><pos>SN</pos></seg>
<seg><mor>년</mor><pos>NNB</pos></seg>
</tok>
<tok>
<orth>외교통상부</orth>
<seg><mor>외 교통 상 부</mor><pos>NNP</pos></seg>
</tok>
<tok>
<orth>장관</orth>
<seg><mor>장 관</mor><pos>NNG</pos></seg>
</tok>
<tok>
<orth>신년사</orth>
<seg><mor>신 년 사</mor><pos>NNG</pos></seg>
</tok>
</head>
<head id="1.1.h3">
<tok>
<orth>연설자</orth>
<seg><mor>연 설 자</mor><pos>NNG</pos></seg>
</tok>

```

<tok>
<orth>한승수</orth>
<seg><mor>한승수</mor><pos>NNP</pos></seg>
</tok>
</head>
<p id="1.1.pl">
<s id="1.1.pl.s1">
<tok>
<orth>국내외에서</orth>
<seg><mor>국내외</mor><pos>NNG</pos></seg>
<seg><mor>에서</mor><pos>JKB</pos></seg>
</tok>
<tok>
<orth>국가이익을</orth>
<seg><mor>국가</mor><pos>NNG</pos></seg>
<seg><mor>이익</mor><pos>NNG</pos></seg>
<seg><mor>을</mor><pos>JKO</pos></seg>
</tok>
<tok>
<orth>수호하기</orth>
<seg><mor>수호</mor><pos>NNG</pos></seg>
<seg><mor>하</mor><pos>XSV</pos></seg>
<seg><mor>기</mor><pos>ETN</pos></seg>
</tok>
<tok>
<orth>위하여</orth>
<seg><mor>위하</mor><pos>VV</pos></seg>
<seg><mor>여</mor><pos>EC</pos></seg>
</tok>
.
.
.
<tok>
<orth>1년</orth>
<seg><mor>1</mor><pos>SN</pos></seg>

```

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

<seg><mor>년</mor><pos>NNB</pos></seg>
</tok>
<tok>
<orth>내내</orth>
<seg><mor>내내</mor><pos>MAG</pos></seg>
</tok>
<tok>
<orth>함께</orth>
<seg><mor>함께</mor><pos>MAG</pos></seg>
</tok>
<tok>
<orth>하기를</orth>
<seg><mor>하</mor><pos>VV</pos></seg>
<seg><mor>기</mor><pos>ETN</pos></seg>
<seg><mor>를</mor><pos>JKO</pos></seg>
</tok>
<tok>
<orth>기원합니다.</orth>
<seg><mor>기원</mor><pos>NNG</pos></seg>
<seg><mor>하</mor><pos>XSV</pos></seg>
<seg><mor>히니다</mor><pos>EF</pos></seg>
<seg><mor>.</mor><pos>SF</pos></seg>
</tok>
</s>
<s id="1.1.p26.s2">
<tok>
<orth>감사합니다.</orth>
<seg><mor>감사</mor><pos>NNG</pos></seg>
<seg><mor>하</mor><pos>XSV</pos></seg>
<seg><mor>히니다</mor><pos>EF</pos></seg>
<seg><mor>.</mor><pos>SF</pos></seg>
</tok>
</s>
</p>

```

</div>
</chunk>
</chunkList>
</ces>

```

- 영어문서: 7RT_06ke.hwp

```

<?xml version="1.0" encoding="EUC-KR"?>
<!--<!DOCTYPE cesCLID SYSTEM "http://clid.yonsei.ac.kr/xml/dtd/cesCLID.dtd"-->
<ces>
  <validation>
    <sejongcorpus>parallel</sejongcorpus>
    <type>English</type>
    <version>1.0</version>
  </validation>
  <cesHeader version="2.3">
    <fileDesc>
      <titleStmt>
        <h,title>New Year Message by Minister of Foreign Affairs and Trade
Han Seung-soo, 형태소 분석 전자파일</h.title>
        <respStmt>
          <respType>헤더붙임, 형태 주석</respType>
          <respName>연세대학교 언어정보연구원</respName>
        </respStmt>
      </titleStmt>
      <extent>
        <wordcount>1300</wordcount>
      </extent>
      <publicationStmt>
        <distributor>국립국어원</distributor>
        <pubAddress>서울시 강서구 방화3동 827번지</pubAddress>
      </publicationStmt>
    </fileDesc>
  </cesHeader>

```


21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

```

    <availability status="restricted">배포 불가</availability>
    <idno>7RT_06ke.hwp, 원시말뭉치 : rj0323ke.hwp</idno>
    <PubDate>2004</PubDate>
  </publicationStmt>
  <sourceDesc>
    <biblStruct>
      <monogr>
        <h.title>New Year Message by Minister of Foreign Affairs
and Trade Han Seung-soo</h.title>
        <h.author></h.author>
        <imprint>
          <pubPlace></pubPlace>
          <publisher></publisher>
          <pubDate></pubDate>
        </imprint>
      </monogr>
    </biblStruct>
  </sourceDesc>
</fileDesc>
<encodingDesc>
  <projectDesc>21세기 세종계획 3단계 1차년도 한영 병렬 말뭉치 구축
</projectDesc>
  <editorialDecl>21세기 세종계획 말뭉치 문헌 입력 지침 및 한영 병렬 영어 형
태소 주석 지침에 따름</editorialDecl>
  <samplingDecl></samplingDecl>
  <refsDecl>본문, 구획, 문단, 문장 단위로 일련번호를 붙여 병렬텍스트와 정렬
시킨 후 형태 주석 정보를 어절의 하위 형태 단위에 태깅
  </refsDecl>
</encodingDesc>
<profileDesc>
  <creation><date>2004/08/01</date></creation>
  <langUsage>
    <language id="en" iso639="en"></language>
  </langUsage>

```

```

<wsdUsage>
  <writingSystem id="iso8859-1"></writingSystem>
</wsdUsage>
<textClass>
  <catRef scheme='SJ21' target="R1564">정부분서: 사회, 외교</catRef>
</textClass>
<translations>
  <translation          trans.loc="7RT_06kk.hwp"          lang="kr"
wsd="EUC-KR(KSC5601)"></translation>
  <translator></translator>
</translations>
</profileDesc>
<revisionDesc>
  <change>
    <changeDate>2004/11/01</changeDate>
    <respName>봉미경, 장윤자</respName>
    <respName>이석재</respName>
    <h.item>본문 마크업, POS 태깅</h.item>
  </change>
</revisionDesc>
</cesHeader>
<chunkList>
<chunk>
<div id="1.1">
<head id="1.1.h1">
<tok>
<orth>연설일자</orth>
<seg><mor>연설일자</mor><pos>FW</pos></seg>
</tok>
<tok>
<orth>2002-01-07</orth>
<seg><mor>2002-01-07</mor><pos>CD</pos></seg>
</tok>
</head>

```

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

<head id="1.1.h2">
<tok>
<orth>제목</orth>
<seg><mor>제목</mor><pos>FW</pos></seg>
</tok>
<tok>
<orth>New</orth>
<seg><mor>New</mor><pos>NNP</pos></seg>
</tok>
<tok>
<orth>Year</orth>
<seg><mor>Year</mor><pos>NNP</pos></seg>
</tok>
<tok>
<orth>Message</orth>
<seg><mor>Message</mor><pos>NN</pos></seg>
</tok>
<tok>
<orth>by</orth>
<seg><mor>by</mor><pos>IN</pos></seg>
</tok>
<tok>
<orth>Minister</orth>
<seg><mor>Minister</mor><pos>NNP</pos></seg>
</tok>
<tok>
<orth>of</orth>
<seg><mor>of</mor><pos>IN</pos></seg>
</tok>
<tok>
<orth>Foreign</orth>
<seg><mor>Foreign</mor><pos>NNP</pos></seg>
</tok>
<tok>

```

<orth>Affairs</orth>
<seg><mor>Affairs</mor><pos>NNPS</pos></seg>
</tok>
<tok>
<orth>and</orth>
<seg><mor>and</mor><pos>CC</pos></seg>
</tok>
<tok>
<orth>Trade</orth>
<seg><mor>Trade</mor><pos>NNP</pos></seg>
</tok>
<tok>
<orth>Han</orth>
<seg><mor>Han</mor><pos>NNP</pos></seg>
</tok>
<tok>
<orth>Seung-soo</orth>
<seg><mor>Seung-soo</mor><pos>NNP</pos></seg>
</tok>
</head>
      .
      .
      .

<tok>
<orth>the</orth>
<seg><mor>the</mor><pos>DT</pos></seg>
</tok>
<tok>
<orth>Year</orth>
<seg><mor>Year</mor><pos>NN</pos></seg>
</tok>
<tok>
<orth>2002</orth>
<seg><mor>2002</mor><pos>CD</pos></seg>

```

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

</tok>
<tok>
<orth>.</orth>
<seg><mor>.</mor><pos>.</pos></seg>
</tok>
</s>
</p>
<p id="1.1.p21">
<s id="1.1.p21.s1">
<tok>
<orth>Thank</orth>
<seg><mor>Thank</mor><pos>VBP</pos></seg>
</tok>
<tok>
<orth>you</orth>
<seg><mor>you</mor><pos>PRP</pos></seg>
</tok>
<tok>
<orth>.</orth>
<seg><mor>.</mor><pos>.</pos></seg>
</tok>
</s>
</p>
</div>
</chunk>
</chunkList>
</ces>

제 3 장

한일 병렬 말뭉치 개발

연구책임자 : 전 형 식

여 백

제 3 장

한일 병렬 말뭉치 개발

1. 연구 개요

이 연구는 한일 양국어의 문법 구조, 어휘 사용의 실태 등 각종 언어적 특징을 비교 연구하기 위한 기초적인 자료를 제공하며, 그 자료의 활용 방안에 대해 검토하는 것을 목적으로 한 것이다.

2003년도(2단계 3차 년도)부터 처음 시작된 한일 병렬 형태소 분석 말뭉치 구축 시에는 생각하지 못 했던 문제점이 많이 나타나 작업 진행에 큰 어려움을 겪었다. 2004년도 작업에서는 전년도 작업 시에 드러난 문제점을 개선하여, 한일 병렬 형태소 분석 말뭉치를 구축함에 있어서 유용한 정보를 제공하는 것도 주된 목표의 하나이다.

이와 같은 대규모 자료 구축 연구에는 다루는 자료의 양이 워낙 큰 만큼 통일된 지침 하에서 진행되어야 하며, 전체적인 문제점 파악이 안 되면 양질의 자료를 구축하는 것은 불가능하다.

따라서 본 보고서에서는 여러 장르의 다양한 문서에서 양질의 한일 병렬 형태소 분석 말뭉치를 구축하는데 있어서 숙지해야할 점을 매뉴얼화 하여 내년도 이후의 작업에 도움이 되는 정보를 남기도록 하겠다. 또한 모든 작업 과정을 통해 기계화·일괄화 연구를 진행하므로 보다 정확하고 효율적인 작업 방법을 찾는 데에도 힘을 기울일 것이다.

2. 한일 형태소 분석 말뭉치 구축

2.1. 형태소 분석 말뭉치 구축량

2004년도 한일 병렬 말뭉치 구축 분과의 형태소 분석 말뭉치의 구축 목표량은 12만 어절이며 자세한 말뭉치 구축 사항을 제시하면 다음과 같다.

- 총 121,335 어절 (100%)
- 한국어 텍스트 60,574 어절 (49.9%) / 일본어 텍스트 60,761 어절 (50.1%)
- 한국어 파일수 44 / 일본어 파일수 44
- 한국어 원문 파일수 58 / 일본어 원문 파일수 30

2.2. 형태소 분석 말뭉치 전체 목록

형태소 분석 말뭉치의 전체 목록은 다음과 같다. 2004년도 작업에서는 시나리오 자료를 제외한 모든 장르를 대상으로 형태소 분석 말뭉치 구축에 적합한 내용의 자료를 선정하였다. 선정 기준은 현대 한국어 및 일본어의 특징을 잘 나타내고 있는 것을 원칙으로 하여 구어체가 지나치게 나타나는 자료는 그 대상에서 제외하기로 하였다.

구어체가 포함된 자료를 형태소 분석 말뭉치로 사용하지 않은 이유는 우선 현대 한국어 및 일본어의 표준적·전형적인 문법·어휘 체계가 아니라 변형적인 요소가 강하다는 점, 그리고 구어체를 형태소 단위로 정확히 분석할 수 있는 기준이 아직 마련되지 않다는 점을 들 수 있다. 구어체의 자료까지 포함한 한일 병렬 형태소 분석 말뭉치를 구축하기 위해서는 구어체를 전문으로 하는 한일 병렬 구어 형태소 분석 말뭉치 구축 분과가 필요할 것으로 생각된다.

자료 내용의 대부분을 대화문(구어체)이 차지하고 있는 시나리오의 장르가 처음부터 선정 대상으로부터 제외된 이유는 그 때문이다.

제3장 한일 병렬 말뭉치 개발

원본 언어	형태소 파일명	원시 파일명	제목	텍스트 언어	장르	출판일	어절수	저작권
일본어	7JT_01jj	je0401jj	こころ-先生と私	일본어	소설	2002/12/06	14,859	-
일본어	7JT_01jk	je0401jk	마음-선생님과 나	한국어	소설	2002/12/06	16,270	○
일본어	7JT_02jj	je0402jj	こころ-両親と私	일본어	소설	2002/12/06	7,833	-
일본어	7JT_02jk	je0402jk	마음-양친과 나	한국어	소설	2002/12/06	8,574	○
일본어	7JT_03jj	jj0401jj	生成語彙論と対訳辞書	일본어	논문	2000/05	3,581	○
소설 종합							47,536	
일본어	7JT_03jk	jj0401jk	생성어휘론과 대역사 전	한국어	논문	2000/05	3,294	○
일본어	7JT_04jj	jj0402jj	韓国における専門用語 の歴史と現状	일본어	논문	2000/02	3,606	○
일본어	7JT_04jk	jj0402jk	한국에 있어서의 전 문용어 역사와 현상	한국어	논문	2000/02	3,603	○
일본어	7JT_05jj	jj0403jj	韓日対訳構文辞書にお ける文型の意味的特徴	일본어	논문	2002/11	3,383	○
일본어	7JT_05jk	jj0403jk	한일대역구문사전에 있어서 문형의 의미 적 특징	한국어	논문	2002/11	3,006	○
일본어	7JT_06jj	jj0404jj	共起関係を用いた同訓 異字の意味弁別	일본어	논문	2003/11	3,384	○
일본어	7JT_06jk	jj0404jk	공기 관계를 이용한 동훈이자(同訓異字)의 의미 변별	한국어	논문	2003/11	3,125	○
일본어	7JT_07jj	jj0405jj	日韓比較語彙研究にお ける意味の問題につい て	일본어	논문	2003/03/31	2,601	○
일본어	7JT_07jk	jj0405jk	일한비교어휘연구에 있어서 의미의 문제 에 대하여	한국어	논문	2003/03/31	2,490	○
한국어	7JT_08kj	jj0411kj	日本の国語情報化	일본어	논문	2002/05/30	4,041	○
한국어	7JT_08kk	jj0411kk	일본의 국어 정보화	한국어	논문	2002/05/30	3,670	○
한국어	7JT_09kj	jj0412kj	第二外国語の習得と対 訳辞書	일본어	논문	2003/06	3,486	○
한국어	7JT_09kk	jj0412kk	제2외국어 습득과 대 역사전	한국어	논문	2003/06	3,221	○
한국어	7JT_10kj	jj0413kj	コーパスの意味情報を 利用した語彙の対照研 究	일본어	논문	2001/12/09	2,496	○
한국어	7JT_10kk	jj0413kk	코퍼스의 의미정보를	한국어	논문	2001/12/09	2,888	○

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

			이용한 어휘의 대조 연구					
논문 종합							51,875	
일본어	7JT_11jj	jb0401jj	映像革命、そして押井守の世界	일본어	잡지	2004/05	417	×
일본어	7JT_11jk	jb0401jk	영상혁명 그리고 오시이마모루의 세계...	한국어	잡지	2004/05	431	×
일본어	7JT_12jj	jb0402jj	韓国社会のトレンドを読むならホームショッピング	일본어	잡지	2004/05	656	×
일본어	7JT_12jk	jb0402jk	한국 사회의 트렌드를 읽으려면 홈쇼핑을 보라	한국어	잡지	2004/05	688	×
일본어	7JT_13jj	jb0403jj	クイズジャパン	일본어	잡지	2004/05	278	×
일본어	7JT_13jk	jb0403jk	퀴즈재팬	한국어	잡지	2004/05	243	×
일본어	7JT_14jj	jb0404jj	テレビ・ラジオ・照明・舞台美術の世界へ	일본어	잡지	2004/05	493	×
일본어	7JT_14jk	jb0404jk	텔레비전, 라디오, 조명, 무대미술의 세계로	한국어	잡지	2004/05	448	×
일본어	7JT_15jj	jb0405jj	本質を見つめ、学問を追究し、真の現代人を育成する	일본어	잡지	2004/05	395	×
일본어	7JT_15jk	jb0405jk	본질의 응시와 깊은 학문의 연구를 통해 여러분을 참된 현대인으로 육성시키는	한국어	잡지	2004/05	390	×
일본어	7JT_16jj	jb0406jj	おにぎり!	일본어	잡지	2004/05	541	×
일본어	7JT_16jk	jb0406jk	오니기리!	한국어	잡지	2004/05	537	×
일본어	7JT_17jj	jb0407jj	韓国ドラマ	일본어	잡지	2004/05	914	×
일본어	7JT_17jk	jb0407jk	한국드라마	한국어	잡지	2004/05	812	×
일본어	7JT_18jj	jb0408jj	フリマやろうよ	일본어	잡지	2004/05	740	×
일본어	7JT_18jk	jb0408jk	프리마켓에 나가자	한국어	잡지	2004/05	634	×
잡지 종합							8,617	
한국어	7JT_19kj	ja0401kj	チャイナショック	일본어	신문	2004/05/02	267	-
한국어	7JT_19kk	ja0401kk	차이나 쇼크	한국어	신문	2004/05/02	239	-
한국어	7JT_20kj	ja0402kj	マックレーカー	일본어	신문	2004/05/03	252	-
한국어	7JT_20kk	ja0402kk	머크레이커	한국어	신문	2004/05/03	226	-
한국어	7JT_21kj	ja0403kj	国益と同盟	일본어	신문	2004/05/04	284	-

제3장 한일 병렬 말뭉치 개발

한국어	7JT_21kk	ja0403kk	국익과 동맹	한국어	신문	2004/05/04	251	-
한국어	7JT_22kj	ja0404kj	宮廷政治	일본어	신문	2004/05/05	287	-
한국어	7JT_22kk	ja0404kk	궁정정치	한국어	신문	2004/05/05	239	-
한국어	7JT_23kj	ja0405kj	アテネ・オリンピック	일본어	신문	2004/05/06	261	-
한국어	7JT_23kk	ja0405kk	아테네 올림픽	한국어	신문	2004/05/06	226	-
한국어	7JT_24kj	ja0406kj	ダイヤモンド	일본어	신문	2004/05/07	237	-
한국어	7JT_24kk	ja0406kk	다이아몬드	한국어	신문	2004/05/07	215	-
한국어	7JT_25kj	ja0407kj	墮胎と犯罪	일본어	신문	2004/05/09	286	-
한국어	7JT_25kk	ja0407kk	낙태와 범죄	한국어	신문	2004/05/09	239	-
한국어	7JT_26kj	ja0408kj	悪妻分析	일본어	신문	2004/05/10	268	-
한국어	7JT_26kk	ja0408kk	惡妻 분석	한국어	신문	2004/05/10	253	-
한국어	7JT_27kj	ja0409kj	8月の砲声	일본어	신문	2004/05/11	266	-
한국어	7JT_27kk	ja0409kk	8월의 대포	한국어	신문	2004/05/11	261	-
한국어	7JT_28kj	ja0410kj	アトラス	일본어	신문	2004/05/12	242	-
한국어	7JT_28kk	ja0410kk	아틀라스	한국어	신문	2004/05/12	242	-
한국어	7JT_29kj	ja0411kj	性悪説	일본어	신문	2004/05/13	246	-
한국어	7JT_29kk	ja0411kk	性惡說	한국어	신문	2004/05/13	228	-
한국어	7JT_30kj	ja0412kj	ヘッジファンド	일본어	신문	2004/05/14	262	-
한국어	7JT_30kk	ja0412kk	헤지펀드	한국어	신문	2004/05/14	242	-
한국어	7JT_31kj	ja0413kj	逆境指数 (A Q)	일본어	신문	2004/05/16	267	-
한국어	7JT_31kk	ja0413kk	역경지수(AQ)	한국어	신문	2004/05/16	233	-
한국어	7JT_32kj	ja0414kj	自殺予防の音楽	일본어	신문	2004/05/17	284	-
한국어	7JT_32kk	ja0414kk	자살 예방 음악	한국어	신문	2004/05/17	256	-
한국어	7JT_33kj	ja0415kj	在韓米軍のイラク派兵	일본어	신문	2004/05/18	307	-
한국어	7JT_33kk	ja0415kk	미군 차출	한국어	신문	2004/05/18	264	-
한국어	7JT_34kj	ja0416kj	発明家大統領	일본어	신문	2004/05/19	300	-
한국어	7JT_34kk	ja0416kk	발명가 대통령	한국어	신문	2004/05/19	264	-
한국어	7JT_35kj	ja0417kj	シク教 (S i k h)	일본어	신문	2004/05/20	275	-
한국어	7JT_35kk	ja0417kk	시크(Sikh)	한국어	신문	2004/05/20	232	-
한국어	7JT_36kj	ja0418kj	サルトシェバ덴協約	일본어	신문	2004/05/21	270	-
한국어	7JT_36kk	ja0418kk	샬트셰바덴 협약	한국어	신문	2004/05/21	228	-
한국어	7JT_37kj	ja0419kj	ほどほどに務められた方	일본어	신문	2004/05/23	286	-
한국어	7JT_37kk	ja0419kk	어지간히 하신 분	한국어	신문	2004/05/23	233	-
한국어	7JT_38kj	ja0420kj	ウォッチタワー	일본어	신문	2004/05/24	259	-
한국어	7JT_38kk	ja0420kk	워치타워	한국어	신문	2004/05/24	250	-
한국어	7JT_39kj	ja0421kj	日本の二重良心	일본어	신문	2004/05/25	296	-
한국어	7JT_39kk	ja0421kk	일본의 이중양심	한국어	신문	2004/05/25	271	-
한국어	7JT_40kj	ja0422kj	移民政策	일본어	신문	2004/05/26	285	-
한국어	7JT_40kk	ja0422kk	이민 정책	한국어	신문	2004/05/26	242	-

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

한국어	7JT_41kj	ja0423kj	チャラビ	일본어	신문	2004/05/27	291	-
한국어	7JT_41kk	ja0423kk	찰라비	한국어	신문	2004/05/27	229	-
한국어	7JT_42kj	ja0424kj	トヨタとホンダ	일본어	신문	2004/05/28	266	-
한국어	7JT_42kk	ja0424kk	도요타와 혼다	한국어	신문	2004/05/28	242	-
한국어	7JT_43kj	ja0425kj	食教育	일본어	신문	2004/05/30	262	-
한국어	7JT_43kk	ja0425kk	食교육	한국어	신문	2004/05/30	219	-
한국어	7JT_44kj	ja0426kj	メソの秘訣	일본어	신문	2004/05/31	251	-
한국어	7JT_44kk	ja0426kk	매추의 비결	한국어	신문	2004/05/31	226	-
신문 종합							13,307	
전체 종합							121,335	

주) ○ : 저작권 승인, × : 저작권 승인 안 됨, - : 저작권과 관련 없는 자료

3. 한일 형태소 분석 말뭉치 구축 도구 및 구축 방법

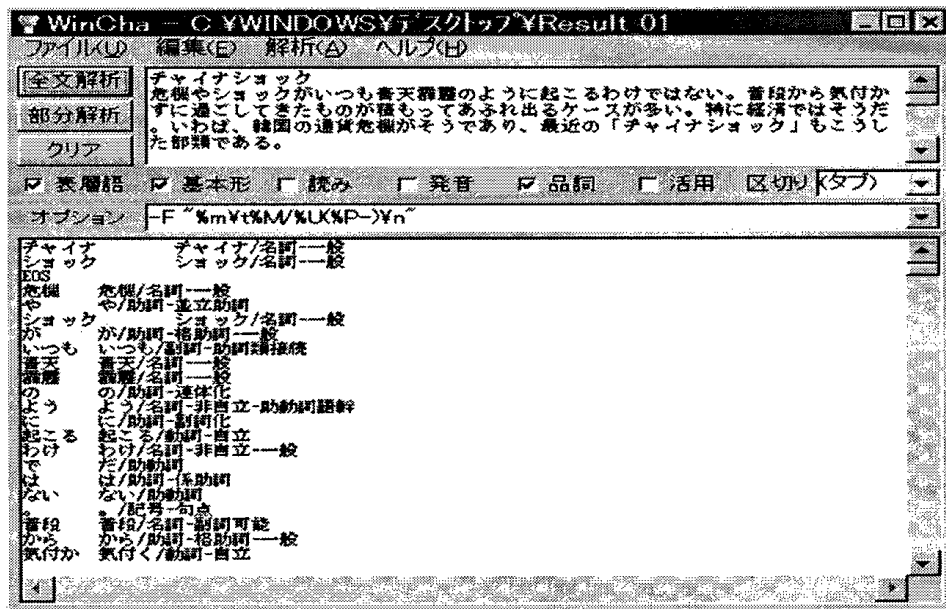
3.1. 구축 도구

한일 형태소 분석 말뭉치 구축 도구로서는 한국어 형태소 분석 말뭉치와 일본어 형태소 분석 말뭉치로 나누어 생각할 수 있다.

우선 한국어 형태소 분석 말뭉치의 경우, 기본적으로 한영 병렬 말뭉치 구축 분과에서 작성한 지침서(한영 병렬 한국어 문서 형태소 분석 말뭉치 구축 지침 <2004-10-국 ver 3.0>)에 의거하여 진행하고 있으므로, 구축 도구도 한영 병렬 말뭉치 구축 분과에서 사용하고 있는 “SORITAG”을 그대로 적용하기로 하였다.

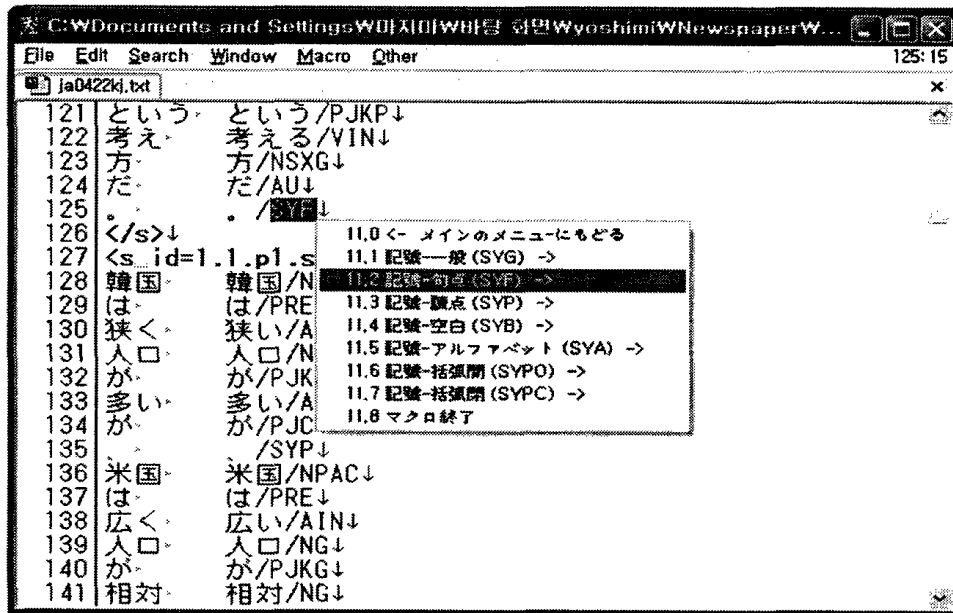
다음으로 일본어 형태소 분석 말뭉치의 경우, 작년도와 마찬가지로 일본 나라(奈良) 첨단기술대학 대학원의 마쓰모토 유지(松本裕治) 연구실에서 개발한 일본어 형태소 분석기 “차센(chasen)”을 사용하기로 하였다.

【일본어 형태소 분석기 “차센(chasen)”의 동작】



또한 검수 작업용 보조 프로그램으로서 일본 텍스트 에디터 “Maruo(마루오)”를 사용하여 필요에 따라 작업용 매크로를 작성해 활용하도록 하였다.

【“Maruo(마루오)”를 사용한 형태소 분석 작업 매크로의 동작】



3.2. 구축 방법

한일 병렬 형태소 분석 말뭉치 구축 방법은 2003년도 최종보고서에 제시한 방법과 크게 달라진 점은 없다. 달라진 점은 작년과 동일한 기준으로 작업을 진행하면서 작업하기 쉬운 매크로가 몇 가지 더 늘어났다는 점이다.

대략적인 작업 순서는 우선 한국어, 일본어 각 형태소 분석기로 1차적으로 작업을 한 뒤, 2차적으로 텍스트 에디터 “Maruo(마루오)”를 이용해 수정 및 편집 작업을 수행하게 된다. 모든 검수 작업을 마친 뒤 헤더를 부여하여 형태소 분석 말뭉치 자료로서의 형식을 갖추게 된다.

4. 논의 사항에 대한 견해

4.1 분석 결과의 “+” 앞뒤의 스페이스 입력

- 기초분과 측에서는 “+” 앞뒤의 스페이스를 입력하면, 본문에 “+”가 나올 경우 구별이 가능하므로 “+” 앞뒤의 스페이스를 입력하는 것이 좋다는 입장

“+” 앞뒤의 스페이스 입력의 유무는 단순한 보기상의 문제가 아니라 분석 시스템상의 기술적인 문제이므로, 시스템 개발자의 의견을 수렴하여 시스템 상 어떤 기술 방법이 적합한지를 알아볼 필요가 있다고 생각한다.

한일 병렬 말뭉치 구축 분과에서는 2004년도 작업 시에는 구어 전사 말뭉치 구축 분과의 기준을 따라 “+” 앞뒤에 스페이스를 넣지 않았으나, 여러 조건을 감안할 때 향후 “+” 앞뒤에 스페이스를 넣는 것이 좋다는 생각을 가지고 있다.

4.2 어떤 사전의 어떤 점을 기준으로 분석할 것인지의 문제

- 사전의 표제항을 기본으로 받아들이는 문제

모든 작업에 공통되는 것이지만, 방대한 작업을 규칙성 있게 진행하는 것은 상당히 어려운 일이다. 어느 정도 통일된 지침이 정해지더라도 그 작업을 모두 수작업으로 해나가면 입력 오류 등으로 인하여 결국 지침대로의 최종결과물이 나오지 않을 것이다. 대략적인 태그세트가 정해지면 1차적으로 그 태그세트로 형태소 태깅을 할 수 있는 툴을 개발하여 그 다음에 문제가 되는 부분을 수작업으로 수정해나가는 방법이 가장 빠르고 정확한 결과를 얻을 수 있다고 본다. 앞으로 개발될 말뭉치 활용 시스템을 실제로 사용할 때도 큰 문제가 생길 우려가 있을 만큼 말뭉치 활용 시스템의 개발뿐만이 아니라 말뭉치 구축 작업 및 검수를 도울 수 있는 툴의 개발 또한 힘을 기울여야 할 것이다.

5. 한일 병렬 형태소 분석 말뭉치 구축의 실례

5.1 일본어 형태소 분석 말뭉치 (7JT_16jj.txt)

```
<ces>
<cesHeader version="2.3">
  <fileDesc>
    <titleStmt>
      <h.title>おにぎり!, 형태소 분석 전자파일</h.title>
      <respStmt>
        <respType>헤더붙임, 형태 주석</respType>
        <respName>고려대학교 한일병렬말뭉치 구축팀</respName>
      </respStmt>
    </titleStmt>
    <extent>
      <wordcount>541</wordcount>
      <bytecount>4096</bytecount>
    </extent>
    <publicationStmt>
      <distributor>국립국어원</distributor>
      <pubAddress>서울시 강서구 방화3동 827번지</pubAddress>
      <availability status="restricted">배포 불가</availability>
      <idno>7JT_16jj.txt, 원시말뭉치 : jb0406jj.txt</idno>
      <PubDate>2004</PubDate>
    </publicationStmt>
    <sourceDesc>
      <biblStruct>
        <monogr>
```

```

<h.title>おにぎり！</h.title>
<h.author></h.author>
<imprint>
    <pubPlace>日本</pubPlace>
    <publisher>月刊留学生</publisher>
    <pubDate>2004/05</pubDate>
</imprint>
</monogr>
</biblStruct>
</sourceDesc>
</fileDesc>
<encodingDesc>
    <projectDesc>21세기 세종계획 3단계 1차년도 한일 병렬 말뭉치 구축
</projectDesc>
    <editorialDecl>21세기 세종계획 말뭉치 문헌 입력 지침 및 한일 병렬
한국어 형태소 주석 지침에 따름</editorialDecl>
    <samplingDecl></samplingDecl>
    <refsDecl>본문, 구획, 문단, 문장 단위로 일련번호를 붙여 병렬텍스트
와 정렬시킨 후 형태 주석 정보를 어절의 하위 형태 단위에 태깅</refsDecl>
</encodingDesc>
<profileDesc>
    <creation><date>2004/11/03</date></creation>
    <langUsage>
        <language id="kr" iso639="kr"></language>
    </langUsage>
    <wsdUsage>
        <writingSystem id="EUC-KR(KSC5601)"></writingSystem>
    </wsdUsage>
    <textClass>

```

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

```
<catRef scheme='SJ21' target='R1320'>책 : 잡지</catRef>
</textClass>
<translations>
  <translation          trans.loc="7JT_16jk.txt"          lang="kr'
wsd="UTF-16"></translation>
  <translator></translator>
</translations>
</profileDesc>
<revisionDesc>
  <change>
    <changeDate>2004/11/03</changeDate>
    <respName>최장원</respName>
    <respName>전형식</respName>
    <h.item>본문 마크업, 형태 주석</h.item>
  </change>
</revisionDesc>
</cesHeader>
<text>
<body id=1>
<div id=1.1>
<head id=1.1.h1>
  おにぎり おにぎり/NG
  !      !/SYG
</head>
<p id=1.1.p1>
<s id=1.1.p1.s1>
  日本    日本/NPAC
  の      の/PCS
  どの    どの/CS
```

コンビニ コンビニ/NG
 に に/PJKG
 行っ 行く/VIN
 て て/PJC
 も も/PRE
 、 、/SYP
 食品 食品/NG
 コーナー コーナー/NG
 に に/PJKG
 は は/PRE
 必ず 必ず/ADCP
 おにぎり おにぎり/NG
 が が/PJKG
 陳列 陳列/NCPV
 さ する/VIN
 れ れる/VSX
 て て/PJC
 いる いる/VDE
 。 。/SYF
 </s>
 <s id=1.1.pl.s2>
 留学生 留学生/NG
 も も/PRE
 、 、/SYP
 誰 誰/NNPG
 も も/PRE
 が が/PJKG
 1 1 /NNR
 度 度/NSXC

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

は は/PRE
 コンビニ コンビニ/NG
 で で/PJKG
 おにぎり おにぎり/NG
 を を/PJKG
 買っ 買う/VIN
 て て/PJC
 食べ 食べる/VIN
 た た/AU
 経験 経験/NCPV
 が が/PJKG
 ある ある/VIN
 だろ だ/AU
 う う/AU
 。 。/SYF
 </s>
 <s id=1.1.pl.s3>
 しかし しかし/CJ
 、 、/SYP
 日本 日本/NPAC
 に に/PJKG
 来 来る/VIN
 て て/PJC
 初めて 初めて/ADG
 おにぎり おにぎり/NG
 を を/PJKG
 買っ 買う/VIN
 た た/AU
 時 時/NDEADP

、	、/SYP
韓国	韓国/NPAC
と	と/PJKG
違っ	違う/VIN
た	た/AU
多様	多様/NAVS
な	だ/AU
包装	包装/NCPV
と	と/PPA
いくつ	いくつ/NNPG
も	も/PRE
の	の/PCS
種類	種類/NG
を	を/PJKG
前	前/NADP
に	に/PJKG
し	する/VIN
て	て/PJC
、	、/SYP
何	何/NNPG
を	を/PJKG
食べ	食べる/VIN
たら	た/AU
いい	いい/ADE
の	の/NDEG
か	か/PAM
、	、/SYP
選ぶ	選ぶ/VIN
の	の/NDEG

に に/PJKG
ずいぶん ずいぶん/ADCP
迷っ 迷う/VIN
た た/AU
記憶 記憶/NCPV
も も/PRE
ある ある/VIN
だろ だ/AU
う う/AU
。 。/SYF
</s>
</p>

----- <중 략> -----

<p id=1.1.p33>
<s id=1.1.p33.s1>
日頃 日頃/NADP
は は/PRE
何 何/NNPG
も も/PRE
考え 考える/VIN
ず ぬ/AU
に に/PJKG
買っ 買う/VIN
て て/PJC
食べ 食べる/VIN
て て/PJC
い いる/VDE

た た/AU
 おにぎり おにぎり/NG
 の の/PCS
 種類 種類/NG
 や や/PPA
 形 形/NG
 など など/PVI
 、 、/SYP
 その その/CS
 多彩 多彩/NAVS
 ふり ふり/NSXG
 を を/PJKG
 知っ 知る/VIN
 て て/PJC
 みる みる/VDE
 と と/PJC
 、 、/SYP
 何故か 何故か/ADG
 ますます ますます/ADCP
 食べ 食べる/VIN
 たく たい/AU
 なる なる/VIN
 。 。/SYF
 </s>
 <s id=1.1.p33.s2>
 今 今/NADP
 まで まで/PVI
 おにぎり おにぎり/NG
 と と/PJKG

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

仲良く 仲良く/ADG

し する/VIN

て て/PJC

い いる/VDE

なかつ ない/AU

た た/AU

留学生 留学生/NG

達 達/NSXG

は は/PRE

、 、/SYP

今日 今日/NADP

、 、/SYP

コンビニ コンビニ/NG

に に/PJKG

寄っ 寄る/VIN

て て/PJC

1つ 1つ/NG

買っ 買う/VIN

て て/PJC

みよ みる/VDE

う う/AU

。 。/SYF

</s>

<s id=1.1.p33.s3>

好み 好み/NG

の の/PCS

おにぎり おにぎり/NG

は は/PRE

どれ どれ/NNPG

```

だろ    だ/AU
う      う/AU
?       ?/SYG
</s>
</p>
</div>
</body>
</text>
</ces>

```

5.2 한국어 형태소 분석 말뭉치 (7JT_16jk.txt)

```

<ces>
<cesHeader version="2.3">
  <fileDesc>
    <titleStmt>
      <h.title>おにぎり!, 형태소 분석 전자파일</h.title>
      <respStmt>
        <respType>헤더붙임, 형태 주석</respType>
        <respName>고려대학교 한일병렬말뭉치 구축팀</respName>
      </respStmt>
    </titleStmt>
    <extent>
      <wordcount>541</wordcount>
    </extent>
    <publicationStmt>
      <distributor>국립국어원</distributor>
      <pubAddress>서울시 강서구 방화3동 827번지</pubAddress>
      <availability status="restricted">배포 불가</availability>
    </publicationStmt>
  </fileDesc>

```

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

```
<idno>7JT_16jk.txt, 원시말뭉치 : jb0406jk.txt</idno>
  <PubDate>2004</PubDate>
</publicationStmt>
<sourceDesc>
  <biblStruct>
    <monogr>
      <h.title>おにぎり ! </h.title>
      <h.author></h.author>
      <imprint>
        <pubPlace>日本</pubPlace>
        <publisher>月刊留学生</publisher>
        <pubDate>2004/05</pubDate>
      </imprint>
    </monogr>
  </biblStruct>
</sourceDesc>
</fileDesc>
<encodingDesc>
  <projectDesc>21세기 세종계획 3단계 1차년도 한일 병렬 말뭉치 구축
</projectDesc>
  <editorialDecl>21세기 세종계획 말뭉치 문헌 입력 지침 및 한일 병렬
한국어 형태소 주석 지침에 따름</editorialDecl>
  <samplingDecl></samplingDecl>
  <refsDecl>본문, 구획, 문단, 문장 단위로 일련번호를 붙여 병렬텍스트와
정렬시킨 후 형태 주석 정보를 어절의 하위 형태 단위에 태깅
</refsDecl>
</encodingDesc>
<profileDesc>
  <creation><date>2004/11/03</date></creation>
```

```

<langUsage>
  <language id="kr" iso639="kr"></language>
</langUsage>
<wsdUsage>
  <writingSystem id="EUC-KR(KSC5601)"></writingSystem>
</wsdUsage>
<textClass>
  <catRef scheme='SJ21' target="R1320">책 : 잡지</catRef>
</textClass>
<translations>
  <translation          trans.loc="7JT_16jj.txt"          lang="jp"
wsd="UTF-16"></translation>
  <translator></translator>
</translations>
</profileDesc>
<revisionDesc>
  <change>
    <changeDate>2004/11/03</changeDate>
    <respName>최장원</respName>
    <respName>전형식</respName>
    <h.item>본문 마크업, 형태 주석</h.item>
  </change>
</revisionDesc>
</cesHeader>
<text>
<body id=1>
<div id=1.1>
<head id=1.1.h1>
오니기리!      오니기리/SL+!/SF

```

</head>

<p id=1.1.p1>

<s id=1.1.p1.s1>

일본의 일본/NNP+의/JKG

어느 어느/MM

편의점에를 편의점/NNG+에/JKB+를/JKO

가더라도 가/VX+더라/EF+도/JX

식품코너에는 식품/NNG+코너/NNG+에/JKB+는/JX

반드시 반드시/MAG

오니기리가 오니기리/SL+가/JKS

진열되어 진열/NNG+되/XSV+어/EC

있다. 있/VA+다/EF+./SF

</s>

<s id=1.1.p1.s2>

그렇기 그렇/VX+기/ETN

때문에 때문/NNB+에/JKB

유학생 유학/NNG+생/NNB

누구나가 누구/NP+나/NNB+가/JKS

한 한/MM

번 번/NNB

이상은 이상/NNB+은/JX

편의점에서 편의점/NNG+에서/JKB

오니기리를 오니기리/SL+를/JKO

사 사/MAG

먹어 먹/VV+어/EC

본 본/NNP

경험이 경험/NNG+이/JKS

있을 있/VX+을/ETM

것이다. 것/NNB+이/VCP+다/EF+./SF

</s>
 <s id=1.1.p1.s3>
 그러나 그러나/MAJ
 일본에 일본/NNG+에/JKB
 와서 오/VV+ㅓ서/EC
 맨 맨/MM
 처음 처음/MAG
 오니기리를 오니기리/SL+를/JKO
 샀을 샀/VV+을/JKO
 때, 때/NNG+,/SP
 한국과 한국/NNP+과/JC
 달리 달리/MAG
 다양한 다양/NNG+하/XSA+ㄴ/ETM
 포장과 포장/NNP+과/JKB
 여러 여러/MM
 종류 종류/NNG
 앞에 앞/NNG+에/JKB
 무엇을 무엇/NP+을/JKO
 먹어야 먹/VV+어야/EC
 할 하/VX+ㄹ/ETM
 지 지/NNB
 고르는데 고르/VX+는데/EC
 무척 무척/MAG
 망설였던 망설/NNG+이/VCP+ㄴ쓰/EP+던/ETM
 기억이 기억/NNG+이/JKS
 있을 있/VX+을/ETM
 것이다. 것/NNB+이/VCP+다/EF+./SF
 </s>
 </p>

----- <중 략> -----

<p id=1.1.p33>

<s id=1.1.p33.s1>

평소에는 평소/NNG+에/JKB+는/JX

아무 아무/MM

생각 생각/NNG

없이 없이/MAG

사먹던 사먹/VX+던/ETM

오니기리의 오니기리/SL+의/JKG

종류나 종류나/NP

모양 /NNG+모양/NNG

등의 등/NNG+의/JKG

그 그/MM

다양함을 다양하/VX+ㅁ/ETN+을/JKO

알게 알/VV+게/EC

되니 되/VV+니/EC

왠지 왠지/NNB

더욱 더욱/MAG

먹고 먹/VV+고/EC

싫어진다. 싫어지/VV+ㄴ다/EF+./SF

</s>

<s id=1.1.p33.s2>

아직까지 아직/MAG+까지/JX

오니기리와 오니기리/SL+와/JC

친하지 친하/VX+지/EC

않은 았/VX+은/ETM

유학생들은 유학/NNG+생/NNB+들/XSN+은/JX

오늘 오늘/NNG
 편의점에 편의점/NNG+에/JKB
 들러 들르/VV+ㅓ/EC
 한두 한두/MM
 개 개/NNB
 사 사/MAG
 보자. 보/VX+자/EF+./SF
 </s>
 <s id=1.1.p33.s3>
 과연 과연/MAG
 내 내/MM
 입맛에 입맛에/JKB
 맞는 맞/VV+는/ETM
 오니기리는? 오니기리/SL+는/JX+?/SF
 </s>
 </p>
 </div>
 </body>
 </text>
 </ces>

5.3 형태소 분석 정렬 파일 (7JT_16a1.txt)

```

<ces>
<cesAlign type=sent version=1.6>
<cesHeader version="2.3">
  <fileDesc>
    <titleStmt>
      <h.title>잡지/おにぎり!</h.title>
  
```



```
<publicationStmt>
  <distributor>국립국어원</distributor>
  <pubAddress>서울시 강서구 방화3동 827번지</pubAddress>
  <availability status=restricted>배포 불가</availability>
  <PubDate>2003</PubDate>
</publicationStmt>
<sourceDesc>
  <biblStruct>
    <monogr>
      <h.title>오니ぎり!</h.title>
      <h.author></h.author>
      <imprint>
        <pubPlace>日本</pubPlace>
        <publisher>月刊留学生</publisher>
        <pubDate>2004/05</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">
  <link xtargets="1.1.p1.s2 ; 1.1.p1.s2">
  <link xtargets="1.1.p1.s3 ; 1.1.p1.s3">
```

```

</linkGrp>
<linkGrp domains="1.1.p2 ; 1.1.p2">
    <link xtargets="1.1.p2.s1 ; 1.1.p2.s1">
</linkGrp>
<linkGrp domains="1.1.p3 ; 1.1.p3">
    <link xtargets="1.1.p3.s1 ; 1.1.p3.s1">
</linkGrp>
<linkGrp domains="1.1.p4 ; 1.1.p4">
    <link xtargets="1.1.p4.s1 ; 1.1.p4.s1">
</linkGrp>
<linkGrp domains="1.1.p5 ; 1.1.p5">
    <link xtargets="1.1.p5.s1 ; 1.1.p5.s1">
</linkGrp>
<linkGrp domains="1.1.p6 ; 1.1.p6">
    <link xtargets="1.1.p6.s1 ; 1.1.p6.s1">
    <link xtargets="1.1.p6.s2 ; 1.1.p6.s2">
    <link xtargets="1.1.p6.s3 ; 1.1.p6.s3">
</linkGrp>
<linkGrp domains="1.1.p7 ; 1.1.p7">
    <link xtargets="1.1.p7.s1 ; 1.1.p7.s1">
</linkGrp>
<linkGrp domains="1.1.p8 ; 1.1.p8">
    <link xtargets="1.1.p8.s1 ; 1.1.p8.s1">
    <link xtargets="1.1.p8.s2 ; 1.1.p8.s2">
</linkGrp>
<linkGrp domains="1.1.p9 ; 1.1.p9">
    <link xtargets="1.1.p9.s1 ; 1.1.p9.s1">
</linkGrp>
<linkGrp domains="1.1.p10 ; 1.1.p10">

```

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

```
<link xtargets="1.1.p10.s1 ; 1.1.p10.s1">
  <link xtargets="1.1.p10.s2 ; 1.1.p10.s2">
</linkGrp>
<linkGrp domains="1.1.p11 ; 1.1.p11">
  <link xtargets="1.1.p11.s1 ; 1.1.p11.s1">
</linkGrp>
<linkGrp domains="1.1.p12 ; 1.1.p12">
  <link xtargets="1.1.p12.s1 ; 1.1.p12.s1">
  <link xtargets="1.1.p12.s2 ; ">
</linkGrp>
<linkGrp domains="1.1.p13 ; 1.1.p13">
  <link xtargets="1.1.p13.s1 ; 1.1.p13.s1">
</linkGrp>
<linkGrp domains="1.1.p14 ; 1.1.p14">
  <link xtargets="1.1.p14.s1 ; 1.1.p14.s1">
  <link xtargets="1.1.p14.s2 ; 1.1.p14.s2">
  <link xtargets="1.1.p14.s3 ; 1.1.p14.s3">
</linkGrp>
<linkGrp domains="1.1.p15 ; 1.1.p15">
  <link xtargets="1.1.p15.s1 ; 1.1.p15.s1">
</linkGrp>
<linkGrp domains="1.1.p16 ; 1.1.p16">
  <link xtargets="1.1.p16.s1 ; 1.1.p16.s1">
  <link xtargets="1.1.p16.s2 ; 1.1.p16.s2">
  <link xtargets="1.1.p16.s3 ; 1.1.p16.s3">
</linkGrp>
<linkGrp domains="1.1.p17 ; 1.1.p17">
  <link xtargets="1.1.p17.s1 ; 1.1.p17.s1">
</linkGrp>
```

```

<linkGrp domains="1.1.p18 ; 1.1.p18">
    <link xtargets="1.1.p18.s1 ; 1.1.p18.s1">
    <link xtargets="1.1.p18.s2 ; ">
</linkGrp>
<linkGrp domains="1.1.p19 ; 1.1.p19">
    <link xtargets="1.1.p19.s1 ; 1.1.p19.s1">
</linkGrp>
<linkGrp domains="1.1.p20 ; 1.1.p20">
    <link xtargets="1.1.p20.s1 ; 1.1.p20.s1">
</linkGrp>
<linkGrp domains="1.1.p21 ; 1.1.p21">
    <link xtargets="1.1.p21.s1 ; 1.1.p21.s1">
</linkGrp>
<linkGrp domains="1.1.p22 ; 1.1.p22">
    <link xtargets="1.1.p22.s1 ; 1.1.p22.s1">
</linkGrp>
<linkGrp domains="1.1.p23 ; 1.1.p23">
    <link xtargets="1.1.p23.s1 ; 1.1.p23.s1">
    <link xtargets="1.1.p23.s2 ; 1.1.p23.s2">
    <link xtargets="1.1.p23.s3 ; 1.1.p23.s3">
    <link xtargets="1.1.p23.s4 ; 1.1.p23.s4">
    <link xtargets="1.1.p23.s5 ; 1.1.p23.s5">
    <link xtargets="1.1.p23.s6 ; 1.1.p23.s6">
</linkGrp>
<linkGrp domains="1.1.p24 ; 1.1.p24">
    <link xtargets="1.1.p24.s1 ; 1.1.p24.s1">
</linkGrp>
<linkGrp domains="1.1.p25 ; 1.1.p25">
    <link xtargets="1.1.p25.s1 ; 1.1.p25.s1">

```

```
</linkGrp>
<linkGrp domains="1.1.p26 ; 1.1.p26">
  <link xtargets="1.1.p26.s1 ; 1.1.p26.s1">
</linkGrp>
<linkGrp domains="1.1.p27 ; 1.1.p27">
  <link xtargets="1.1.p27.s1 ; 1.1.p27.s1">
  <link xtargets="1.1.p27.s2 ; 1.1.p27.s2">
</linkGrp>
<linkGrp domains="1.1.p28 ; 1.1.p28">
  <link xtargets="1.1.p28.s1 ; 1.1.p28.s1">
  <link xtargets="1.1.p28.s2 ; 1.1.p28.s2">
</linkGrp>
<linkGrp domains="1.1.p29 ; 1.1.p29">
  <link xtargets="1.1.p29.s1 ; 1.1.p29.s1">
</linkGrp>
<linkGrp domains="1.1.p30 ; 1.1.p30">
  <link xtargets="1.1.p30.s1 ; 1.1.p30.s1">
</linkGrp>
<linkGrp domains="1.1.p31 ; 1.1.p31">
  <link xtargets="1.1.p31.s1 ; 1.1.p31.s1">
</linkGrp>
<linkGrp domains="1.1.p32 ; 1.1.p32">
  <link xtargets="1.1.p32.s1 ; 1.1.p32.s1">
  <link xtargets="1.1.p32.s2 ; 1.1.p32.s2">
  <link xtargets="1.1.p32.s3 s4 ; 1.1.p32.s3">
  <link xtargets="1.1.p32.s5 ; 1.1.p32.s4">
</linkGrp>
<linkGrp domains="1.1.p33 ; 1.1.p33">
  <link xtargets="1.1.p33.s1 ; 1.1.p33.s1">
```

```
<link xtargets="1.1.p33.s2 ; 1.1.p33.s2">  
  <link xtargets="1.1.p33.s3 ; 1.1.p33.s3">  
</linkGrp>  
</linkGrp>  
</linkGrp>  
</linkList>  
</cesAlign>  
</ces>
```

〈부록〉 한일 병렬 일본어 문서 형태소 분석 말뭉치 구축 지침 (2004-11-일 ver 2.0)

1. 기본 원칙

- 한일 병렬 말뭉치의 경우, 한국어 문서는 한영 형태소 분석 말뭉치의 한국어 문서 지침에 따르기로 하고, 일본어 문서는 한일 병렬 말뭉치 구축 분과에서 규정하는 지침으로 작업을 진행하기로 한다. 단, 문서 기술 방식에 있어서는 약간의 차이가 있어 이에 유의해야 한다. 한영 병렬 말뭉치 구축 분과의 경우 원시 말뭉치는 SGML 방식으로, 형태소 분석 말뭉치는 XML 방식으로 되어 있으나, 한일 병렬 말뭉치 구축 분과의 경우, 한영 병렬 말뭉치 구축 분과에서 채택된 두 개의 기술 방식을 통합하여 모두 SGML 방식으로 기술하기로 한다.
- 형태소 주석 부여 방법은 일본 나라(奈良) 첨단기술 대학 대학원의 마쓰모토 유지(松本裕治) 연구실에서 개발한 일본어 형태소 분석기 “차센(chasen)”의 분석 결과를 기반으로 한다.
- 한국어 문서와 일본어 문서의 분석 결과 기술 방식은 각각의 어휘 체계 및 문법 체계 등에 맞추어 작성하기로 한다. 따라서 한국어 문서는 어절 단위로 형태소 분석이 이루어지며, 일본어 문서의 경우 형태소 단위로 분석이 이루어진다.
- 형태소 분석 기준과 사용 용어 문제는 각 언어 간의 문법 용어를 최대한 사용하여 영문 약자 표기를 고안하되, 명사는 N, 동사는 V와 같이 형태소의 대분류에 있어서는 가능한 한 용어 통일을 하도록 고려한다.
- 작업 시에 발견된 새로 기준을 적용해야할 부분에 대해서는 한일 병렬 말뭉

치 구축 분과 내 정기회의 때 검토를 한 후 지침에 추가하여 계속적으로 지침서를 보완해나가기로 한다.

2. 기본 태그세트

- 이 태그세트는 일본어 형태소 분석기 “차센(chasen)”의 분석 결과를 기반으로 작성한 것이다.
- 태그세트의 영문 표기는 한일 병렬 말뭉치 구축 분과가 고안한 것으로 “차센(chasen)”의 분석 결과에 충실하게 영문 약자화한 것이다.
- 태그세트 중에는 “차센(chasen)”에서 분석 불가능한 새 항목도 포함되어 있으며, 그 분류 기준 및 영문 표기 등에 대해서는 한일 병렬 말뭉치 구축 분과 내에서 검토하여 설정한 것이다. 추가적으로 설정한 태그세트에 대해서는 “한일 병렬 일본어 문서 형태소 분석 말뭉치 구축 지침(2004-11-일 ver 2.0)”의 보완 작업을 통해 반영하도록 하여, 세부 기준의 변경 사항은 다음 버전의 지침서에 명기하기로 한다.
- 일본어 태그세트 일람
 - 1.1 名詞—一般 (NG)
 - 1.2 名詞—固有名詞—一般 (NPG)
 - 1.3 名詞—固有名詞—人名—一般 (NPNG)
 - 1.4 名詞—固有名詞—人名—姓 (NPNL)
 - 1.5 名詞—固有名詞—人名—名 (NPNF)
 - 1.6 名詞—固有名詞—組織 (NPO)
 - 1.7 名詞—固有名詞—地域—一般 (NPAG)
 - 1.8 名詞—固有名詞—地域—国 (NPAC)
 - 1.9 名詞—代名詞—一般 (NNPG)
 - 1.10 名詞—代名詞—縮約 (NNPR)
 - 1.11 名詞—副詞可能 (NADP)
 - 1.12 名詞—サ変接続 (NCPV)

- 1.13 名詞-形容動詞語幹 (NAVS)
- 1.14 名詞-ナイ形容詞語幹 (NAS)
- 1.15 名詞-数 (NNR)
- 1.16 名詞-非自立-一般 (NDEG)
- 1.17 名詞-非自立-副詞可能 (NDEADP)
- 1.18 名詞-非自立-助動詞語幹 (NDEAUS)
- 1.19 名詞-非自立-形容動詞語幹 (NDEAVS)
- 1.20 名詞-特殊-助動詞語幹 (NEAUS)
- 1.21 名詞-接尾-一般 (NSXG)
- 1.22 名詞-接尾-人名 (NSXN)
- 1.23 名詞-接尾-地域 (NSXA)
- 1.24 名詞-接尾-サ変接続 (NSXCPV)
- 1.25 名詞-接尾-助動詞語幹 (NSXAUS)
- 1.26 名詞-接尾-形容動詞語幹 (NSXAVS)
- 1.27 名詞-接尾-副詞可能 (NSXADP)
- 1.28 名詞-接尾-助数詞 (NSXC)
- 1.29 名詞-接尾-特殊 (NSXE)
- 1.30 名詞-接續詞的 (NCJ)
- 1.31 名詞-動詞非自立的 (NVDE)
- 1.32 名詞-引用文字列 (NQ)

- 2.1 接頭詞-名詞接続 (PXCN)
- 2.2 接頭詞-数接続 (PXCC)
- 2.3 接頭詞-動詞接続 (PXCv)
- 2.4 接頭詞-形容詞接続 (PXCA)

- 3.1 動詞-自立 (VIN)
- 3.2 動詞-非自立 (VDE)
- 3.3 動詞-接尾 (VSX)

4.1 形容詞-自立 (AIN)

4.2 形容詞-非自立 (ADE)

4.3 形容詞-接尾 (ASX)

5.1 副詞-一般 (ADG)

5.2 副詞-助詞類接統 (ADCP)

6.1 連體詞 (CS)

7.1 接統詞 (CJ)

8.1 助詞-格助詞-一般 (PJKG)

8.2 助詞-格助詞-引用 (PJKQ)

8.3 助詞-格助詞-連語 (PJKP)

8.4 助詞-接統助詞 (PJC)

8.5 助詞-係助詞 (PRE)

8.6 助詞-副助詞 (PVI)

8.7 助詞-間投助詞 (PIJ)

8.8 助詞-並立助詞 (PPA)

8.9 助詞-終助詞 (PEN)

8.10 助詞-副助詞／並立助詞／終助詞 (PAM)

8.11 助詞-連体化 (PCS)

8.12 助詞-副詞化 (PAD)

8.13 助詞-特殊 (PE)

9.1 助動詞 (AU)

10.1 感動詞 (EX)

- 11.1 記号-一般 (SYG)
- 11.2 記号-句点 (SYF)
- 11.3 記号-読点 (SYP)
- 11.4 記号-空白 (SYB)
- 11.5 記号-アルファベット (SYA)
- 11.6 記号-括弧開 (SYPO)
- 11.7 記号-括弧閉 (SYPC)

- 12.1 その他-間投 (OAM)
- 12.2 その他-ファイラー (F)
- 12.3 その他-外国語 (FL)
- 12.4 その他-ホームページ住所 (URL)

3. 일본어 태그세트 분류 기준 및 세부 지침

3.1 명사

3.1.1 名詞-一般 (NG)

보통명사 혹은 하위분류가 미정인 명사.

3.1.2 名詞-固有名詞-一般 (NPG)

일반적인 고유명사 혹은 하위분류가 미정인 고유명사.

3.1.3 名詞-固有名詞-人名-一般 (NPNG)

성과 이름으로 나눌 수 없는 것. 외국인 이름. 혹은 성, 이름의 결정이 미정인 인명.

【例】お市の方

3.1.4 名詞-固有名詞-人名-姓 (NPNL)

주로 일본인 성에 해당됨.

【例】山田, 田中, 佐藤

3.1.5 名詞-固有名詞-人名-名 (NPNF)

주로 일본인 이름 부분에 해당됨.

【例】太郎, 花子, 一郎

3.1.6 名詞-固有名詞-組織 (NPO)

조직을 나타내는 명칭.

【例】通産省, NHK

3.1.7 名詞-固有名詞-地域-一般 (NPAG)

국명 이외의 지명을 나타내는 것.

【例】アジア, バルセロナ, 京都

3.1.8 名詞-固有名詞-地域-国 (NPAC)

나라 이름을 나타냄.

【例】日本, オーストラリア

3.1.9 名詞-代名詞-一般 (NNPG)

소위 말하는 대명사, 불특정한 語.

【例】それ, ここ, あいつ, あなた, あちこち, いくつ, どこか, なに

3.1.10 名詞-代名詞-縮約 (NNPR)

대명사와 係助詞「は」의 복합형으로 축약된 것. 구어체로 사용됨.

【例】ありゃ, こりゃ, こりゃあ, そりゃ, そりゃあ

3.1.11 名詞-副詞可能 (NADP)

요일, 월 등, 시간을 나타내는 부사적인 용법을 가진 명사. 양이나 정도를 나타내며 부사적으로 사용이 가능한 명사.

【例】金曜, 一月, 午後, 少量

3.1.12 名詞-サ変接続 (NCPV)

격 요소를 요구하며 「する」「できる」「なさる」「くださる」등이 후행하는 것.

【例】「インプット」「愛着」「悪化」「悪戦苦闘」「一安心」

3.1.13 名詞-形容動詞語幹 (NAVS)

소위 말하는 形容動詞 어간으로 「な」앞에 나타나는 것.

【例】「健康」「安易」「駄目」「だめ」

3.1.14 名詞-ナイ形容詞語幹 (NAS)

助動詞「ない」직전에 나타나며 형용사적인 용법을 가진 것.

【例】「申し訳」「仕方」「とんでも」「違い」

3.1.15 名詞-数 (NNR)

한자로 표기된 숫자, 산용 숫자. 혹은 「何(回)」 「数(%)」.

【例】「0」「1」「2」「何」「数」「幾」

3.1.16 名詞-非自立-一般 (NDEG)

연체사, 「の(격조사)」, 활용어의 기본형에 접속해서 사용되는 것 중, 이하의 하위분류에 해당되지 않은 것. 소위 말하는 형식 명사를 포함함.

【例】 「あかつき」 「暁」 「かい」 「甲斐」 「気」 「きらい」

3.1.17 名詞-非自立-副詞可能 (NDEADP)

연체사, 「の(격조사)」, 활용어의 기본형에 접속해서 사용되는 것 중, 부사적인 기능을 가진 것.

【例】 「あいだ」 「間」 「あげく」 「挙げ句」 「あと」 「後」

3.1.18 名詞-非自立-助動詞語幹 (NDEAUS)

연체사, 「の(격조사)」, 활용어의 기본형에 접속해서 사용되는 것 중, 일본 학교 문법에서 助動詞로 취급되는 「ようだ」의 어간 부분.

【例】 「よう」 「やう」 「様(よう)」

3.1.19 名詞-非自立-形容動詞語幹 (NDEAVS)

연체사, 「の(격조사)」, 활용어의 기본형에 접속해서 사용되는 것 중, 「な(助動詞「だ」)의 체언 접속」과 접속이 가능한 것.

【例】 「みたい」 「ふう」

3.1.20 名詞-特殊-助動詞語幹 (NEAUS)

기본형에 접속하는 것으로 일본 학교문법에서 助動詞로 분류되어 있는 「そうだ(伝聞)」의 어간 부분.

【例】 「そう」

3.1.21 名詞-接尾-一般 (NSXG)

명사, 혹은 기타 품사의 어간 [ガル접속] 이나 [連用タイ접속] 에 접속해서 복합 명사를 형성하는 것 중, 하위분류에 해당되지 않은 것. 일반적으로 「接尾語」 보다 범위가 넓으며, 복합명사 뒤 부분의 요소로서 사용되는 것이 일반적인 것.

【例】 「おき」 「かた」 「方」 「甲斐 (がい)」 「がかり」 「ぎみ」

3.1.22 名詞-接尾-人名 (NSXN)

명사 중 특히 인명에 접속해서 명사를 형성하는 접미어.

【例】 「君」 「様」 「著」

3.1.23 名詞-接尾-地域 (NSXA)

명사 중 특히 지역명에 접속해서 명사를 형성하는 접미어.

【例】 「町」 「市」 「県」

3.1.24 名詞-接尾-サ変接続 (NSXCPV)

명사에 접속해서 명사를 형성하는 접미어 중, 「スル」 앞에 접속할 수 있는 것.

【例】 「化」 「視」 「分け」 「入り」 「落ち」 「買い」

3.1.25 名詞-接尾-助動詞語幹 (NSXAUS)

다른 품사의 연용형에 접속해서 일본 학교문법에서 助動詞의 어간으로 분류되어 있는 것. 「そうだ(樣態)」의 어간 부분.

【例】 「そう」

3.1.26 名詞-接尾-形容動詞語幹 (NSXAVS)

다른 명사나 활용어의 연용형에 접속하는 접미어로, 「だ」 ([助動詞特殊・タ]) 앞에 접속하는 것.

【例】 「的」 「げ」 「がち」

3.1.27 名詞-接尾-副詞可能 (NSXADP)

다른 명사에 접속하는 접미어로 부사적인 기능을 가진 것.

【例】「後(ご)」 「以後」 「以降」 「以前」 「前後」 「中」 「末」

3.1.28 名詞-接尾-助数詞 (NSXC)

숫자에 접속해서 명사를 형성하는 접미어. 일반적인 조수사보다 범위가 넓으며, 숫자에 접속하는 보통명사도 포함된다.

【例】「個」 「つ」 「本」 「冊」 「パーセント」 「c m」 「カ月」

3.1.29 名詞-接尾-特殊 (NSXE)

주로 용언에 접속하는 특수한 접미사로 새로이 정의한 것.

【例】「(樂し) さ」 「(考え) 方」

3.1.30 名詞-接統詞的 (NCJ)

주로 용언에 접속하는 특수한 접미사로 새로이 정의한 것 중에서 접속사적인 용법을 가지는 것.

【例】「(日本) 対 (アメリカ)」 「(3) 対 (5)」

3.1.31 名詞-動詞非自立的 (NVDE)

[助詞-接統助詞]의 「て」에 접속하는 것으로 의미적으로는 동사적인 것.

【例】「ごらん」 「ご覧」 「御覧」 「頂戴」

3.1.32 名詞-引用文字列 (NQ)

인용문에서 문자열로 사용되는 것.

3.2 접두사

3.2.1 接頭詞-名詞接統 (PXCN)

명사 (形容動詞 어간을 포함) 앞에 접속하는 접속어 중, 숫자에 접속하지 않은 것.

【例】「お (水)」 「某 (氏)」 「同 (社)」 「高 (品質)」

3.2.2 接頭詞-数接統 (PXCC)

명사 앞에 접속하는 것 중, 숫자에 접속하는 것.

【例】「約」 「およそ」 「毎時」

3.2.3 接頭詞-動詞接統 (PXCv)

동사의 명령형 혹은 [動詞連用形] + 「なる/なさる/くださる」 앞에 접속하는 접두어.

【例】「お (読みなさい)」 「お (座り)」

3.2.4 接頭詞-形容詞接統 (PXCA)

형용사 앞에 접속하는 접두어.

【例】「お (寒いですねえ)」 「バカ (でかい)」

3.3 동사

3.3.1 動詞-自立 (VIN)

3.3.2 動詞-非自立 (VDE)

3.3.3 動詞-接尾 (VSX)

3.4 형용사

3.4.1 形容詞-自立 (AIN)

3.4.2 形容詞-非自立 (ADE)

3.4.3 形容詞-接尾 (ASX)

3.5 부사

3.5.1 副詞-一般 (ADG)

반드시 뒤 부분에서 끊어지는 것. 연체수식이 불가능한 것.

【例】「あいかわらず」「多分」

3.5.2 副詞-助詞類接続 (ADCP)

뒤 부분에 「の」「は」「に」「な」「する」「だ」 등이 접속할 수 있는 부사.

【例】「こんなに」「そんなに」「あんなに」「なにか」

3.6 연체사

3.6.1 連体詞 (CS)

명사를 수식하는 형태만 가지는 것.

【例】「この」「その」「あの」「どの」「いわゆる」「こんな」

3.7 접속사

3.7.1 接続詞 (CJ)

독립적으로 나타나는 접속사.

【例】「が」「けれども」「そして」「じゃあ」「それどころか」

3.8 조사

3.8.1 助詞-格助詞-一般 (PJKG)

소위 말하는 격조사.

【例】「から」「が」「で」「と」「に」「へ」「より」「を」

3.8.2 助詞-格助詞-引用 (PJKQ)

명사, 인물의 대사, 회의의 결정 사항, 이유, 판결, 추측 표현 직후에 나타나는 「と」.

【例】「(～だ)と(述べた。)」

3.8.3. 助詞-格助詞-連語 (PJKP)

격조사와 동사와의 연어로 주로 격조사에 상당하는 기능을 가지는 것.

【例】「という」「といった」「とかいう」「として」「とともに」

3.8.4 助詞-接統助詞 (PJC)

조사 중 접속적인 용법을 가지는 것.

【例】「から」「からには」「が」「けれど」「けれども」「けど」

3.8.5 助詞-係助詞 (PRE)

소위 말하는 계조사.

【例】「こそ」「さえ」「しか」「すら」「は」「も」「ぞ」

3.8.6 助詞-副助詞 (PVI)

소위 말하는 부조사.

【例】「がてら」「かも」「くらい」「位」「ぐらい」「しも」

3.8.7 助詞-間投助詞 (PIJ)

문어적인 용법을 가지는 것.

【例】「(松島) や」

3.8.8 助詞-並立助詞 (PPA)

어떤 사항을 병립적인 의미로 나타낼 수 있는 용법을 가진 조사.

【例】「と」「たり」「だの」「だり」「とか」「なり」「や」

3.8.9 助詞-終助詞 (PEN)

소위 말하는 종조사.

【例】「かい」「かしら」「さ」「ぜ」「(だ)つけ<口語>」

3.8.10 助詞-副助詞／並立助詞／終助詞 (PAM)

조사「か」중에서 부조사, 병립조사, 종조사 어느 것인지 구별할 수 없는 것. 예를 들어 다음 a. b. c. 와 같은 것.

a. 「AかBか」형.

【例】「(国内で運用する) か、(海外で運用する) か (。)」

b. 부사절 중에 나타나는 것.

【例】「(幸いという) か、(死者はいなかった。)」

c. 「かのように」.

【例】「(何もなかった) か (のように振る舞った。)」

3.8.11 助詞-連体化 (PCS)

[名詞]에 접속해서 체언에 연결하는「の」.

3.8.12 助詞-副詞化 (PAD)

의음어, 의성어, 의태어, 혹은 이에 속하는 명사나 부사 직후에 나타나는「に」, 「と」.

【例】「に」 「と」

3.7.13 助詞-特殊 (PE)

조사 중에서 이상의 분류에 속하지 않은 것. 短歌나 俳句 등에 사용되는 조사 등을 포함함.

【例】「かな」 「けむ」 「(～したろう) に」 「(俺) ん (家)」

3.9 조동사

3.9.1 助動詞 (AU)

3.10 감동사

3.10.1 感動詞 (EX)

감동사. 인사말 등.

【例】「おはよう」「おはようございます」「こんにちは」

3.11 기호

3.11.1 記号-一般 (SYG)

이하의 분류 이외의 일반적인 기호.

【例】「○」「◎」「@」「\$」「〒」「→」「+」 등

3.11.2 記号-句点 (SYF)

소위 말하는 마침표.

【例】「。」「.」

3.11.3 記号-読点 (SYP)

소위 말하는 쉼표.

【例】「、」「,」

3.11.4 記号-空白 (SYB)

전각의 공백 문자. 화면상으로는 보이지 않음.

3.11.5 記号-アルファベット (SYA)

영어의 알파벳. 대문자, 소문자.

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

【例】「A」 「a」

3.11.6 記号-括弧開 (SYPO)

괄호 기호의 시작.

【例】「(」 「{」 「『」 「“」 「『」 「【」

3.11.7 記号-括弧閉 (SYPC)

괄호 기호의 종료.

【例】「)」 「}」 「』」 「”」 「』」 「】」

3.12 기타

3.12.1 その他-間投 (OAM)

[名詞-接尾]나 [助詞-終助詞]로 하기 어려운 것.

【例】「(だ) ア」

3.12.2 フィラー (F)

대화시의 맞장구나 삽입적인 음성적인 말.

【例】「あの」 「うんと」 「えと」

3.12.3 外国語 (FL)

영어, 한국어, 중국어 등 일본어가 아닌 모든 언어.

【例】「Watchtowe」 「안녕하세요」

3.12.4 ホームページ住所 (URL)

인터넷 홈페이지 주소를 나타냄.

【例】 「<http://www.korea.ac.kr/>」

여 백

제 4 장

북한 및 해외 한국어 말뭉치 개발

연구책임자 : 소 강 춘

여 백

제 4 장

북한 및 해외 한국어 말뭉치 개발

1. 주석 말뭉치 개발

1.1 대상 자료

금년도 대상 자료는 그동안 구축되었던 북한어 기초자료 말뭉치와 금년도에 구축한 북한 소학교 국어 교과서와 중국 소학교 조선어문 교과서 자료를 중심으로 선정했다.

1.2 개발 내용

1) 주석말뭉치 구축 목표 : 35만 어절

- 구축현황 : 북한 문화어 366,875 어절

2) 내용 :

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

2004년 북한 및 해외 한국어 형태소 분석 말뭉치 개발에서는 35만 어절 구축을 목표로 하고 있다. 금년도 과제는 북한 자료를 중심으로 기존에 구축되었던 자료를 고려하여 분야별 균형성을 확보하려 하였고, 특히 2000년에 개정한 북한의 소학교와 고등중학교 교과서를 원시말뭉치로 구축하고 있는데, 그 중에서 소학교의 국어 교과서를 주석하였다.

현재까지의 진행은 아래 표에 제시된 바와 같이 주석과 교정이 다 완료된 상태이다. 다만, ETI 헤더만 수정을 요한다.

[주석 말뭉치 구축 목록]

파 일 명	원 파 일 명	책이름	출판년도	어절수
▪ 7TNA02V1.HWP	5NA02V01.HWP	로동신문	2002	31,835
▪ 7TNB01F1.HWP	6NB01F02.HWP	조선어문	2001	50,428
▪ 7TNCA001.HWP	7NCA00A1.HWP	소학교 국어 1	2003	4,770
▪ 7TNCA002.HWP	7NCA00A2.HWP	인민학교 국어 2	2002	14,707
▪ 7TNCA003.HWP	7NCA00A3.HWP	인민학교 국어 3	2002	18,860
▪ 7TNCA004.HWP	7NCA00A4.HWP	인민학교 국어 4	2000	24,686
▪ 7TNE00B1.HWP	6NE00B01.HWP	붉은 산줄기	2000	109,659
▪ 7TNE00B2.HWP	4NE00001.HWP	터전	2000	111,930
				총 : 366,875 어절

2. 북한어 말뭉치 주석 지침(2004-수정본)

2.1. 기본 원칙

원칙적으로 '21세기 세종계획 국어 기초자료 구축 분과'의 '형태소 분석 말뭉치 구축 지침(2000-1)'을 따른다.

다만, 북한어에서 나타나는 특징을 보여주어야 할 경우에는 아래의 각 세분류의 목록 및 구축 지침에서 수정·보완하였다.

띄어쓰기는 북한어 자료에 충실히 따랐다.

어절의 분석은 조선말대사전에 표제항으로 등재되어 있는 경우에는 더 분석하지 않고 표제항의 형태를 기본으로 따랐다.

어절의 분석은 표면형을 토대로 하였으며, 원칙적으로 태그를 제거했을 때 원어절로 복원 가능하도록 하였다.

원칙적으로 단어형성에 참여하는 접사는 인정하지 않는다. 다만 북한어의 특성을 감안해서 접미사 '-적(的)', '-들', '-네', '-뿐', '-째' 만 인정한다.

축약형과 생략형은 원어절로 복원이 정확하게 가능하도록 하였다.

분석할 필요가 없는 특수한 기호가 한 어절을 이루는 경우 주석 말뭉치에 포함시키지 않았다.(따라서 그 기호에 해당하는 일련번호가 생략되었다.)

2.2. 어절 분석 표지(2004-수정본)

		기본품사		확장품사		
품사 단위	대단위	소단위	저단위	고단위		
	체언	명사	일반명사	NN		
			고유명사	NQ		
			의존명사	일반 단위	NB NC	
		대명사		NP		
		수사	NR			
	용언	일반동사	VV			
		일반형용사	VA			
		보조용언	VX			
		지정사	VC			
	관형사		MM			
	부사		MA			
	감탄사		IC			
교착소 단위	저단위		중단위	고단위		
	조사		J			
	어미	어말어미	종결어미	EF		
			연결어미	EC		
			전성어미	명사형	ETN	
		관형사형		ETM		
		선어말어미	시제	T		
	주체존대		H			
어기-접 사 단위	어기		WB			
	접사		접미사	XS		
기타 범주	문장부호		마침표	SF		
			쉼표,가운데점	SP		
			따옴표	SS		
			기타,논리수학	SW		
			명사추정	NF		
			용언추정	NV		
			분석불능	NA		

전성어미에서 인정하던 부동사형어미(EX)를 인정하지 않는다. 부동사형어미는

원칙적으로 한 어절 안에서 본용언과 보조용언을 연결하는 것으로 구분해서 사용했었다. 따라서 어절 내부의 연결어미(EC) 만으로도 이를 구분할 수가 있어 부동사형어미를 인정하지 않게 되었다.

문장부호에서 '기타(논리, 수학, 화폐단위 등)'를 위해 'SW'를 설정한다.

2.3. 표지별 분류 기준 및 세부 지침

2.3.1. 체언

체언은 조사가 연결되거나 단독으로 다른 체언이나 용언과의 문법적 관계를 나타내거나 문장 성분이 될 수 있다. 체언은 명사, 대명사, 수사 등 소단위로 나눈다.

2.3.1.1. 명사

명사는 일반명사, 특정한 사물을 지칭하는 고유명사, 관형어의 수식을 받고 형식적 의미를 지니는 의존명사로 세분한다.

(1) 일반명사 (NN)

'명사+명사'의 구조일 때, '명사+명사'가 조선말대사전에 표제항으로 실려 있는 경우에는 분석을 하지 않고 복합어로 처리하였다.

가극혁명	가극혁명[NN]
가내작업반	가내작업반[NN]
가을건이전투계획을	가을건이[NN]+전투계획[NN]+을[]
논흙갈이를	논흙갈이[NN]+를[]

(2) 고유명사 (NQ)

고유명사는 의미적으로 결정되는 표지이기 때문에 일관된 작업을 유지하기 위해서 그 범위를 명확히 한다. 다음과 같은 경우에만 고유명사로 인정한다.

인명: 성과 이름이나 호가 함께 쓰이면 하나의 분석 대상으로 하지만, 성이나 이름, 혹은 호와 함께 쓰인 관직명, '씨(氏), 공(公), 군(君), 양(孃), 웅(翁)' 등은 뜻에 관계없이 각기 다른 분석 대상이 된다.

김일성동지께서	김일성[NQ]+동지[NN]+께서[J]
김정일동지께서는	김정일[NQ]+동지[NN]+께서는[J]
김일성원수	김일성[NQ]+원수[NN]
림수경학생	림수경[NQ]+학생[NN]

지명 : 나라, 내륙, 바다, 강, 산, 산맥, 호수, 섬, 만, 계곡, 늪, 주 등의 이름이나 지역 이름은 고유명사로 한다. 지명명 뒤에 붙은 '도(道)', '군(郡)', '리(里)' 등도 지역명과 함께 전체가 고유명사가 된다.

남조선을	남조선[NQ]+을[J]
남천강도	남천강[NQ]+도[J]
마안산의	마안산[NQ]+의[J]

건축물이나 시설물 혹은 구조물의 이름, 창작물이나 작품의 이름은 고유명사로 간주한다.

김종태전기기관차연합기업소와	김종태전기기관차연합기업소[NQ]+와[J]
김일성저작선집	김일성저작선집[NQ]

(3) 의존명사

의존명사는 비단위성 의존명사와 단위성 의존명사로 나눈다.

비단위성 의존명사 (NB)

때문이다 때문[NB]+이[VC]+다[EF]
 만들어나갈줄 만들[VV]+어[EC]+나가[VX]+르[ETM]+줄[NB]
 맞물려줄것은 맞물리[VV]+어[EC]+주[VX]+르[ETM]+것[NB]+은[]
 못하신채 못하[VX]+시[H]+ㄴ[ETM]+채[NB]

단위성 의존명사 (NC)

두개	두[MM]+개[NC]
두번이나	두[MM]+번[NC]+이나[]
마흔두해전	마흔두[MM]+해[NC]+전[NN]
매+해	매[MM]+해[NC]
몇가지	몇[MM]+가지[NC]
몇달이라는	몇[MM]+달[NC]+이[VC]+라는[ETM]
몇되박을	몇[MM]+되박[NC]+을[]
몇백마디의	몇[MM]+백[MM]+마디[NC]+의[]
몇번의	몇[MM]+번[NC]+의[]

2.3.1.2 대명사 (NP)

"지시관형사+것(의존명사)"은 따로 분석하지 않고 통합형을 대명사로 간주한다.

자기	자기[NP]
이것은	이것[NP]+은[]
그것이	그것[NP]+이[]

‘어데’는 따로 분석하지 않고 대명사로 표지한다.

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

어데	어데[NP]]
어데서	어데[NP]+서[]
어데까지	어데[NP]+까지[]

‘뭐’는 ‘무엇’으로 복원하지 않는다.

뭐	뭐[NP]
뭐니	뭐[NP]+이[VC]+니[EF]
뭐봐	뭐[NP]+르[]+봐[VV]
뭐들해요	뭐[NP]+들[XS]+해[VV]+요[EC]

2.3.1.3. 수사 (NR)

서수사는 따로 분석하지 않고 통합형으로 처리한다.

세번째	세번째[NR]
넷을	넷[NR]+을[]
둘째도	둘째[NR]+도[]

2.3.2 용언

용언은 일반동사(VV)와 일반형용사(VA), 보조용언(VX)으로 구분한다.

‘용언+용언’ 구조의 경우, 조선말대사전에 표제항으로 등재된 경우에는 분석하지 않고 복합용언으로 처리한다.

돌이켜보게	돌이켜보[VV]+게[EC]
떨쳐나가도록	떨쳐나가[VV]+도록[EC]

밀고나가고있다	밀고나가[VV]+고[EC]+있[VX]+다[EF]
끓어넘치고있다	끓어넘치[VV]+고[EC]+있[VX]+다[EF]
끓어번지고있다	끓어번지[VV]+고[EF]+있[VA]+다[EF]

‘명사+용언’ 구조인 경우, 조선말대사전에 표제항으로 등재된 경우에는 분석하지 않고 복합용언으로 처리한다.

기운차게	기운차[VA]+게[EC]
보람차고	보람차[VA]+고[EC]
김매는	김매[VV]+는[ETM]
나래치게	나래치[VV]+게[EC]

‘부사+용언’의 구조인 경우, 조선말대사전에 표제항으로 등재된 경우에는 분석하지 않고 복합용언으로 처리한다.

잘되도록	잘되[VV]+도록[EF]
같이하고	같이하[VV]+고[EF]
그만두거나	그만두[VV]+거나[EF]
깊이하고	깊이하[VV]+고[EF]
끄떡하지	끄떡하[VV]+지[EC]

보조용언은 보조동사와 보조형용사로 구분하지 않는다. 이는 보조동사와 보조형용사가 본동사와 본형용사에 기인하는 경우가 많고, 동일 형태가 보조동사로도 보조형용사로도 분석되기 때문이다.

가까와오도록	가깝[VA]+아[EC]+오[VX]+도록[EF]
가다듬고있다	가다듬[VV]+고[EC]+있[VX]+다[EF]
굳어지게	굳[VA]+어[EC]+지[VX]+게[EC]
리해해준다니	리해하[VV]+여[EC]+주[VX]+다[EF]+니[EC]

불규칙용언이 활용한 경우에는 기본형으로 분석한다.

날라다줄것을	나르[VV]+아다[EF]+주[VV]+르[ETM]+것[NB]+을[J]
노는	놀[VV]+는[ETM]
놀라운	놀랍[VA]+은[ETM]
들으면서	듣[VV]+으면서[EF]
무어진	못[VV]+어[EC]+지[VX]+ㄴ[ETM]
이어	잇[VV]+어[EF]
이으시면서	잇[VV]+으시[H]+면서[EF]

지정사[VC] 기능을 하는 아래의 것들은 통합해서 지정사로 분석한다.

리춘금이에요	리춘금[NQ]+이예[VC]+요[EC]
별고웨다	별고[NN]+웨[VC]+다[EF]
글썸말이웨다	글썸[IC]+말[NN]+이웨[VC]+다[EF]
나누만요	나[NP]+누[VC]+만[J]+요[EF]
몸수색허던눔이우다	몸수색허[VV]+던[EC]+눔[NN]+이우[VC]+다[EF]

2000년 이후에 적용되고 있는 띄어쓰기 규정에 따르면 기존에 '용언+용언', '명사+용언', '부사+용언', '본용언+보조용언' 등의 형태로 실현되던 예들을 모두 띄어 쓰고 있다. 따라서 조선말대사전이나 남한의 표준국어대사전에서도 붙여쓰고 있는 어형을 띄어쓰고 있어 선행 어형의 품사를 밝힐 수 없는 경우도 있다.

도저히 선행 어형의 품사를 밝힐 수 없는 경우를 제외하고는 띄어쓰기를 반영한다. 단, 이 경우에는 선행하는 어사의 품사를 용언으로 볼 수 있는 경우에 후행하는 용언은 보조용언(VX)으로 처리한다.

내 딴다	내[VV]+어[EC]+딛[VX]+다[EF]
올려 다 보다	올리[VV]+어[EC]+다[EF]+보[VX]+다[EF]

들여다 보다 들이[VV]+어다[EC]+보[VX]+다[EF]

눈여겨 보다 눈[NN]+여기[VV]+어[EC]+보[VV]+다[EF]

그러나 선행하는 어형이 '명사, 부사'일 경우에는 본동사처럼 처리하는 것이 원칙일 것이다. 그러나 2000년도 이전 자료와의 기계적인 처리를 위해서 보조용언으로 분석한다. 이는 프로그램적으로 조선말대사전으로 복원이 가능하게 하기 위한 조치이다.

청승 맞다 청승[NN] 맞[VX]+다[EF]

열 떠다 열[NN] 떠[VX]+다[EF]

가로 막다 가로[NN] 막[VX]+다[EF]

가득 차다 가득[MA] 차[VX]+다[EF]

새빨개 지다 새빨개[MA] 지[VX]+다[EF]

아로 새기다 아로[MA] 새기[VX]+다[EF]

2.3.3. 수식언

2.3.3.1. 부사 (MA)

부사는 크게 성분 부사와 문장 부사로 나누고, 문장 부사는 다시 양태 부사와 접속 부사로 나뉜다. 여기서는 이들을 세분하지 않고 모두 부사로 통일한다.

'그래서'는 접속부사로 인정하지만, '그리하여'는 접속부사로 인정하지 않는다.

그래서 그래서[MA]

그리하여 그리하[VV]+여[EF]

'용언어간+부사파생접미사'구조를 보이는 부사는 접사를 따로 인정하지 않으므

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

로 이를 분석하지 않고 통합형태를 부사로 인정한다.

감명깊이	감명깊이[MA]
------	----------

기존에 한 단어로 붙여쓰던 부사가 두 단어로 나뉘는 예들이 있다. 이 경우 앞 어사의 경우는 다른 품사로 기술할 수 있지만, 후행하는 어사는 부사[MA]로 인정한다.

배불리[MA]	배[NN]	불리[MA]
---------	-------	--------

2.3.3.2. 관형사 (MM)

관형사는 지시관형사, 수관형사, 성상관형사로 세분될 수 있으나, 본 표지에서는 세분하지 않고 관형사로만 표시한다.

두차례의	두[MM]+차례[NN]+의[J]
그날로	그[MM]+날[NN]+로[J]
모든분야에서	모든[MM]+분야[NN]+에서[J]
새세계를	새[MM]+세계[NN]+를[J]

"이렇게 할"의 준말인 '이럴'은 더 이상 분석하지 않고 통합형을 관형사로 분석한다.

이럴	이럴[MM]
----	--------

'그러하+ㄴ'의 준말인 '그런'은 더 이상 분석하지 않고 통합형을 관형사로 인정한다.

그런것만큼	그런[MM]+것[NB]+만큼[J]
-------	--------------------

‘어떠하+ㄴ’의 준말인 ‘어떤’은 더 이상 분석하지 않고 통합형을 관형사로 인정한다.

어떤것을	어떤[MM]+것[NB]+을[J]
------	-------------------

2.3.4. 관계언

조사는 격조사와 보조사로 세분될 수 있다. 본 표지에서는 격조사와 보조사의 표지를 따로 마련하지 않고 조사 표지(J)로 통일한다.

생산량을	생산량[NN]+을[J]
각방으로	각방[NN]+으로[J]
각오밑에	각오[NN]+밑[NN]+에[J]
생산량만	생산량[NN]+만[J]
거두는것도	거두[VV]+는[ETM]+것[NB]+도[J]

‘격조사+격조사’는 따로 분석하지 않고 통합형태를 단일조사로 처리한다.

비관적으로가	비관[NN]+적[XS]+으로가[J]
사상분야에서의	사상[NN]+분야[NN]+에서의[J]
말로써가	말[NN]+으로써가[J]

‘격조사+보조사’의 경우는 따로 분석하지 않고 통합형태를 조사로 처리한다.

북으로부터	북[NQ]+으로부터[J]
북에까지	북[NQ]+에까지[J]
대상으로만	대상[NN]+으로만[J]

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

다함으로써만	다함[VV]+ㅁ[ETN]+으로써만[J]
단위들에서도	단위[NN]+들[XS]+에서도[J]
다음해에도	다음해[NN]+에도[J]

‘보조사+격조사’의 경우는 따로 분석하지 않고 통합형태를 조사로 처리한다.

내용만이	내용[NN]+만이[J]
당만을	당[NN]+만을[J]
때만이	때[NN]+만이[J]
모통이마다에	모통이[NN]+마다에[J]

‘보조사+보조사’의 경우, 분석하지 않고 통합형태를 조사를 처리한다.

가운데서도	가운데[NN]+서도[J]
문제까지도	문제[NN]+까지도[J]
4월중순까지는4월[NN]+중순[NN]+까지는[J]	

‘명사+보조사’의 경우, 보조사 앞에 생략된 격조사를 따로 복원하지 않고 보조사에 조사 표지를 넣는다.

준비하는데서	준비하[VV]+는[ETM]+데[NB]+서[J]
가운데서	가운데[NN]+서[J]

2.3.5. 의존형태

2.3.5.1. 어말어미

어말어미는 종결어미, 연결어미, 전성어미로 구분한다. 전성어미는 다시 명사형 전성어미와 관형사형 전성어미로 구분한다.

어말어미는 환경에 따라 이형태로 실현되는데, 이때 기본형태로 분석하지 않고 표면형에 따라 분석한다.

‘어말어미+격조사’는 따로 분석하여 표지를 준다.

다하겠는가에	다하[VV]+겠[T]+는가[EF]+에[]
되었는가를	되[VX]+였[T]+는가[EF]+를[]
떨쳐나서는가에	떨치[VV]+어[EC]+나서[VX]+는가[EF]+에[]
마련된것인가를	마련되[VV]+ㄴ[ETM]+것[NB]+이[VC]+ㄴ가[EF]+를[]
의존해서가	의존하[VV]+여서[EF]+가[]
의해서가	의하[VV]+여서[EF]+가[]
있겠는가에	있[VA]+겠[T]+는가[EF]+에[]

‘어말어미+보조사’도 따로 분석하지 표지를 준다.

높아져야만	높아지[VV]+어야[EC]+만[]
놓고도	놓[VV]+고[EC]+도[]
담고있으면서도	담[VV]+고[EC]+있[VX]+으면서[EF]+도[]
당면해서는	당면하[VV]+여서[EC]+는[]

‘어말어미+관형사형 어미’는 따로 분석하여 표지를 준다.

계속하려는데로부터	계속하[VV]+려[EF]+는[ETM]+데[NB]+으로부터[]
과업이라는것을	과업[NN]+이[VC]+라[EF]+는[ETM]+것[NB]+을[]
관철하자는데	관철하[VV]+자[EF]+는[ETM]+데[NB]
따라배운다는것은	따라배우[VV]+ㄴ다[EF]+는[ETM]+것[NB]+은[]

(1) 종결어미 (EF)

용언의 어간이나 선어말 어미 뒤에 연결되어 용언의 형식을 완성시키는 어미로 문장을 끝맺는 역할을 한다.

이끌어주는것이다 이끌[VV]+어[EC]+주[VX]+는[ETM]+것[NB]+이[VC]+다[EF]

이끌어오시였다 이끌[VV]+어[EC]+오[VX]+시[H]+였[T]+다[EF]

(2) 연결어미 (EC)

이형태에 대한 기본형을 제시하지 않고 표면형을 토대로 분석한다.

각성되어 각성되[VV]+여[EC]

가하여 가하[VV]+여[EC]

담아 담[VV]+아[EC]

결어 결[VV]+어[EC]

‘연결어미+보조사’의 경우는, 따로 분석하지 않고 통합형태를 연결어미로 처리한다.

무겁고도 무겁[VA]+고도[EC]

살아도 살[VV]+아도[EC]

짜고들어야 짜고들[VV]+어야[EC]

밀어주어야만 밀[VV]+어[EC]+주[VX]+어야만[EC]

담고있으면서도 담[VV]+고[EC]+있[VX]+으면서도[EC]

‘종결어미+연결어미’는 각각 분석한다.

가지고있다면 가지[VV]+고[EC]+있[VX]+다[EF]+면[EC]

가한다고 가하[VV]+ㄴ다+[EF]+고[EC]

(3) 전성어미

전성어미는 명사형 전성어미(ETN)과 관형사형 전성어미(ETM)으로만 구분한다.

움직임도	움직이[VV]+ㅁ[ETN]+도[J]
이끄심으로	이끌[VV]+시[H]+ㅁ[ETN]+으로[J]
유지하기	유지하[VV]+기[ETN]
이끌어주시기에	이끌[VV]+어[EC]+주[VX]+시[H]+기[ETN]+에[J]
궁지로움이여	궁지롭[VA]+음[ETM]+이여[EF]

어간 말음이 'ㄹ'인 용언에 관형사형 어미 '-ㄹ'이 결합된 경우에는 표면형에 관형사형 어미가 나타나지 않으나 분석하여 표지한다.

머물수 머물[VV]+ㄹ[ETM]+수[NB]

2.3.5.2. 선어말어미

선어말어미는 시제(T)를 나타내는 '-었-', '-겠-', '-리'와 주체존대선어말어미(H) '-시-'만을 인정한다.

가리키시는	가리키[VV]+시[H]+는[ETM]
감수하시며	감수하[VV]+시[H]+며[EF]
같으신	같[VA]+으시[H]+ㄴ[ETM]
가로놓여있었다	가로놓이[VV]+어[EC]+있[VX]+었[T]+다[EF]
갈라주었으며	가르[VV]+아[EC]+주[VX]+었[T]+으며[EC]

과거시제 선어말어미가 선행어간에 따라 형태를 달리하는 경우에는 기본형으로

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

설정하지 않고 표면형에 따라 분석한다.

감행되었으며	감행되[VV]+였[T]+으며[EC]
강대하였으며	강대하[VV]+였[T]+으며[EC]
교시하시였다	교시하[VV]+시[H]+였[T]+다[EF]
가다듬었으며	가다듬[VV]+였[T]+으며[EC]
내놓았다	내놓[VV]+았[T]+다[EF]

2.3.6. 어기-접사

2.3.6.1. 어기 (WB)

2000년에 개정된 북한의 띄어쓰기 규정에 따라 기존에 하나의 단어로 표시되던 용언의 어간들이 분리되어 표기되었다. 이를 표기하기 위해서 어기를 설정한다.

빨개 진	빨개[WB]+지[VX]+ㄴ[ETM]
들어 왔습니다	들어[WB]+오[VX]+았[T]+습니다[EF]

2.3.6.2 접사 (XS)

접사는 접미사와 접두사로 나눌 수 있는데, 본 표지에서 접사를 따로 표지하지 않는다. 다만 북한어의 특성을 감안하여 접미사 ‘-적(的)’, ‘-들’, ‘-네’, ‘-뿐’, ‘-째’는 표지한다.

감동적인	감동[NN]+적[XS]+이[VC]+ㄴ[ETM]
강령적과업을	강령[NN]+적[XS]+과업[NN]+을[J]
거인적풍모속에는	거인[NN]+적[XS]+풍모[NN]+속[NN]+에는[J]

결정적국면을	결정[NN]+적[XS]+국면[NN]+을[J]
농사일들을	농사일[NN]+들[XS]+을[J]
경공업공장들을	경공업[NN]+공장[NN]+들[XS]+을[J]
경제일군들이	경제일군[NN]+들[XS]+이[J]

‘어근+-적(的)’ 형태는 통합형태를 관형사로 인정하지 않고 ‘어근’과 접미사 ‘-적(的)’으로 분석하여 표지한다.

다각적인	다각[NN]+적[XS]+이[VC]+ㄴ[ETM]
구체적으로	구체[NN]+적[XS]+으로[J]

가족을 나타내는 접미사 ‘-네’, 한정을 나타내는 접미사 ‘-뿐’, 순서를 나타내는 접미사 ‘-째’를 표지한다.

범나비네가	범나비[NN]+네[XS]+가[J]
생각뿐이었어	생각[NN]+뿐[XS]+이[VC]+였[T]+어[EF]
말뿐아니라	말[NN]+뿐[XS]+아니[VC]+라[EC]
두번째	두[MM]+번[NC]+째[XS]

2.3.7. 생략형, 축약형, 융합형

2.3.7.1. 생략형

생략형은 원칙적으로 복원하지 않는다.

‘연결어미+보조사’의 구조인데, 연결어미가 표면형으로 드러나지 않을 경우에는 따로 복원하지 않고 ‘보조사’에 어미 표지를 한다.

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

이끌고나아가야	이끌[VV]+고[EF]+나아가[VV]+야[EC]
돌아서서	돌아서[VV]+서[EC]
가도	가[VV]+도[EC]
가나	가[VV]+나[EC]

‘용언어간+연결어미’의 구조인데, 용언어간의 말음과 연결어미가 같은 형태여서 연결어미가 표면형으로 드러나지 않을 경우에는 연결어미를 복원하여 표지한다.

나가	나가[VV]+아[EC]
내려가	내려가[VV]+아[EC]
떠나	떠나[VV]+아[EC]

‘명사+하+지’의 구조에서 ‘하’가 완전히 생략된 경우에는 다음과 같이 분석한다.

생각지	생각[NN]+지[EC]
-----	--------------

2.3.7.2. 축약형

‘용언어간’의 끝음절과 어미의 첫음절이 결합하여 한 음절로 축약된 경우에 축약형은 분석하여 표지한다.

높여	높이[VV]+어[EC]
해서	하[VV]+여서[EC]
내세워야	내세우[VV]+어야[EC]
가르쳐주고	가르치[VV]+어[EC]+주[VX]+고[EC]
단합시켜	단합시키[VV]+어[EC]
높아졌다	높[VA]+아[EC]+지[VX]+았[T]+다[EF]
전진시켜왔다	전진시키[VV]+어[EC]+오[VX]+았[T]+다[EF]
쌔워먹여야겠는데	쌔우[VV]+어[EC]+먹이[VX]+어야[EC]+겠[T]+는데[EC]

다음의 축약형들은 어간에 'ㅎ'을 설정한다.

심상치	심상ㅎ[VA]+지[EC]
온당치	온당ㅎ[VA]+지[EC]

다음의 예들을 구분하여 분석한다.

선도하여	선도하[VV]+여[EC]
선도해	선도하[VV]+어[EC]

2.3.7.3. 융합형

융합형은 그대로 단독 어절로 인정한다.

그래야	그래야[MA]
그랬단	그래[VV]+쓰[T]+다[EF]+ㄴ[ETM]
그랬됐군요	그래[VV]+쓰[T]+됐[VV]+군요[EF]
그러기	그러[VA]+기[ETN]
그러자면	그러자면[MA]
그러지	그러[VA]+지[EC]

2.3.7.4. 기타 북한 표현어

북한어의 특이형으로 볼 수 있는 예들은 그 원형대로 인정한다.

그러쉬우	그러[VV]+쉬[H]+우[EC]
몸수색허던눔이우다	몸수색허[VV]+던[EC]+눔[NN]+이우[VC]+다[EF]
별고웨다	별고[NN]+웨[VC]+다[EF]

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

동무우다	동무[NN]+우[VC]+다[EF]
그랬됐군요	그래[VV]+쓰[T]+됐[VV]+군요[EF]
느들허구	느[NP]+들[XS]+허구[J]
돌려부내줍세	돌려부내[VV]+주[VX]+ㅁ세[J]
늣었나부다	늣[VA]+엇[T]+나[EC]+부[VX]+다[EF]
무스거	무스거[MA]

제 5 장

역사자료 말뭉치 개발

연구책임자 : 임 용 기

여 백

제 5 장

역사자료 말뭉치 개발

1. 품사 주석 말뭉치

1.1 목표

3단계 1차년도인 2004년 역사자료 주석 말뭉치에서는 총 5만 어절 규모의 원시 말뭉치를 선정하여 형태정보 주석을 더하는 것을 목표로 한다.

1.2 대상문헌 및 파일명

1.2.1 대상문헌

<월인석보>, <석보상절>, <독립신문>

1.2.2 주석 결과물 총 어절수

<월인석보 권21, 23, 25>.....	34,842 어절
<석보상절 권6, 9, 13>.....	11954 어절
<독립신문>.....	3,685 어절

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

소계 50,481 어절

1.2.3 대상 원본 파일(2단계 1차년 파일명, 원시말뭉치:파일명)

<월인석보>

6PT_0001.HWP, 원시말뭉치:P5CD0020.HWP 34842 어절

6PT_0002.HWP

6PT_0003.HWP

<석보상절>

6PT_0004.HWP, 원시말뭉치:P5CD0007.HWP 11954 어절

6PT_0005.HWP

6PT_0006.HWP

<독립신문>

6PT_0007.HWP, 원시말뭉치:P9BC0017.HWP 3,685 어절

소계 : 50,481 어절

이상을 표로 정리하면 다음과 같다.

파일명	원시파일명	제목	언어	장르	연도	어절수	저작권
6PT_0001	P5CD0020.HWP	월인석보	한국어	종교서류	1447	13,583	없음
6PT_0002	P5CD0020.HWP	월인석보	한국어	종교서류	1447	9,298	없음
6PT_0003	P5CD0020.HWP	월인석보	한국어	종교서류	1447	11,970	없음
6PT_0004	P5CD0007.HWP	석보상절	한국어	종교서류	1459	4,013	없음
6PT_0005	P5CD0007.HWP	석보상절	한국어	종교서류	1459	3,122	없음
6PT_0006	P5CD0007.HWP	석보상절	한국어	종교서류	1459	4,819	없음
6PT_0007	P9BC0017.HWP	독립신문	한국어	산문기록류	1889	3,685	없음
합계						50,481	

월인석보, 석보상절의 주석 대상이 되는 파일은 디지털 한글 박물관에서 구축한 것을 대상으로 하였다. 그 이유는 방점과 한자음까지 모두 입력이 되어 있어 좀더 완벽한 주석 말뭉치를 구축할 수 있기 때문이다.

2. 역사자료 말뭉치 주석 체계

주석체계의 확정을 위해서는 다음과 같은 세 가지가 결정되어야 한다(리치 Leech 1997).

- 첫째, 문법정보 표지.
- 둘째, 주석 대상 단위.
- 셋째, 표지 부여 기준.

문법 정보 표지란 말뭉치에 대한 문법 정보 주석을 위한 표지의 집합 즉 태그 집합tag set을 뜻하는 것으로, 주석 대상이 되는 언어 단위의 해당 언어 안에서의 계열적 분석paradigmatic analysis과 통합적 분석syntagmatic analysis의 결과를 반영한다. 주석 대상 단위의 확정이란 문법정보 주석을 위해 결정된 표지tag를 배정할 단위를 구체적으로 결정하는 것을 의미하는 바, 일반적으로 문법정보 주석에서는 어절 분석의 결과가 그 대상 단위가 되지만, 어절 단위의 분석만으로는 해결되지 않는 문제를 안고 있는 것들도 있다. 표지 부여 기준은 구체적인 말뭉치 분석 과정에서 하나의 단위에 여러 가지 표지를 부여할 수 있을 때, 구체적으로 어떤 표지를 부여할 것인가를 결정하는 것을 의미한다.

또한 이러한 주석 체계 외에, 주석 결과물의 형태에 대한 원칙도 결정할 필요가 있다. 금년도의 연구에서는 현대 국어 주석 말뭉치 구축의 원칙을 그대로 준용하여 수직 말뭉치의 형태로 만들었지만, 실제 수직 말뭉치 형태는 확장성과 자료의 가독성 그리고 검색의 편의성이라는 측면에서 그대로 유지할 수 없는 것으로 판단된다. 웹에의 게시나 DB화 등을 고려할 때에 XML 형식의 문서로 바꾸는 것이 바람직하다고 판단된다. 그러나 여기서는 자세히 다루지 않는다.

여기의 말뭉치 주석 체계는 어절 단위를 대상으로 한 역사자료 말뭉치의 문법정보 주석에 적용되는 일련의 약속들을 의미한다. 문법정보 주석이란 어절 단위에 문법적 정보를 더하는 작업을 가리키는데, 역사자료를 대상으로 한 문법정보 주석 체계는 현대국어에 대한 문법정보 주석체계와는 달라질 수 밖에 없다. 여기서는 역

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

사자료 말뭉치를 대상으로 한 주석체계를 결정하는 데에 고려해야 할 문제들을 정리하기로 한다.

2.1 문법 정보 주석 태그 집합 선정

2.1.1 문법 정보 주석 태그 집합 결정 기준

- 21세기 세종 계획의 현대국어 어절 분석 정보 표지 표준안에 준하는 것을 원칙으로 하되, 역사 자료의 특성을 반영할 수 있도록 한다.
- 단어구성정보 관련 표지는 1차 주석 말뭉치 구축에서는 반영하지 않는다. 따라서 의존 형태 중 접미사, 접두사 관련 태그는 반영하지 않는다.
- 방점의 경우, 숫자로 표시한다.

2.1.2 형태 정보 주석 태그 집합

대분류(7개)	소분류(18개)	세부분류	태그	
(1) 체언	명사	일반명사	NNG	1
		고유명사	NNP	2
		의존명사	NNB	3
	대명사		NP	4
	수사		NR	5
(2) 용언	동사		VV	6
	형용사		VA	7
	보조용언		VX	8
	지정사	긍정지정사	VCP	9
		부정지정사	VCN	10
(3) 수식언	관형사		MM	11
	부사	일반부사	MAG	12
		접속부사	MAJ	13
(4) 독립언	감탄사		IC	14
(5) 관계언	격조사	주격조사	JKS	15
		관형격조사	JKG	16
		목적격조사	JKO	17
		부사격조사	JKB	18
		호격조사	JKV	19
		인용격조사	JKQ	20
	보조사		JX	21
	접속조사		JC	22
(6) 의존형태	어미	선어말어미	EP	23
		종결어미	EF	24
		연결어미	EC	25
		명사형전성어미	ETN	26
		관형형전성어미	ETM	27
(7) 기호	방점	평성	0	28
		거성	1	29
		상성	2	30
	첨표		SP	31
	괄호표		SS	32
	한자		SH	33
	기타기호		SW	34

2.2 역사자료 말뭉치 주석 대상 단위

2.2.1 어절 단위의 확정과 분석 : 띄어쓰기와 기타 기호 문제

- 원시 말뭉치의 구축에서 띄어쓰기는 원본에 있는 대로 하는 것을 원칙으로 하되, 띄어쓰기가 전혀 되어 있지 않은 자료는 현대국어 띄어쓰기에 따라 띄어 썼다. 따라서 주석 작업도 원시 말뭉치의 어절 단위에 따라 수직 주석을 단다. 아주 드물게 원시 말뭉치의 어절 구별에 이상이 있을 때에는 원본의 띄어쓰기나 특성에 맞게 정정하여 주석을 단다.

- 원문에 있는 특수 기호에 대해서도 주석을 단다.

020590	●독립협회와	●	SW	독립협회NNP	와	JC
020348	○종일품	○	SW	종일품	NNG	

- 3단계 1차년도 역사 자료는 띄어쓰기를 기준으로 두 가지 유형의 텍스트로 분류할 수 있다.

i) 띄어쓰기가 전혀 되어 있지 않은 텍스트

: <월인석보>, <석보상절>

ii) 현대국어 어법과 완전히 일치하지는 않지만 어절별로 띄어쓰기가 되어 있는 경우

: <독립신문>

☞ 이미 띄어쓰기가 되어 있는 경우, 주석 작업 시에 현대국어 어법에 맞는 띄어쓰기를 임의로 하면, 결과적으로 원문으로 회복이 불가능하게 된다. 따라서 우선 원문의 띄어쓰기를 그대로 반영하여 주석 작업을 한다. 이 작업이 이루어지고 다시 2차 가공을 하게 된다면 현대국어 어법에 맞는 띄어쓰기를 따라 주석을 달아야 할

것이다.

- 3단계 1차년도 역사 자료는 방점을 기준으로 두 가지 유형의 텍스트로 분류할 수 있다.

i) 방점 자료 : <월인석보, 석보상절>

☞ 방점 자료의 방점 태그는 숫자로 한다. 평성은 0, 거성은 1, 상성은 2. 주석 부착 순서는 원어절, 방점, 어간, 선어말어미, 어미/조사의 순이다.

000075	잇느·니	001	잇	VV	느	EP	니
	EC						
034634	丹단霞향·익	001	단향	丹霞	NNG	익	JKB

ii) 무방점 자료 : <독립신문>

☞ 무방점 자료의 주석부착 순서는 방점 자료의 순서에 준하며, 다만 방점 정보가 없다.

000060	교우여든	교우	NNG	이	VCP	ㄷ	EC
020540	업스오니	업	VA	스오	EP	니	EC

2.3 주석 지침

2.3.1 체언

- 일반명사→NNG

i) 일반명사와 한자 : (본문일 때)어린 百姓이→NNG
(협주일 때) 國은 나라히라→SH

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

ii) 특수어간 교체 : 나모와 나모 NNG 와 JC
 남기 남 NNG | JKS

• 고유명사→NNP

*迦강文문 · 이 001 강문 迦文 NNP 이 JKS

• 의존명사→NNB

- i) 보편성 : 바, 것, 녀, 득, 던, 분, 썬, 스, 앓, 이, 적, 히
- ii) 주어성 : 디, 슷
- iii) 서술성 : 썸
- iv) 부사성 : 득, 디, ㄹ장, 거기, 게, 그에, 괴, 다비, 듯, 동, 만, 손디, 양, 자히
- v) 단위성 : 디위, 마리, 치, 리(里), 설(歲), 번, 날/낫, 량, 볼....
 썸, 말, 되, 흠, 자ㅎ, 히, 돌, 날, 돈, 사롬, 즈릭(柄)

• 대명사→NP

- i) 대명사 : 나(우리), 너(너희), 그디~그디~그디, 더, 누, 아모, 저(저희), 즈가~
 즈개
- ii) 지시 대명사 : 이, 그, 더, 어느, 므속~므슴~므스것, 아모것, 이어괴, 그어
 괴~그에~게, 덩어괴~더괴~더에~데, 어딴, 아모더, 어드메, 아모드.

• 수사→NR

- i) 둘, 셋....아흔, 온, 즈믄, 몇, 여럿
- ii) 一, 二, 三....九十, 百, 千, 萬, 億
- iii) ㅎ나차히, 둘차히....닐흔차히, 여든차히-분석하지 않음.
- iv) 第一, 第二....

2.3.2 용언

- 동사→VV

- i) 보다, 가다…,
- ii) 조치다, 머기다…,
- iii) 졌다, 뵈다, 굴다, 탄다, 버히다, 홀다…,
- iv) 잇/이시다

- 형용사→VA

- i) 붉다, 희다, 길다, 높다, 모딜다, 아롭답다, 조하다, 곧하다…
- ii) 이러하다, 그러하다, 더러하다, 엇더하다, 아므라하다…
- iii) 없다

- 보조용언→VX

- i) 진행 : 가다, 오다, 잇다
- ii) 종결 : 나다, 내다, 브리다
- iii) 시행 : 보다
- iv) 보유 : 두다, 놓다
- v) 피동 : 드외다, 디다
- vi) 사동 : (-게/긔) 하다
- vii) 부정 : 아니하다, 못하다, 말다
- viii) 당위 : (-어샤) 하다
- ix) 희망 : (-고져) 식브다
- x) 추측 : (-ㄴ/ㄹ가) 식브다
- xi) 상태 : (-어) 잇다/겨시다

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

- 지정사→VC

i) 긍정지정사→VCP : -이라, -ㅣ라

ii) 부정지정사→VCN : 아니라

2.3.3 수식언

- 관형사→MM

i) 외, 상, 새

ii) 므슷, 어누

iii) 이, 그, 더, 어느, 므슴

iv) 혼, 두, 세/석/서, 네/넉/너, 다섯/닷, 여섯/엣, 닐굽, 여덥, 아홉, 열……스믈 /스므……온, 즈믄/첿/첸, 몇, 현, 여러

- 부사→MA

i) 일반부사→MAG : ㄹ, 기리, 노피, 다뭇, 어루, 절로, ㄱ장, 그르, 뭇; 설설, 썸썸, 구물구물; 이리, 그리, 더리, 어드리, 엇데, 아믄리; 아니, 몬; 모로매, 모더, 반드기.

ii) 접속부사→MAJ : 그러나, 그러면, 그럴씨, 그런드로, 이런드로, 이럴씨, 헛다가, 오히려, 출하리, 흥물며, 도로혀, 비록, 쑤, 밋

2.3.4 독립언(감탄사)→IC

i) 감정 : 익, 아으, 오호라, 오오, 오히라

ii) 의지 : 앵, 아소, 응

2.3.5 관계언

- 주격조사→JKS
- 목적격조사→JKO
- 관형격조사→JKG
- 부사격조사→JKB
- 호격조사→JKV
- 인용격조사→JKQ
- 접속조사→JC
- 보조사→JX

2.3.6 의존형태

- 선어말어미→EP(선어말어미의 통합형은 더 이상 분석하지 않음)
 - i) 습/줍/습, 시/샤, 이
 - ii) ㄴ, 더, 리
 - iii) 오/우
 - iv) 거/어
 - v) 뒛/웃/ㅅ
 - vi) 니
- 어말어미→EF
 - i) 다, 나라
 - ii) 녀, ㄴ다, ㄴ가, 뇨
 - iii) 라
 - iv) 저
- 연결어미→EC

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

- i) 대등 : 고, 며, 나, 건마룬, 거니라, 거나
- ii) 종속 : 니, 오디, 르썸, 판디, 거든, 거늘, 고저
- iii) 보조 : 어/아, 괴/게, 디, 고

- 명사형 전성어미→ETN

- i) 음/움, 기, 디

- 관형사형 전성어미→ETM

- i) ㄴ, ㄷ

2.3.7 기호

- 쉽표(,)→SP
- 괄호표(())→SS
- 한자→SH
- 기타기호(○, ●)→SW

2.4 주석 작업시 문제점

2.4.1 작업 외적인 문제

- 원시말뭉치의 통일성과 관련된 문제
 - i) 띄어쓰기 조절.
 - ii) 오자 및 방점 오류기 검토.
 - iii) 쪽페이지 조절.
 - iv) 2차 가공 작업분-어간형 통일(깁/끓→끓), 선어말어미 분석, 접사 분석 등은 작업하지 않는다.

2.4.2 작업 내적인 문제

- 분석은 그 결과가 원어간형과 동일하도록 함. 즉 복원했을 때(원어간형이 그 대로 나와야 함)

000070	닐 · 오 · 더	011	<u>닐</u>	<u>VV</u>	오 더	EC
000150	깁 · 고	01	<u>깁</u>	<u>VA</u>	고	EC

☞ 어간을 '닐-, 깁-'으로 잡지 않는다.

:업 · 스니 · 라	2101	<u>업</u>	<u>VA</u>	—니라	EF
:업다	20	<u>업</u>	<u>VA</u>	다	EF

☞ 동일 어간일지라도 표기형에 따라 어간의 표기를 결정한다.

- 한자/한자음 병기 어간 처리

034631	日 · 齡月 · 齡 · 이	111	<u>齡齡</u>	<u>日月</u>	<u>NNG</u>	이
	JKS					
000145	名명相 · 상 · 이	011	<u>명상</u>	<u>名相</u>	<u>NNG</u>	이
	JKS					

- 주격, 목적격, 지정사, ᄃ-, 공동격조사 '-와/과' 생략형은 따로 정보를 주지 않는다.

- 협주의 '漢字는 무엇이나라'에서 한자의 태그는 SH.

- 격조사와 보조사/접속조사와 보조사 연속체 처리-분석하지 않는 것을 원칙. 선어말어미 구조체도 마찬가지.

2.4.3 개별적으로 처리해야 하는 문제들

- ‘와’ 생략됨. 생략된 조사에는 태그를 달지 않는다.

· 히 1 히 NNG

- ‘ $\text{ㅁ} > \emptyset$ ’ 유형-복원이 쉬운 형으로 어간을 처리.

갓가·을 001 갓가오 VA ㄹ ETM

- 어+잇 구성의 처리-그대로 둠.

머·갓·고 011 덕 VV 갓 EP 고 EC

- 동사+동사 구성은 복합동사로 보고 분석하지 않는다.

:우·듯ㅎ·니·라 21011 우듯ㅎ VV 니라 EF

• 원어절이 현대국어 띄어쓰기를 따르지 않고 붙어 있을 때, 부득이하게 띄어쓰기를 가정하고 후행어절의 품사에 따라서 주석을 단다. 이는 2차 가공에서 다시 현대국어에 따라 분석하여 주석을 줄 것이다.

093092	가셔죄인을	<u>가셔죄인</u>	NNG	을	JKO
017664	가두논거시	<u>가두논것</u>	NNB		JKS
080013	갓치ㅎ니죄가너희요	<u>갓치ㅎ니죄가너희</u>	NP	요	EF
123520	가기를말ㅎ리요	<u>가기를말ㅎ</u>	VV	리	EP 요
	EF				
047534	가셔자고	<u>가셔자</u>	VV	고	EC

- 어간의 일부가 생략된 경우에는 그 형태에 대한 주석을 주지 않는다. 이는 복원을 했을 때, 생략된 형태의 표기를 충분히 추론할 수 있기 때문이다.

055360 거역지 거역 NNG 지 EC

2.5 주석 결과물의 형태 : 수직적 주석 말뭉치

주석 작업의 원칙은 원자료를 최대한 반영하는 것이다. 따라서 형태분석이 직관과 어긋나더라도 복원가능성이 높은 형태를 중심으로 분석한다.

- TEI 태그들은 모두 삭제하고 작업한다. 단 검수 혹은 검색의 편의를 위해서 페이지 태그는 살려둔다.
- 주석들 간은 한 개의 탭으로 간격을 둔다.
- 접사류는 분석 대상에서 제외한다. 따라서 접두사와 접미사에는 따로 태그를 설정하지 않는다. 한편, 선어말어미 구성체나 격조사와 보조사, 접속조사와 보조사 연속체도 분석하지 않는다.

2.6 실제 주석 말뭉치의 예

```
<!DOCTYPE tei.2 SYSTEM "c:\Wsgml\Wdtd\Wtei2.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>釋譜詳節13, 전자 파일</title>
        <author>?</author>
        <sponsor>대한민국 문화관광부</sponsor>
        <respStmt><resp>국립국어원 말뭉치 입수, 표준화, 헤더 붙임</resp>
          <name>국립국어원</name>
        </respStmt>
```

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

```
</titleStmt>
<editionStmt>
  <edition><date>19?</date>전산입력</edition>
</editionStmt>
  <extent>4819 어절</extent>
<publicationStmt>
<distributor>국립국어원</distributor>
<idno>6PT_0001.hwp</idno>
<availability><p>배포 불가</p></availability>
</publicationStmt>
<notesStmt>
<note><p>이 텍스트는 국립국어원 작업 내용 유지</p></note>
</notesStmt>
<sourceDesc>
<bibl><author>?</author>
  <title>釋譜詳節</title>
  <pubPlace>?</pubPlace>
  <publisher>?</publisher>
  <date>19?</date>
</bibl>
</sourceDesc>
</fileDesc>
<encodingDesc>
<projectDesc><p>21세기 세종계획 3차년도 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>1447</date></creation>
```

```

<langUsage>
  <language id=KO usage=99>한국어, 고어</language>
</langUsage>
<textClass>
  <catRef scheme='SJ21' target='P5CD'>역사자료: 15세기, 언해/번역자료, 종
교서류</catRef>
</textClass>
</profileDesc>
<revisionDesc>
  <change>
    <date></date>
    <respStmt>
      <resp>입력자</resp><name>송기중</name>
    </respStmt>
  </change>
  <change>
    <date>1995/07/15</date>
    <respStmt>
      <resp>교정자</resp><name>송기중</name>
    </respStmt>
    <item>교정</item>
  </change>
  <change>
    <date>2004/11</date>
    <respStmt>
      <resp>프로젝트 책임자</resp><name>임용기</name>
    </change>
  </revisionDesc>
</teiHeader>

<text>
<body>

```

21세기 세종계획 국어기초자료 구축-특수자료 구축 소분과

00001 <釋詳13:001a>

00002 釋·석譜:봉詳쌍節·諍第·땡十·씩삼 1201110 석봉쌍절땡씩삼 釋
譜詳節第十三

00003 부:테 02 부터 NNG | JKS

00004 王왕솜·상城쌍耑깁閨쌍堀·꿍山산 0100010 왕상쌍깁쌍꿍산 王
舍城耑閨堀山 NNP

00005 中똥·에 01 똥 中 NNB 에 JKB

00006 :겨·샤 21 겨 VV 샤 EC

00007 [·이 1 이 MM

00008 브·터 01 블 VV ㅏ EC

00009 法·법華쌍經經 100 법쌍經 法華經 NNG

00010 니르·시느 0010 니르 VV 시 EP 느
ETM

00011 靈령山산會·쌍·라] 0011 령산쌍 靈山會 NNG 라 EF

00012 :굴·근 21 굴 VA 一ㄴ ETM

00013 比·땡丘쿨衆·중 101 땡쿨중 比丘衆 NNG

00014 一·萬萬·먼 11 萬먼 一萬 MM

00015 二·쌍千천 10 쌍천 二千 MM

00016 :사름·과 201 사름 NNG 과 JC

00017 혼·딕 01 혼딕 MAG

00018 잇·더시·니 0101 잇 VV 더 EP 시 E P
니 EC

00019 [이 0 이 MM

00020 ·뵤 1 뵤 NNG

00021 부텃 00 부터 NNG ㅅ ETM

00022 ·나·히 11 나히 NNG

00023 널·흔 01 널흔 NR

00024 흥나·히·러시·니 001101 흥나히 NR 러 EP 시 E P
니 EC

00025 穆·목왕 10 목왕 穆王 NNP

00026 마순 00 마순 NR

00027 다·숫샷 010 다숫샷 NR

00028 ·히 1 히 NNB

</body>

</text>

</tei.2>

여 백

제 3 부

전문가 초청 특강 발표 요지

여 백

제 1 장

전문가 초청 특강 발표 요지

1. 전문가 초청 특강 일지

연구의 결과를 심도 있게 점검하고 향후 방향에 대해 논의하기 위하여, 매 워크숍에 각 분야 전문가를 초청하여 전문가 초청 특강을 개최하였다.

1차 전문가 초청 강연(6월 11일 개최)

- 장소 : 연세대학교
- 발표자: 박병선(KORTERM, KAIST)
- 주제 : 세종계획 분과별 형태표지 활용에 대하여

2차 전문가 초청 강연(8월 20일 개최)

- 장소 : 연세대학교
- 발표자: 안동연(전북대학교)
- 주제 : 한국어 정보처리에서의 태그 세트 연구 현황

3차 전문가 초청 강연(11월 12일 개최)

- 장소 : 고려대학교

- 발표자: 김성묵(한국 IBM 기술연구소 HCT 팀장)
이석재(연세대학교)
- 주제 : 언어 복합 정보 DB와 언어태깅(LT)

4차 전문가 초청 강연(12월 18일 개최)

- 장소 : 연세대학교
- 발표자: 장원재(계명대학교)
- 주제 : 한일 대역코퍼스로 본 한일어휘의 유사점과 차이점

제 1 차

전문가 초청 특강

박 병 선(KORTERM, KAIST)

여 백

KAIST

세종계획 분과별 형태표지 활용에 대하여

박병선(KORTERM, KAIST)

언어자원은행
Bank of Language Resources
한국과학재단

차례

- 다룰 내용
- 분과별 표지 개관
- 기초자료 분과 형태 표지의 특징
- 특수자료 분과 형태 표지의 특징
- 품사별 태깅의 실제와 문제
- 자료 호환의 예
- 앞으로의 방향
- 토론

오늘 다룰 내용

- 형태(품사) 표지 기호
- 분과별 형태 표지 특징
- 형태(품사)별 실제 태깅 문제 고찰
 - 기초자료 분과 vs. 특수자료 분과
 - 기초/특수 vs. 병렬
 - 기초/특수 vs. 북한/해외
 - 실제 태깅 원칙 비교
 - 표준화 방안
- 자료 호환의 예
 - ◆ KAIST vs. 세종기초자료
- 형태(품사) 표지 적용 표준화에 대한 제언

3

분과별 표지 개관 - 기초자료 분과/병렬 분과

대분류	소분류	세분류	대분류	소분류	세분류
(1) 채언	명시NN	일반명시NNG	(6) 의존형태	어미E	선어말어미EIP
		고유명시NNP			종결어미EIP
		의존명시NNB			연결어미EC
	대명시NP	대명시NP			명사형전성어미ETN
	수시NR	수시NR			관형형전성어미ETM
(2) 용언	동사VV	동사VV	(7) 기초	접두사XP	채언접두사XPN
	형용사VA	형용사VA		접미사XS	명사파생접미사XSN
	보조용언VX	보조용언VX			동사파생접미사XSV
	지정사VC	규정지정사VCP			형용사파생접미사XSA
(3) 수식언	관형사IMM			어근XR	어근XR
	부사MA	일반부사MAG		마침표, 물결표, 느낌표	SF
(4) 독립언	관탄사IC	관속부사MAJ		쉼표, 가운뎃점, 줄표, 빗금	SP
		관탄사IC		따옴표, 괄호표, 줄표	SS
(5) 관계언	격조사JK	주격조사JKS		물음표	SE
		보격조사JKC		붙임표(붙임, 순김, 특김)	SO
		관형격조사JKG		외국어	SL
		목적격조사JKO		한자	SH
		부사격조사JKB		기타 기호(논리 수학기호, 화폐 기호) 등)	SW
		호격격조사JKV		명사추정법주	NF
		인용격조사JKQ		용언추정법주	NV
		보조사JX		숫자	SN
	접속조사JC	접속조사JC		본적불능법주	NA

4

분과별 표지 개관 - 특수자료(구어) 분과

(7) 억양 기호	마침표, 물음표, 느낌표	SF
	쉽표	SP
(8) 분석불능범주	(끊어진 어절)	UNT
	(불분명한 어절)	UNC
	(기타)	UNA

5

분과별 표지 개관 - 특수자료(북한/해외) 분과

		기본종사		특정종사				기본종사		특정종사		
종사 단위	대단위	소단위	지단위	고단위		교학소 단위	교학소 단위	지단위	중단위	고단위		
	재민	영사	일반영사		NN			조사	J			
고유영사			NQ		여기	WB						
의존영사			일반		NB		어미	어말어미	종결어미		EF	
			단위		NC				연결어미		EC	
대명사			NP		경성어미				영사형		ETN	
수사		NR		관형사형		ETM						
용언	일반종사		VV		가타 병주	가타 병주	접사	접미사	XS			
	일반형용사		VA						마침표		SF	
	보조용언		VX						쉼표, 가운대 점		SP	
	지칭사		VC						따옴표		SS	
	관형사		MM						기타, 논리수 박		SW	
부사		MA		명사추정		NF						
감탄사		IC		용언추정		NV						
				분석불능		NA						

6

분과별 형태표지 기호체계 문제

□ 문어 vs. 구어

- ◆ 큰 문제 없음
- ◆ 분석 불능 표지 NA 하위 범주화가 바람직...
 - 예) 불분명한 어절: UNT -> NAT

□ 문/구어 vs. 북한/해외

- ◆ 일부 범주에서 차이 많이 보임
- ◆ 품사 설정 체계에도 차이 많이 나타남
 - 명사, 지정사, 부사, 조사, 어기, 선어말어미, 접두사, 접미사
 - 일반명사: NNG vs. NN
 - 고유명사: NNP vs. NQ
 - 의존명사: NNB vs. NB, NC
 - 지정사: VCP, VCN vs. VC
 - 부사: MAG, MAJ vs. MA
 - 조사: {JKS, JKC, JKG, JKO, JKB, JKV, JKQ}, JX, JC vs. J
 - 어근(어기): XR vs. WB
 - 선어말어미: EP vs. T, H
 - 접두사: XP vs. 없음
 - 접미사: XSN, XSV, XSA vs. XS

7

형태표지의 특징 - 기초자료

□ 국어학적, 전산학적 문제 모두 고려

- 결과 자동처리 가능시 하위 세부 태깅하지 않음
 - 단계적 분석 가능
- 국어학적 논란거리 (접두사, 한자접사 등) 최소 태깅
- 자동 태거 구현 불가능 부분 고려

□ 다년간의 실제 경험 반영

- 태깅 숙련자 (1996년부터 현재까지)
- 각종 태깅 도구와 노하우

□ 이해결 문제 아직 많음

- 국어학적 논란 여지
- 자동화 어려운 부분

8

형태 표지의 특징 - 특수자료(구어)

- 기초자료 분과와 기본 방향 일치
- 구어 말뭉치에 필요한 부분 추가
 - ◆ 원칙은 유지
- 태깅 실제 지침 -기초자료 분과와 차이
오늘 주로 다룰 내용

9

표지별 실제 태깅 문제 - 일반명사

- 대체로 일치된 원칙
- 일부 차이 보임
 - 시간 명사(명사적/부사적 용법)
 - 기초: 표준국어사전 기준
 - 특수: 격조사/보조사 결합 여부
 - 복합 명사
 - 기초: 표준국어사전 기준
 - 특수: 복합어 구조 인정(예: 월요일날)
일관성 문제 고려-사전보다 우선 고려(예: 구내-)
 - 음차 외국어
 - 기초: 명사로 분석
 - 특수: 분명한 경우 해당 품사로 분석
 - 고사성어
 - 기초: 원문 띄어쓰기 고려
 - 특수: 하나로 - 구어 말뭉치 특징

10

표지별 실제 태깅 문제 - 고유명사, 의존명사

□ 고유명사

- ◆ 대체로 일치
- ◆ 구어 말뭉치 특성상 띄어쓰기 문제 없음
- ◆ 차이점: 특수분과에서 상품명 고유명사로 인정

□ 의존명사

- ◆ 대체로 일치
- ◆ 구어말뭉치에서는 다양한 이형태 인정

11

표지별 실제 태깅 문제 - 대명사, 수사

□ 대명사

- ◆ 대체로 일치
- ◆ 기초자료 분과: 표준국어사전 기준

□ 수사

- ◆ 대체로 일치
- ◆ 일부 명사와 혼동되는 것들은 목록화하여 일괄처리
 - 예: 하나 -> 무조건 수사로
- ◆ 수사와 수관형사 구분은 임흥빈(1998)을 따름
 - 수관형사 목록 정함

12

표지별 실제 태깅 문제 – 동사, 형용사, 지정사

□ 동사, 형용사

◆ 기초자료: 표준국어사전 기준

◆ 특수자료: 사전 기준

일부는 일관성 고려, 복합용언 목록 작성

기초자료 분과와 차이 유발(예: 짜증나다)

□ 지정사

◆ 세종 표지의 특징

◆ 구어 태깅 현실 잘 반영

◆ 대체로 일치

13

표지별 실제 태깅 문제 – 보조용언

□ 기초자료 분과

◆ 보조용언 목록 작성 -> 문맥으로 최종 판단

◆ 자동처리 어려움

◆ 특수한 경우 고려(예: 버릇하다)

□ 특수자료 분과

◆ 보조용언 목록 작성 -> 기준(일부 사전 참조)

◆ 자동처리 가능

14

표지별 실제 태깅 문제 - 관형사, 부사, 감탄사,

□ 관형사

- ◆ 기초자료
 - 표준국어사전 목록과 일치
 - 일부 문맥 참조 필요(예: 울)
- ◆ 특수자료
 - 자체 목록
 - 기초자료 분과와 목록 차이 존재(예: 울, 후)

□ 부사, 감탄사

- ◆ 차이 없음 - 표준국어사전 기준
- ◆ 특수자료 - 다양한 이형태 인정

15

표지별 실제 태깅 문제 - 조사

□ 격조사

- ◆ 큰 차이 없음
- ◆ 구어 말뭉치 특성 반영

□ 접속조사

- ◆ 두 분과는 일치
- ◆ 표준국어사전과는 일부 차이(예: -하며)
- ◆ 문맥으로 판단 필요(예: 철수와 같이 놀았다. 철수와 영희가 왔다.)

□ 보조조사

- ◆ '이다'의 활용어미와 혼동 유발
 - 기준 1: 대상 형태가 용언의 어미로 사용되는가
 - 기준 2: 대상 형태가 체언에 후행할 때 서술어의 자격을 갖는가
- ◆ 기초자료/특수자료 일부 목록 차이
 - 예: -그러

16

표지별 실제 태깅 문제 - 어미

□ 선어말어미

- ◆ 설정한 목록만 인정 - 기초, 특수 일치

□ 어말어미

- ◆ 지침에서 제시한 일부 목록 차이 - 원칙 일치
- ◆ 구어말뭉치 특성 반영

□ 전성어미

- ◆ 대체로 일치
- ◆ 관형형전성어미 '-런': 특수 자료 지침에 없음

17

표지별 실제 태깅 문제 - 접사, 어근

□ 접두사

- ◆ 체언접두사만 인정 - 기초/특수 일치
- ◆ 목록화

□ 접미사

- ◆ 명사파생
 - 일부 목록 차이(예: -가(價) -> 기초자료만 있음)
- ◆ 동사파생 - 일치
- ◆ 형용사파생
 - 일부 목록 차이(예: -갈, -만하 -> 특수자료만 있음)

□ 어근

- ◆ 기초: 목록화
- ◆ 특수: 원칙만 제시 -> 목록화 필요

18

표지별 실제 태깅 문제 - 기타

- 기호, 준말, 분석불능
 - ◆ 대체로 일치
 - ◆ 구어말뭉치 특성 반영
- 합성어 분석
 - ◆ 표준사전 최대한 반영
 - ◆ 띄어쓰기 정보 활용
 - ◆ 말뭉치 원본 정보 반영
- 1음절어 처리 - 복잡한 원칙

19

실제 태깅 문제: 문어/구어 vs. 병렬

- 기본 원칙은 문어 원칙과 동일
- 세부 지침은 차이 보임
 - ◆ 원본 띄어쓰기 정보 - 원칙에 따라 수정
 - ◆ 명사의 부사 인정(예: 최근-병렬-> 부사적 용법)
 - ◆ 어근 인정 문제(예: 노골, 간혈을 어근으로 봄)
 - ◆ 조사 목록 일부 차이(예: 부사격조사 목록)
 - ◆ 지침에서 구체적인 예나 목록 제시 더 필요

20

실제 태깅 문제: 문어/구어 vs. 북한/해외

□ 표지체계 차이 큼

- 명사, 지정사, 부사, 조사, 어기, 선어말어미, 접두사, 접미사
- 일반명사: NNG vs. NN
- 고유명사: NNP vs. NQ
- 의존명사: NNB vs. NB, NC
- 지정사: VCP, VCN vs. VC
- 부사: MAG, MAJ vs. MA
- 조사:
 - {JKS, JKC, JKG, JKO, JKB, JKV, JKQ}, JX, JC vs. J
- 어근(어기): XR vs. WB
- 선어말어미: EP vs. T, H
- 접두사: XP vs. 없음
- 접미사: XSN, XSV, XSA vs. XS

21

품사 표지 표준화

□ 각 분과별 비교 – 실무자

- ◆ 각 세부 표지별 지침과 적용 방법 차이 파악

□ 기구축된 각 분과별 자료의 표준화 방안 모색

- ◆ 정책적 차원
- ◆ 조율 필요

□ 각 분과별 언어자원 일관성 확보

□ 자원의 호환 사용을 위한 실무자 모임

□ 실제 사용 가능 언어자원 활용도구 개발

22

언어자원 호환의 문제

□ 세종계획 기타 분과 자원과의 호환 문제

- ◆ 전자사전 분과
 - 표지, 품사 체계의 차이
 - 말뭉치 활용
 - 품사별 대상 목록 차이

□ 기타 기관 자원과의 호환 문제

- ◆ 카이스트 자원
- ◆ ETRI 자원
- ◆ 기타...

23

호환 예: BOLA의 필요성

□ 국내외 언어자원의 효율적 활용

- ◆ KAIST에서 10여 년간 축적한 각종 자료
- ◆ 국내외에 산재한 언어자료 파악과 활용
- ◆ 외국의 LDC, ELDA 등이 모델

24

호환 예: 태깅의 실제

자료 호환은 별 문제가 아니다.

ㄱ. KAIST: 자료/ncn 호환/ncn+은/jxc 별/mma

문제/ncn+가/jcc 아니/paa+다/ef+./sf

(ncn:비서술성명사, jxc:보조사, mma:성상관형사,
jcc:보격조사, paa:성상형용사, ef:종결어미, sf:마침표)

ㄴ. 세종계획: 자료/nng 호환/nng+은/jx 별/mm

문제/nng+가/jkc 아니/vcn+다/ef+./sf

(nng:일반명사, jx:보조사, mm:관형사,
jkc:보격조사, vcn:부정지정사, ef:종결어미, sf:마침표)

25

호환 예: 품사표지 비교 - 개관

□ 세종 형태분석 말뭉치 구축 원칙

- ◆ 의미, 형태, 기능 중에서 형태/통사적인 것을 더 중시
- ◆ 어문규정과 일반 문법 규정에 더 맞는 것
- ◆ 형태소보다는 형태 분석을 더 중요하게 보는 것

□ KAIST 품사부착 말뭉치 구축 원칙

- ◆ 세종계획의 원칙과 유사
- ◆ 일부 표지 세종계획에 비해서 세분화

26

호환 예: 품사표지 비교 - 체언

카이스트				세종		
체언	보통명사	서술성	동작성, 상태성	체언	명사	일반명사
		비서술성				고유명사
	고유명사					의존명사
	의존명사	단위성, 비단위성				
	대명사	인칭, 지시			대명사	
	수사	양수사, 서수사			수사	

27

호환 예: 품사표지 비교 - 체언

- 카이스트에서 분류한 체계가 세종보다 세분화
- 일관성의 문제를 비롯한 여러 어려움이 예상
- 세종 일반명사에 대하여 다시 태깅
- 동일한 단어가 파생접사의 결합 여부에 따라서 품사가 다르다는 것은 논란의 여지
 - ◆ ㄱ. 세종: 변환/NNG+하/XSV+다/EF → 카이스트: 변환/NCPA+하/XSV+다/EF
 - ◆ ㄴ. 세종: 변환/NNG+은/JX → 카이스트: 변환/NCN+은/JX

28

호환 예: 품사표지 비교 - 용언

카이스트			세종		
용언	동사	지시, 일반	용언	동사	
	형용사	지시, 성상		형용사	
	보조용언			보조용언	
지정사				긍정, 부정	

29

호환 예: 품사표지 비교 - 용언

- 세종에서 지정사를 하나의 독립된 단위
- 카이스트에서는 동사와 형용사를 하위 분류
- 자동 재태깅에도 큰 문제는 없다고 보임
 - ◆ 세종: 이러다(이러/VV(일반동사))+다) 어쩌려고 그러니?
 - 카이스트: 이러다(이러/PVD(지시동사))+다) 어쩌려고 그러니?

30

호환 예: 품사표지 비교 - 수식언

카이스트			세종		
수식언	관형사	지시, 성상	수식언	관형사	
	부사	지시, 접속, 일반		부사	일반, 접속

31

호환 예: 품사표지 변환 - 수식언

- 카이스트가 더 하위분류
- 해당 단어들을 목록화하여 자동처리가 가능
 - ◆ 지시관형사 ‘이, 그, 저’
 - ◆ 지시부사 ‘이리, 그리, 저리, 이리저리, 어디’

32

호환 예: 품사표지 비교 - 독립언

카이스트		세종	
독립언	감탄사	독립언	감탄사

33

호환 예: 품사표지 비교 - 관계언

카이스트			세종		
관계언	격조사	주격, 목적격, 보격, 관형격, 호격, 부사격, 접속격, 공동격, 인용격	관계언	격조사	주격, 보격, 관형격, 목적격, 부사격, 호격, 인용격
	보조사	통용, 중결		보조사	
	서술격 조사			접속조사	

34

호환 예: 품사표지 비교 - 관계언

- 상하체계는 일부 차이
- 각 표지는 같은 대상에 표시
- 세종의 용언에 해당하는 긍정지정사를 카이스트의 서술격조사로 변환
- 카이스트의 경우 통용과 종결보조사를 구분
 - ◆ 세종: 단풍이 들었어요.(들/VA+었/EP+어요/EF+./SF)
 - ◆ → 카이스트: 단풍이 들었어요.(들/PVG+었/EP+어/EF+요/JXF+./SF)

35

호환 예: 품사표지 비교 - 어미

카이스트			세종	
어미	선어말어미		어미	선어말, 종결, 연결, 명사형전성, 관형형전성
	연결어미	대등, 종속, 보조적		
	전성어미	명사형, 관형사형		
	종결어미			

36

호환 예: 품사표지 비교 - 어미

- 카이스트의 경우가 더 하위분류
- 표지 체계의 상하체계에서 차이
- 자동처리에 문제를 일으킬 소지
- 세종 연결어미에 대해서는 전면적인 재태깅이 필요

37

호환 예: 품사표지 비교 - 접사

카이스트			세종		
접사	접두사	*체언만 인정	의존 (접사, 어근)	접두사	체언접두사
	접미사	명사파생, 동사파생, 형용사파생, 부사파생		접미사	명사파생, 동사파생, 형용사파생
				어근	

38

호환 예: 품사표지 비교 - 접사, 어근

- 접사는 카이스트 표지와 세종이 큰 차이 없음
- 세종에서는 어근 범주를 설정
- 실제 지침에서 제시하는 접사의 목록에서 차이
- 부사파생 접사는 세종에서 인정하지 않음
- 목록을 비교하여 재태깅 대상을 선정

39

호환 예: 품사표지 비교 - 기타

카이스트		세종	
기호	첨표, 마침표, 여는 따옴표, 닫는 따옴표, 이음표, 줄임표, 단위, 기타	기호	문장종결 (마침표, 물음표, 느낌표), 첨표, 붙임표, 외국어, 한자
외국어		분석불능	끊어진 어절, 불분명 어절, 기타

40

호환 예: 품사표지 비교 - 기타

- ☐ 세종에서 분석불능 표지가 있음
- ☐ 카이스트에서 사용할 수 없는 표지
- ☐ 코퍼스에서 차지하는 비율이 미약

41

결론

- ☐ 각 기관별 자원의 일관성 확보
- ☐ 세종계획 하위 분과별 표준 실무 회의
 - ◆ 기타 자원 관리자로 확대
- ☐ 자료 호환에서 발생하는 문제점 정리
- ☐ 자료 호환 방안 모색 - 현실적으로 반드시...
- ☐ 앞으로의 새로운 구축에서는 표준안 필요
- ☐ 정책적 표준 권고안 마련 필요
- ☐ 자원 통합 관리
- ☐ 자원의 효율적 활용

42

여 백

제 2 차

전문가 초청 특강

안 동 언(전북대학교)

여 백

한국어 정보처리에서의 태그 세트 연구 현황

2004. 8. 20

전북대학교 전자정보공학부

안 동 언

duan@chonbuk.ac.kr

목차



- 한국어 정보처리의 역사
- 말뭉치
- 태그 세트
- 태깅

한국어 정보처리의 역사



- 한글 처리 시스템
- 기계번역
- 국어정보베이스
- 우리말 정보처리 규격 심포지움
- 자연어 정보처리 표준화
- 대용량 국어 정보 심층처리 및 품질관리 기술 개발
- 대용량 음성(음향)/언어/영상 DB 구축 및 표준화 사업 - **KSURIMAL** -
- 정보검색

3

한글 처리 시스템

한국어 정보처리의 역사



- **Samsung/NEC S-100**
 - 김길창 교수 (KAIST)
 - 1982
 - 한글 코드
 - 한글 폰트
 - 한글 에디터
 - 한글 포맷터
- 명필
 - 이기식 부장, 박동인 부장 (SERI)
 - 1983
 - 고려시스템

4

기계번역(1)

한국어 정보처리의 역사



○ SERI

- ATLAS/JK('90)
- 박동인 부장(KISTI), 황도삼 교수(영남대), 김태완 교수(인제대)

○ KAIST

- PIVOT/EK('86), MATES/EK('92), ALKOL('96)
- 김길창 교수, 최기선 교수(KAIST), 이종혁 교수(포항공대), 안동언 교수(전북대)

5

기계번역(2)

한국어 정보처리의 역사



○ 서울대

- KEMTS('85), KSHALT('87), EKTRAN('90)
- 김영택 교수(서울대), 권혁철 교수(부산대)

○ 포항공대

- COBALT/JK('94)
- 이종혁 교수(포항공대)

○ ETRI

- 한/영/일/중 기계번역, 자동통역
- 박상규 박사(ETRI)

6

국어정보베이스(2)

한국어 정보처리의 역사



○ 국어정보베이스

- 최기선 교수(KAIST), '95 ~ '97
- 핵심 S/W 기술개발 STEP2000
국어정보처리기술개발 통합국어정보베이스
- 우리말 정보처리 규격 심포지움, '96, '97

9

국어정보베이스(3)

한국어 정보처리의 역사



○ 세부 과제

- 국어정보베이스 통합시스템
- 텍스트 코퍼스 및 전자사전 관리 시스템
- 한국어 형태 태깅 시스템
- 한국어 구문 태깅 시스템 및 구문트리 뱅크
- 한국어/영어 정렬 시스템
- 한국어 형태소 해석기 및 형태소 해석 사전 개발
- 한국어 정보 획득 도구
- 문서구조 표현을 위한 표준화에 관한 연구
- 한글 정보처리를 위한 표준 라이브러리 개발
- 균형화 코퍼스 구축 표준 방법론 및 시범 패키지
- TreeBank 및 전문용어 데이터베이스
- 한글 필기체 데이터베이스
- 한국어 음성 데이터베이스
- 분류체계에 의한 영한 대역 사전

10

우리말 정보처리 규격 심포지움

한국어 정보처리의 역사



- 국어정보베이스
- KAIST, '96, '97
- 주제
 - 국어 코퍼스 구축 방안
 - 국어 형태·통사 태그 규격
 - 국어 구문트리뱅크 규격
 - 전자사전 및 텍스트 마크업 기술 규격

11

자연어 정보처리 기술 표준화

한국어 정보처리의 역사



- ETRI, '98
- 전자사전 분과: 전자사전 표제어 선정 기준안
- 태그세트 분과: 표준 태그세트 설정 및 지침
- 태깅방식 분과: 품사 부착 말뭉치 구축 지침
- 원시말뭉치 분과: 평가용 말뭉치 구축을 위한 지침
- 분석/검증 분과: 형태소 태깅 평가 및 검증 지침

12



- **KAIST, '98**
- 사전 및 코퍼스 관리개발 환경
- 전문용어 다국어 통합정보사전 연구
- 코퍼스 품질관리 방법 연구
- 범용적 확률문법개발환경 연구
- 문장분석/생성용 전자사전 개발/관리 시스템
- 국어정보처리 품질관리 시스템
- 분산 언어자원 관리 기술



- **KAIST, 2000**
- **KSURIMAL**
- 언어 부문
 - 언어자원 표준화 분과
 - 기계번역 분과
 - 정보검색 분과
- 목표
 - 표준화
 - **DB** 구축
 - 평가
 - 테스트 세트 구축



○ 색인

□ Bigram

□ 명사 중심의 형태소 분석

□ 모든 품사의 형태소 분석

○ 요구 사항

□ 속도

□ 과분석

□ 띄어쓰기/철자가 틀린 문장 처리

□ 매우 간단한 태크 세트

15

말뭉치 (1)



○ 말뭉치 종류

□ 연세대

□ 고려대

□ KAIST

• 국어정보베이스

• 태깅

□ ETRI

□ 국립국어연구원

□ 21세기 세종계획

• 세종 말뭉치

16

말뭉치 (2)



- 국어정보베이스 말뭉치 설계 방향
 - 1980년 이후의 현재 언어 용법을 보여줄 수 있는 대표성을 지닌 균형 코퍼스 지향
 - 출판물의 양과 독서 경향에 따른 텍스트 선정
 - 비용 효과성
 - 다목적: 사전 구축, 문법 구축, 기계 번역, 문서 이해
 - 표준화
 - 자원 공유
 - 객관적 연구평가 유도

17

태그 세트



- 한국어 정보처리의 태그 세트
 - 국어정보베이스
 - 1996
 - 한국어 형태·통사 태그 규격
 - 한국어 구문 태그 규격

18

한국어 형태·통사 태그 규격 (1)

태그 세트



○ 9가지 최상위 분류와 54개의 형태·통사 태그

○ 설정 원칙

- 한국어를 대상으로 하나, 한국어 문장에서 두루 사용하는 외국어나 특수기호들도 대상으로 삼는다.
- 품사태그 집합의 설정은 학교문법을 최대한 반영하며, 통사론적 분석 위주보다는 형태론적 분석을 위주로 한다.
- 품사태그는 여러 분야(형태소 레벨과 통사적 레벨)에서 다양한 용도로 사용할 수 있도록 계층적으로 분류한다.

19

한국어 형태·통사 태그 규격 (2)

태그 세트



○ 문제점

- 규격 제정의 기준이 명확하지 못하여 태그 집합의 일관성이 결여되어 있다.
- 태깅을 위한 지침서가 명확하게 마련되지 못했다.
- 자동사와 타동사의 분류가 필요하다.
- “상태성”과 “동작성”의 구분이 명확하지 못하다.
- 지시사의 분류가 불필요하다.
- “수관형사” 태그를 추가할 필요가 있다.
- 격조사 소분류의 기준이 모호하다.
- 접두사의 세분류가 필요하다.
- 보통명사 소분류와의 일관성 유지를 위하여 서술성 부사와 비서술성 부사의 구분이 필요하다.

20

한국어 형태·통사 태그 규격 (3) 태그 세트



○ 표준화의 목적

- 국어정보처리의 효율적인 연구협력체계 확립
- 국어 형태 통사 태그의 보급
- 태깅된 코퍼스의 보급
- 태그가 부착된 사전의 보급
- 국어 언어정보획득 도구의 보급
 - 형태소 분석기
 - 태거
 - 용례 추출기
 - 통계 정보 추출기

21

한국어 형태·통사 태그 규격 (4) 태그 세트



○ 토의 사항

- 인간을 위한 것인가? 컴퓨터를 위한 것인가?
- 언어학 연구를 위한 것인가? 전산학에서의 자연어 처리를 위한 것인가?
- 정보획득과 생성 두 관점에 맞는 태그 규격이 필요하다.
- 여러 응용 분야에 따라 필요한 태그 세트의 복잡도가 달라지므로 이에 대한 고려가 있어야 한다.

22



○ 합의 사항

- 한국어 정보처리 시스템의 구현 및 성능 향상이라는 관점에서 정의되어야 한다.
 - 확장성, 호환성, 효율성, 정밀성, 경제성, 일관성
- 표준화는 중분류까지 권고안으로 정한다.
 - 응용 분야에 중립적이어야 한다.
 - 세분류를 제안하지만 개별 시스템 및 응용 분야에 따라서 확장할 수 있도록 한다.
- 태깅시 일관성을 유지하고 중의성을 최소화할 수 있어야 한다.



- 8개 구문 태그
- 구구조 문법 기반

태깅 (1)



○ 태깅의 문제점

- 태깅 표기 형식의 표준화
- 다어절의 태깅
 - 제목, 복합어
- 태깅된 코퍼스에서 원시 코퍼스의 획득 방안 필요
 - 불규칙 현상
 - 음운 현상
 - 축약, 탈락, 준말
- 자동 태깅에서 모호성이 배제되도록 국어 형태 태그 규칙의 표준화 추진
 - 세분화된 태그로 인해 자동 태깅이 어렵다.

25

태깅 (2)



○ 태깅 방식 표준안 결정 요인 (1)

- 지침의 명확성
 - 정확한 지침이 태깅 일관성 유지에 효과적
- 지침의 일관성
 - 일관적이고 예외 조항이 적어야 기억이 용이
- 직관과의 부합성
 - 사람의 언어직관에 부합되어야 가독성이 좋고 기억이 용이
- 표현의 간결성
 - 보다 간결한 표현이 가독성이 좋고 기계처리에 효율적임

26

태깅 (3)



○ 태깅 방식 표준안 결정 요인 (2)

- 표현의 완전성
 - 모든 사례를 표현할 수 있어야 함
- 정보 획득의 용이성
 - 정보(표제어, 품사, 규칙)의 획득이 용이해야 실용적
- 정보의 양
 - 정보의 양이 적어야 사람 및 기계 처리에 효율적
- 수작업의 용이성
 - 수작업이 쉬운 방향으로
- 기계처리의 용이성
 - 도구의 도움을 받기 위해

27

태깅 (4)



○ 품사부착말뭉치 구축을 위한 지침서

- 품사부착말뭉치의 형태
 - 세로형, 주석, 원시어절오류, 태깅오류
- 품사태그 표현
 - 단일 태그, 미지 태그, 복합어 태그, 다중어 태그
- 품사분류 지침
 - 각 품사별 분류 기준, 예, 참고사항
- 형태소분석 지침
 - 원형(사전표제어) 결정 기준
 - 음운현상 처리 기준: 규칙활용, 불규칙활용, 축약 및 탈락, 준말
 - 형태소 분리 기준: 복합형 분리 방법
 - 표준어 오류에 대한 처리 기준: 형태소 분석 및 태깅 지침
- 품사 부착 예

28

참고문헌(1)

○ 연구보고서

- 우리말 컴퓨터를 위한 개발 계획 수립, SERI, 1995
- 국어정보베이스, KAIST, 1995~1997
- 대용량 국어 정보 심층처리 및 품질관리 기술개발, KAIST, 1998

○ 심포지움/워크샵 자료

- 우리말 정보처리 규격 심포지움, 1996, 1997
- 자연어 정보처리 기술 표준화 워크샵, 1998
- 대용량 음성(음향)/언어/영상 DB 구축 및 표준화 과제 워크샵, 2000

29

참고문헌(2)

○ 논문지

- “우리말 표준규격” 특집, 인지과학, 제7권 제4호, 1996
 - 한국어정보베이스를 위한 형태·통사 태그 표준에 관한 연구
 - 구문 트리 부착 코퍼스 구축을 위한 한국어 구문 태그

○ 학회지

- “한글 및 한국어 정보처리” 특집, 정보과학회지, 제12권 제8호, 1994
- “컴파일러 및 자연어 처리” 특집, 정보과학회지, 제14권 제7호, 1996
- “한글 및 한국어 처리” 특집, 전자공학회지, 제24권 제9호, 1997
- “국어공학” 특집, 정보과학회지, 제15권 제10호, 1997
- “정보검색” 특집, 정보과학회지, 제16권 제8호, 1998
- “사용자 중심의 정보검색” 특집, 정보과학회지, 제22권 제4호, 2004

30

여 백

제 3 차

전문가 초청 특강

김 성 목(한국 IBM 기술연구소 HCT 팀장)

여 백

Linguistic Tag for HLT

Nov. 12, 2004

Dr. Seong Mook Kim

**Human Language Technology
&
Country Unique Support**

IBM Korea Technology Lab.

Contents

- ☐ **Background**
- ☐ **Why are Linguistic Tags(LT) Needed?**
- ☐ **LT-support Tools**
- ☐ **LT examples**
- ☐ **Prospects**
- ☐ **Conclusion**

Background

- We have been confronted with various difficulties in the development of Human Language Technology systems.
- In addition to immaturity of HLT, there are some problems with existing tags.
- Sejong project has and will continue to compile a large Human Language/Linguistic Knowledge Database.

KTIL

IBM

Why are LT Needed?

Level of Language

- Phonetics/phonology : sound/phoneme/prosody
- Morphology : morpheme/lexical rules
- Syntax : phrase structure/grammatical features
- Semantics : word meaning/sentence meaning
- Pragmatics : context/deeper meaning
- Discourse and world knowledge : background knowledge

KTIL

IBM

Why are LT Needed?

Language Ambiguities

- St. Martin St. ; Dr. Huang Dr.
- 1 oz. / 2 oz. ; 1 km. / 2km.
- can(n) / can(v) ; saw(n) saw(v)
- bank(1) / bank(2) ; case(1) / case(2)
- learned(v) / learned(a) ; subject(n) / subject(v)
- school(n) / school(v) ; help(n) / help(v)
- eye / I ; made / maid ; knight / night
- She saw a man with a telescope.
- Flying planes can be dangerous.

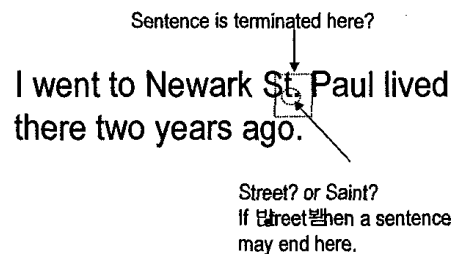
KT

IBM

Why are LT Needed?

Sentence Boundary:

A period is not necessarily a sentence breaker.



KT

IBM

Why are LT Needed?

Proper nouns:

- Bill Clinton
- Apple
- Java

Subject Area: (ex. Golf)

- putting
- eagle

Titles

Navigation Links

Headline

KTL

IBM

Why are LT Needed?

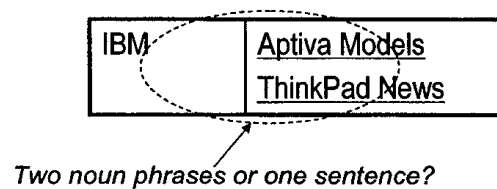
HTML
 Tag Problem:

*A
 tag does not necessarily a sentence terminator.*

HTML Source

```
<table>
<tr>
<td>IBM:</td>
<td><a href=뵁? Aptiva Models</a><br>
<a href=뵁? ThinkPad News</a>
</td>
</tr>
</table>
```

HTML Rendering Image



KTL

IBM

LT Improve HLT

- We need linguistic knowledge or knowledge of language.
- There are ambiguity problems to be dealt with properly, using current Human Language Technologies.
- HLT Programs are confronted with several scope-related problems:
 - Sentence scope
 - Segment scope to be processed



Linguistic Tags can solve these problems.

KTL

IBM

LT-support Tools

Parser:

- ESG (English parser by IBM Watson Research Center)
- KSG(Korean parser by IBM Korea Technology Lab.)
- ...

MT Engine:

- EtoF,G,I,S MT engines (by IBM Watson Research Center)
- EtoK MT engine (by IBM Korea Technology Lab.)
- ...

KTL

IBM

LT example

Scope:

Sentence:	<lal:s>This is a sentence.</lal:s> <lal:s type=발화? Sejong 2004</lal:s>
Phrase:	I have been to <lal:seg para=발화? Taipei and Beijing</lal:seg>.
Word:	I have two <lal:w pos=발화 발화 ket 발화 rs=발화? tickets</lal:w>.

KTL

IBM

LT example

Multiple Parses:

She <lal:w id=발화1? saw</lal:w> a <lal:w id=발화2? man</lal:w> with a <lal:w id=발화3? od=발화1:w1 p2:w2? telescope</lal:w>.
<i>parse identifiers</i>

Reference:

He bought a new <lal:w id=발화1? car</lal:w> yesterday. <lal:w ref=발화1? It</lal:w> was very expensive.

KTL

IBM

Prospect

Tag-based HLT:

- Wide-spread approach
- Standardization is needed.

Tag Editing Environment:

- Editing environment is mandatory.
- Infrequent user interventions are needed

LT-aware tools :

- Development tools must be LT-aware

KTL

IBM

Conclusion

Linguistic Tags (LT)

- XML compliant tag set for annotating linguistic knowledge

LT-aware Tools

- Parsers and MT engines have been developed.

LT-editor

- GUI-based easy-to-use editing environment for LT

IBM MT Experiment

- Promising results, but need more work

KTL

IBM

제 4 차

전문가 초청 특강

장 원 재(계명대학교)

여 백

한일 대역코퍼스로 본 한일어휘의 유사점과 차이점

장원재(계명대학교)

1. 들어가며

지금까지 한일 양국어의 어휘조사를 살펴보면 일본어는 주로 국립국어연구소가 여러 가지 분야의 언어작품을 달리하여 많은 대규모의 어휘조사를 내 놓고 있는 반면 한국어는 문교부(1956) 『우리말 말수 사용의 찾기 조사』을 시작으로 몇 가지 결과가 나와 있기는 하지만 일본어의 어휘조사에 비해 역사도 짧고 체계적인 조사가 아직 이뤄져 있지 않다¹⁾. 또한 조사단위도 일본어는 조사목적에 따라 다양하게 짧은 단위(M단위, β단위, 단단위 등)와 긴 단위(W단위, α단위, 장단위 등)를 설정하여 조사하고 있으나 한국어는 어절(文節)단위가 대부분이다. 물론 최근 코퍼스를 대상으로 형태소분석기로 조사한 연구결과(김홍규외(2000))가 있지만 조사단위의 규칙 등 미미한 점이 적지 않다.

한편 한일양국어 어휘에 관한 대조연구는 최근 들어 활발하게 조사²⁾되고 있어

1) 김광해(1993)pp99-103를 참조. 김광해(1993)이후 조사된 것으로는 김홍규외(2000)과 국립국어연구원(2002)가 있다.

2) 한국에서 발표된 일본어 어휘 관련 논문의 시대적 경향 (조사대상: 224편)

1965-69	70-74	75-79	80-84	85-89	90-94	95-99	00-04
3	3	4	11	17	28	58	100

다음 논문은 일본에서 연구된 논문으로 []은 논문의 경향을 나타낸 것이다. [어휘의 성격, 단어의 종류, 자료의 성격, 전체어수/개별어수]의 순이다.

· 中村暎枝 (1968) 「現代朝鮮語の辭典見出語における語彙の分布狀況」 『朝鮮學報』 49

[전어휘, 전어종, 사전, 개별]

· 李漢燮 (1984) 「日韓同形の漢字表記語彙」 『日本語学』 VOL3-8明治書院

[기본, 한자어, 어휘표, 개별]

· 조희철 (1991) 「日韓同形漢語の語義・用法の相違」 近代語研究会 『日本近代語研究』 1 ひつじ書房

[기본, 한자어, 어휘표, 개별]

· 宋永彬 (1993) 「『分類語彙表』による日韓基本語彙の対照」 『早稲田大学大学院文学研究科紀要 文学・芸術編』 別冊20

한일 각각의 언어적 특징들이 밝혀지고는 있지만 이것도 질적(조사단위, 조사자료, 조사방법)으로나 양적으로나 아직 충분하지 않다. 본 발표는 대역코퍼스를 자료로 단어의 운용면(전체어수)도 포함하여 한국어와 일본어 어휘(단어의 집합체)의 유사점과 차이점을 밝히는 것을 목표로 한다.

2. 조사자료

조사자료로서는 주된 자료로 한일대역신문([조선일보]의 한국어판과 일본어판)을 대상으로 한다. 대역자료를 조사대상으로 한 이유는 대역자료가 같은 내용을 한일 양국의 언어로 각각 표현된 것으로 조사자료의 내용에 의한 한일 간 어휘의 차이점을 배제할 수 있기 때문이다. 한편 신문을 자료로 한 것은 지금까지의 한일 어휘조사가 기본어휘(교과서 어휘)에 치중되어 있어 어휘의 주변적인 경향도 시야에 넣어야 할 필요성이 있기 때문이다.

대역코퍼스에는 1)일본어를 한국어로 2)한국어를 일본어로 3)제3국의 언어를 한일 각각의 언어로 번역된 것이 있다. 1)과 2)는 원본의 언어가 번역되어지는 언어에 영향을 끼치는 경향이 있기 때문에 3)의 번역데이터가 가장 바람직하나 현재 입수할 수 있는 코퍼스의 양과 질이 제한되어 있어 충분한 조사량을 확보할 수 없다. 그렇다면 1)과 2)중에서 조사자료를 선택하여야 하는데 이들 자료도 충분한 실정이 아니다. 대부분의 대역코퍼스는 영일 혹은 영한대역에 치중되어 있기 때문이다.

이런 코퍼스의 양과 질이 부족한 상황에서 2000년 1월 10일부터 조선일보 신문사는 자사 사이트에서 조선일보의 기사를 각 분야별로 나누어 각 분야 중에서 중요한 기사를 골라 일본어로 번역해 그 기사를 사이트에 게재하는 서비스를 개시하였다. 이는 현재까지 서비스가 계속되고 있어 향후 통시적인 조사 또한 가능하리라 본다. 각각의 데이터는 아래의 사이트에서 열람할 수 있다.

조선일본 한국어판→<http://www.chosun.com>

[기본, 고유어/한자어, 교과서, 개별]

- ・塩田雄大 (1999) 「日本・韓国・中国の専門用語－日本語とはどのくらい似ているか－」 『国文学解釈と鑑賞』 64-1至文堂

[전문, 한자어, 사전, 개별]

- ・申玟澈 (2001) 「日韓語彙の比較研究－「小学生基本語彙」を対象として－」 『比較語彙研究の試み』 (田島毓堂編、名古屋大学大学院国際開発研究科) 7

[기본, 전어종, 어휘표, 개별]

조선일보 일본어판→<http://japan.chosun.com>

현재 한일대역코퍼스 중에 이 정도의 대규모 코퍼스는 없기 때문에 본고의 조사자료로 한다.

보조자료로는 어린왕자 (La Petite Prince) 를 자료로 하여 신문자료에서 밝혀진 특징에 대해 검증한다. 中野洋 (1976) 의 6개 국어(일본어, 영어, 독일어, 프랑스어, 스페인어, 포르투갈어) 조사와 大河内康憲 (1990) 의 중국어 조사를 참고하고 한국어판(고려대학교 불문학과 황현산교수가 번역, <http://mycyberworld.net/>)은 발표자가 조사했다.

3. 조사방법

3.1 조사대상과 범위

대역신문의 조사대상 범위는 일본어 대역판을 서비스하기 시작한 2000년 1월 10일부터 2000년 7월 11일까지 6개월 간을 조사범위로 삼았다. 우선 이 기간 중에 일본어로 번역된 모든 기사를 일람표로 만들고 일련번호를 달았다. 추출한 단위는 한 개의 기사단위로 하고 추출비율은 1/20로 했다. 추출비율을 1/20로 한 이유는 예비조사를 통해 이 정도의 비율이라면 한일 각각 전체어수 약 12만 5천단어(조사, 어미, 부호 포함)로 합계 약 25만 단어로 예상되어 개인적인 조사로는 더 이상 조사하기 어렵다고 판단했기 때문이다. 그리고 추출단위는 기사내용에 좌우되지 않는 문(文)단위가 바람직하나 여러 가지 어려운 점이 있어 기사단위로 하였다. 본 조사는 용어용자조사(用語用字調査)가 아니고 같은 조건에서 한일 간의 언어적 차이점을 고찰하는 것이 목적이므로 어느 정도는 기사 내용에 좌우되어도 조사결과에는 큰 영향을 미치지 않을 것이라 판단한 것도 기사단위를 채용한 이유이기도 하다.

추출방법은 난수를 발생시켜 스타트번호를 매긴 후 20개 기사 간격으로 추출했다. 그 다음 일본어로 번역된 기사에 대응하는 한국어 기사를 찾았다. 이하(표1)은 모집단과 각 분야별 샘플수(비율)을 나타낸 것이다.

표1 조사대상의 모집단과 샘플수(단위: 기사)

기사	문화	경제	정치	스포츠	사회	논설	총계
모집단	319	852	854	165	517	345	3052
샘플	16	43	43	8	26	17	153
%	10.5	28.1	28.1	5.2	17.0	11.1	100

어린왕자의 조사범위는 [어린왕자] 제 1장만을 조사대상으로 한다. 그 이유는 中野洋 (1976) 의 조사범위가 제 1장에서 5장까지이지만, 6개국어판 동음이어(同音異語)판별작업은 제1장만을 했기 때문이다. 한국어와 정확한 비교를 위해서는 동음이어 판별작업이 필수적인 요소이다.

3.2 조사단위

신문자료의 조사단위로 국립국어연구소에서 채용한 β 단위³⁾를 따르기로 하고 한다. 단, 동음이어판별에 있어서는 『잡지90종1(3)』의 기준을 바꾼 것이 한 가지 있다 (「物」「者」(もの)). 『잡지90종1(3)』에서는 같은 음으로 한 개의 표제어로 판별하고 있지만, 한국어를 생각하면 다른 표제어로서 취급하는 것이 좋을 것이라 판단했기 때문이다. 한국어에서는 각각 다른 음이기 때문이다.

어린왕자의 조사단위는 中野洋 (1976) 가 채용한 α 단위를 이용했다.

4. 어휘량

지금까지의 선행연구(中野洋 (1976), 真田治子 (1992))에 의하면 일본어는 서양 언어에 비해 많은 단어가 필요하다고 한다(서양언어가 누적사용률이 높음). 그럼 한일 간에는 어떠한가?. 본 조사에서 한일 전체어수와 개별어수는 한국어(이하, K)가 각각 33647어, 4990어이고 일본어(이하, J)가 31399어, 4803어로 한국어가 단어수가 많다.

표2는 사용률순으로 정렬해서 각 개별어수의 누적사용률을 비교한 것이다. 이미 언급했듯이 전체 어휘량은 한국어보다 일본어가 적지만, 250단어까지 사용률이 높

3) 国立国語研究所 (1962) 『現代雑誌九十種の用語用字第一分冊 (三分冊)』 (이하 『잡지90종1(3)』)

은 고빈도 어군에서는 한국어의 누적사용률이 일본어의 그것보다 높다는 것을 알 수 있다⁴⁾.

표2 사용률순 개별어수의 한일누적사용률의 비교(단위:퍼밀)

단어수	J누적사용률	K누적사용률
50	277.1744	281.1209
100	358.7057	360.8214
250	496.7674	498.7370
500	622.1854	620.5164
1000	760.7567	754.7175
1500	836.6190	829.4850
2000	886.2065	879.2606
2500	918.0547	912.8703
3000	942.5459	940.4178
3500	958.4700	955.2762
4000	974.3940	970.1346
4500	990.3181	984.9930
4804	1000	994.0269

그럼 어린왕자의 데이터로 어휘량을 살펴보면 아래의 표3이다. 선행조사 7개국어와 한국어 조사를 포함한 8개국어 조사결과이다.

- 4) 참고로 국립국어연구소의 W단위(장단위에 속함)로 조사단위를 설정하여 한일간 어휘량을 비교한 결과 β단위의 결과와는 달랐다. 즉 고빈도 어군에서도 일본어가 한국어보다 누적사용률이 높다. 이와 같이 단위를 달리함으로써 초래하는 결과의 원인은 명확하게는 말할 수 없지만 사용률이 높은 접미사들은 적어도 이 현상에 영향을 끼치지 않았을까 라고 생각되어 진다. 왜냐하면 접미사의 경우 β단위에서는 「학생/들」 「선생님/들」 「아이/들」 처럼 조사단위로 나뉘어져 「들」 (3회)로 고빈도어가 되나 W단위에서는 어기(語基)와 접미사를 포함한 형태를 1단위로 설정하기 때문에 「학생들」 「선생님들」 「아이들」 각각 빈도수 1회로 저빈도어가 되기 때문이다. 본 조사에서의 빈도수는 한국어의 [들]은 293빈도수로 사용률순위 6위이나 일본어의 「たち」는 36빈도수 122위로 한일 양국간에 빈도수와 순위에 많은 차이점이 보인다.

표3 8개국어판 어린왕자의 어휘량

	한국어	일본어	중국어	영어	독일어	프랑스어	스페인어	포르투갈어
전체어수	642	601	505	500	415	457	403	350
전체어수(삭제)	362	323	392	387	325	353	306	272
개별어수	224	208	257	202	192	184	167	169
개별어수(삭제)	182	172	221	184	177	169	145	152

(삭제)는 조사, 조동사(어미)를 삭제한 결과를 의미

이 데이터에서도 전체어수 개별어수 또는 조사 조동사를 제외한 수치 모두 한국어보다 일본어가 적다(일본어가 누적사용률이 높음). 대역신문 전체 한일 어휘량의 경향과 부합한다. 그럼 사용률순 상위 100위까지 전체어수의 비율을 보면 다음과 같다.

표4 사용률순 상위100위가 전체어수를 커버하는 비율

	한국어	일본어	독일어	영어	프랑스어
상위100단어가 차지하는 비율	77.3(80.8)	61.6(75.6)	68	70.3	73.3

()안은 조사, 조동사를 포함한 수치

표4에서 알 수 있듯이 상위 100위까지의 단어가 전체를 커버할 수 있는 비율이 일본어에 비해 한국어가 높다. 이것도 대역신문 조사의 고빈도 어군의 경향과 일치한다.

田中章夫 (1996) 는 「일본어의 경우 사용률이 높은 단어군이 반복해서 사용되어 전체어수를 커버할 수 있는 비율이 외국어에 비해 매우 높다」라고 中野 (1976) 의 결과인 표4(한국어 제외)를 증거로 언급하고 있는데 여기서 외국어라 하면 서양어를 의미하는 것으로 한국어와 대조해 보면 그 양상은 다르다. 지금까지 한일어휘 대조연구가 이루어지지 않은 결과이다. 한국어와 대조를 통해 일본어 어휘의 특징이 좀 더 명확해 질 수 있다는 것을 단적으로 말해주고 있다.

5. 품사구성

5.1 대역신문

품사분류 기준은 국립국어연구소 (1964) 『国立国語研究所資料6 分類語彙表』

秀英出版 (p.4) 를 따르기로 한다⁵⁾. 1단어에 2개 이상의 품사가 있을 경우는 「잡지90종1」에 따랐다.

표5와 표6에서 알 수 있듯이 한일 품사구성은 전체적으로 큰 차이는 보이지 않는다. 단, 미미하기는 하지만 用言類는 일본어가 높고 相言類는 한국어가 높다. 이것은 어린왕자의 데이터(그림 1)에서도 宋永彬 (1993) 의 교과서 조사(用言-J:30.1%, K:28.1%, 相言類-J:12.8%, K:13.5%)에서도 같은 경향을 보여주고 있다. 그리고 빈도 구간과 품사와의 관계도 거의 같은 경향을 띄고 있다.

5.2 어린왕자

대역신문 조사에서는 품사분류를 하지 않았지만, [어린왕자] 데이터에서는 대명사의 비율에 한일 간 차이점이 보인다(그림1). 개별어수에서는 한일 간 차이점이 없지만, 전체어수에서는 한국어가 일본어보다 많다. 바꿔 말하면 한일 간 대명사의 종류는 비슷하지만, 대명사의 운용면에서는 차이점이 있다는 것이다.

다음은 한국어판에 나타나는 대명사의 출현상황을 나타낸 것이다.

나 (27:4.21) 그 (4:1.11) 그것 (2:0.55) 아무것 (1:0.28)

(빈도수 : 사용률)

대명사 중에 가장 빈도가 높은 [나]에 대해 살펴보면 일본어가 13회에 비해 한국어의 [나]는 27회 나타난다. 한국어판에는 출현하고 일본어판에는 나타나지 않는 부분을 보면 8개의 예가 주격에 위치하는 곳이다. 이것은 일본어의 주어가 생략하는 특징을 반영하고 있는 것으로 파악된다.

J : おとなの人たちに、そういわれて、こんどは、これなら、なるほどとわかってくれるだろう、と思って、ウワバミのなかみをかいてみました(어른들에게 그렇게 말해져 이번은 이것이라면 잘 알 수 있을 것이라고 생각해 보아뱀의 속을 그려보았다).

5) 품사의 명칭은 国立国語研究所 (1964) 『分類語彙表』秀英出版에 따라 体言類(명사, 대명사, 수사), 用言類(동사), 相言類(형용사, 형용동사, 일부분의 부사, 연체사), その他(접속사, 감탄사, 일부분의 부사)로 집계하고 분석하였다. 본고에서는 품사명을 한국어로 번역하기 곤란한 점이 있으므로 일본어 명칭 그대로 사용하는 것으로 한다.

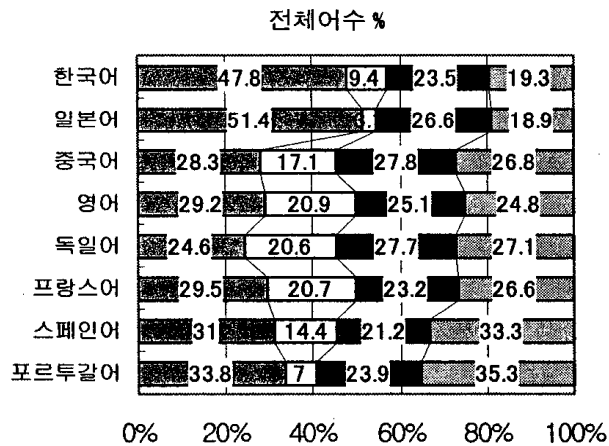
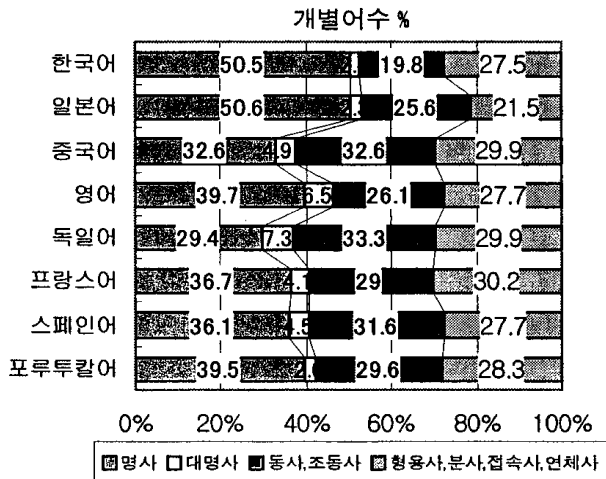
K : 그래서 나는 어른들이 알아볼 수 있도록 보아뱀의 속을 그렸다.

J : そして、世界じゅうを、たいてい、どこもとびあるきました(그리고 온 세상을 대부분 어디라도 날아다녔다).

K : 나는 세계의 여기저기 제법 많은 곳을 날아다녔다.

()필자 번역(가능한 한 직역)

그림1 8개국어판 [어린왕자]의 품사구성



이 외에 한일 간 품사구성에 영향을 끼칠 가능성이 있는 예는 아래와 같다.

혼종어 相言類

한국어 : 18例

間間이 (時々) 强하다 (強い) 決코 (決して) 貴하다 (貴い) 그間
(その間) 글러벌化 (グローバル化) 急하다 (急ぐ) 別다르다 (特に變
わっている) 數없다 (多い) 實로 (実に) 甚하다 (甚だしい) 弱하다
(弱い) 全혀 (全く) 正말 (本当) 眞짜 (本物) 限없다 (限りない)
幸여 (若しかしたら) 凶하다 (醜い)

일본어 : 4例

だめ 愛くるしい 決して むやみ

혼종어 用言類

한국어 : 36例

加하다 (加える) 功들다 (念が入る) 關하다 (関する) 求하다 (求め
る) 勸하다 (勧める) 達하다 (達する) 當하다 (当面する) 對하다
(對する) 亡하다 (滅びる) 免하다 (免ずる) 番갈다 (順番をかえる)
變하다 (変わる) 比하다 (比する) 傷하다 (傷つく) 屬하다 (属する)
要하다 (要する) 願하다 (願う) 爲하다 (為にする) 應하다 (応じる)
依하다 (依る) 因하다 (因る) 轉하다 (転じる) 接하다 (接する) 定하
다 (定める) 處하다 (処する) 取하다 (取る) 漆하다 (色を塗る)
通하다 (通じる) 避하다 (避ける) 限하다 (限る) 合치다 (合わ
せる) 合하다 (合わせる、混ぜる) 害치다 (害する) 行하다 (行な
う) 向하다 (向かう) 欠잡다 (けちをつける)

일본어 : 22例

對する 通じる 達する 転じる 脱する 属する 愛し合う 位置付ける 課
する 辞する 祝する 称する 処する 制する 接する 呈する 適する 瀕
する 封切る 有する 要する 欲張る

연체사

일본어	한국어
この (158)	이 (227)
その (62)	그 (96)

복합동사

한일간 대응되는 예

[呼び集める-불러모은다.]

J : まだ多くの漁船を浜へ引き上げる掛け声で騒がした.

K : 아직도 많은 어선들을 모래사장으로 끌어올리는 소리로 떠들썩했다.

[聞き流す-흘려듣다, 着替える-갈아입다.]

J : 何かいい考えはないものかと、当たりを見回した.

K : 무슨 좋은 생각이 없을까하고 둘레를 둘러보았다.

한일간 대응하지 않는 예

[読み終える- 다 읽다. 求めつづける-계속 구하다.]

J : 久しぶりにあまり歩きすぎたのがいけなかった.

K : 오랜만에 너무 많이 걸은 것이 잘못되었던 것이다.

J : とにかく読み出したら最後まで読むというのは、うまい読書法の一つだと思ふ.

K : 아무튼 읽기 시작하면 끝까지 읽는 것이 좋은 독서법의 하나라고 생각한다.

6. 어종구성

한일 간 어종구성은 표5와 표6에서 알 수 있듯이 전체어수 개별어수 모두 [일본어는 고유어와 외래어의 비율이 높고 한국어는 한자어와 혼종어의 비율이 높]다. 표7은 일본어의 고유어와 외래어가 한국어의 한자어와 혼종어로 대응하는 예이다. ②는 한국어에 고유어가 있지만 빈도수에서 알 수 있듯이 거의 한자어로 대응한다. 또한 ③은 한자어 [비교]도 있지만, 혼종어[비하다]에 대응하는 것이 보통이다. ⑤⑥은 20세기 초기에 일본(고유어)에서 한국에 들어와 한자어가 된 예이고, ⑧⑨

와 같이 부사나 관용적인 표현도 한자어가 사용되고 있다. ⑩⑪⑫는 한국어에도 외래어가 있으나 언어생활에서 외래어보다는 한자어가 많이 쓰여지고 있다(신문 데이터에서는 외래어가 0예).

표7 일본어에 대응하는 한국어의 번역어

日本語	韓国語
①答える	對答/하다
②進む 21 (進行 5)	進行/하다 23 (나아가다 1)
③比べる 14 (比較 3, 比する 0)	比較/하다 1, 比하다 15
④認める 12 (認定 4)	認定/하다 15
⑤売り切れ	賣盡
⑥引き下げ	引下
⑦封切り	開封
⑧すぐ	當場
⑨目を引く	關心/을 끌다
⑩プレッシャー	負擔
⑪メイク	化粧
⑫スト	罷業
⑬伝える	傳하다
⑭急ぐ	急하다

그럼 조금 전에 언급한 한일 간 어종구성의 특징, [일본어는 고유어와 외래어의 비율이 높고 한국어는 한자어와 혼종어의 비율이 높]다는 주제로 되돌아가 생각해 보도록 한다. 이 경향을 빈도구간별로 나누어 살펴보면 지금까지의 양상과는 다르다.

표8-1 빈도수별 어종구성(전체어수)

度数		固		漢		混		外		總計
1-32	K	3951	20.6%	14050	73.3%	265	1.4%	899	4.7%	19165
	J	4719	25.2%	12669	67.7%	233	1.2%	1098	5.9%	18719
33-	K	6928	47.8%	6893	47.6%	321	2.2%	340	2.4%	14482
	J	5674	44.7%	6418	50.6%	104	0.8%	484	3.8%	12680

표8-2 빈도수별 어종구성(개별어수)

度数		固		漢		混		外		總計
1-32	K	1080	22.4%	3321	68.8%	104	2.2%	322	6.7%	4827
	J	1214	26.0%	2967	63.6%	79	1.7%	404	8.7%	4664
33-	K	52	31.9%	102	62.6%	4	2.5%	5	3.1%	163
	J	36	25.9%	96	69.1%	1	0.7%	6	4.3%	139

표8-1과 8-2를 정리해 보면 전체어수 개별어수 모두 다음과 같은 경향이다.

	固	漢	混	外
전체	J	K	K	J
빈도수低	J	K	—	J
빈도수高	K	J	—	—

7. 마치며

1. 고빈도 어군의 한일 어휘량은 한국어가 일본어보다 적은 단어로 일정 범위를 커버하는 경향이 있다.
2. 한일 간의 품사구성 차이점이 어종구성에 비해 많은 차이점은 보이지 않고 매우 유사하다고 말할 수 있다. 단, 세부적으로 품사별로 나누어 조사한다면 한일간 품사구성의 차이점이 부각될 것으로 예상된다.
3. 한편 어종구성은 일본어는 고유어와 외래어의 비율이 높고 한국어는 한자어

와 혼종어의 비율이 높다. 그런데 고빈도어군에서는 고유어와 한자어의 비율이 한일간 역전현상이 보인다. 어휘량의 분포경향과 유사한 결과이다.

참고문헌

국립국어연구원편(2002) 『현대 국어 사용 빈도조사-한국어학습용 어휘선정을 위한 기초조사-』

김광해(1993) 『국어어휘론 개설』 집문당

김홍규외(2000) 『한국어형태소 및 어휘사용빈도의 분석』 고려대학교 민족문화연구원

大河内康憲 (1990) 「日本語と中国語の語彙の対照」 『講座日本語と日本語教育7日本語の語彙・意味 (下)』 明治書院

真田治子 (1992) 「諸言語の基本語彙の有効性の比較」 『計量国語学』 18-7

田中章夫 (1996) 「基本語彙の選定」 『国文学解釈と教材の研究』 41-11

中野洋 (1976) 「『星の王子さま』6か国語版の語彙論的研究」 『計量国語学』 79

표5-1 일본어의 품사와 어종구성(전체어수)

	固		漢		混		外		總計	
体言類	3098	13.7	17836	78.6	175	0.7	1574	6.9	22683	100
		29.9		93.4		50.0		99.5		72.2
用言類	5720	97.2	0	0.0	165	2.8	0	0.0	5885	100
		55.2		0.0		47.1		0.0		18.7
相言類	1421	52.8	1256	46.6	9	0.3	8	0.3	2694	100
		13.7		6.6		2.6		0.5		8.6
その他	133	97.1	3	2.2	1	0.7	0	0.0	137	100
		1.3		0.0		0.3		0.0		0.4
總計	10372	33.0	19095	60.8	350	1.1	1582	5.0	31399	100
		100		100		100		100		100

표5-2 한국어의 품사와 어종구성(전체어수)

	固		漢		混		外		總計	
体言類	3066	12.7	19732	81.6	142	0.6	1236	5.1	24176	100
		28.2		94.2		24.2		99.8		71.9
用言類	5585	93.3	0	0.0	403	6.7	0	0.0	5988	100
		51.3		0.0		68.8		0.0		17.8
相言類	2087	62.7	1201	36.1	36	1.1	3	0.1	3327	100
		19.2		5.7		6.1		0.2		9.9
その他	141	90.4	10	6.4	5	3.2	0	0.0	156	100
		1.3		0.0		0.9		0.0		0.5
總計	10879	32.3	20943	62.2	586	1.7	1239	3.7	33647	100
		100		100		100		100		100

표6-1 일본어의 품사와 어종구성(개별어수)

	固		漢		混		外		總計	
体言類	484	12.9	2822	75.0	53	1.4	404	10.7	3763	100
		38.7		92.1		66.3		98.5		78.4
用言類	533	96.0	0	0.0	22	4.0	0	0.0	555	100
		42.6		0.0		27.5		0.0		11.6
相言類	212	45.9	240	52.0	4	0.9	6	1.3	462	100
		17.0		7.8		5.0		1.5		9.6
その他	21	91.3	1	4.4	1	4.4	0	0.0	23	100
		1.7		0.0		1.3		0.0		0.5
總計	1250	26.0	3063	63.8	80	1.7	410	8.5	4803	100
		100		100		100		100		100

표6-2 한국어의 품사와 어종구성(개별어수)

	固		漢		混		外		總計	
体言類	435	11.0	3145	79.5	51	1.3	324	8.2	3955	100
		38.4		91.9		47.2		99.1		79.3
用言類	441	92.5	0	0.0	36	7.6	0	0.0	477	100
		39.0		0.0		33.3		0.0		9.6
相言類	241	45.1	273	51.0	18	3.4	3	0.6	535	100
		21.3		8.0		16.7		0.9		10.7
その他	15	65.2	5	21.7	3	13.0	0	0.0	23	100
		1.3		0.2		2.8		0.0		0.5
總計	1132	22.7	3423	68.6	108	2.2	327	6.6	4990	100
		100		100		100		100		100

21세기 세종계획 국어 특수 자료 구축

2004년 12월 인쇄

2004년 12월 발행

발행자: 문화관광부
서울시 종로구 세종로 82-1
전화: 3704-9437

인쇄처: 학위사
전화: 719-9567

ISBN: 89-7820-125-3 93710

<비매품>