

국립국어원 2015-01-41

발간등록번호

11-1371028-000580-01

2015년 한국어 학습자 말뭉치 구축 지원 도구 개발 연구

연구 책임자

곽 용 진



국립국어원

제 출 문

국립국어원장 귀하

“2015년 한국어 학습자 말뭉치 구축 지원 도구 개발 연구”에
관하여 귀 원과 체결한 연구용역 계약에 의하여 연구보고서를
작성하여 제출합니다.

2015년 12월 11일

연구 책임자: 곽용진((주)아이와즈)

연구 기관 (주)아이와즈

연구 책임자 곽용진((주)아이와즈)

공동 연구원 박진호(서울대), 김한샘(연세대), 유현조(서울대),
이화진(홍익대), 권택수((주)아이와즈), 오지훈((주)
아이와즈), 황용주(국립국어원)

연구 보조원 이지연(연세대), 최지명(연세대)

2015년 한국어 학습자 말뭉치 구축 지원 도구 개발 연구

본 연구의 목적은 대규모 한국어 학습자 말뭉치 자료를 효과적으로 수집, 관리, 보급, 활용하기 위한 한국어 학습자 말뭉치 기초 연구 및 구축 사업에 필요한 지원 도구의 개발 방안과 그 실례를 제시하는 데 있다. 이를 위하여 한국어 교육 관련 기관 및 교강사, 학습자에 대한 한국어 학습자 말뭉치 구축 방안에 대한 설문 조사, 국내외 유사 서비스 분석, 한국어 학습자 말뭉치 구축 실무자들의 요구 사항 분석을 통해, 사용자 접근성이 우수한 말뭉치 구축, 관리, 배포, 활용 시스템을 설계하고, 목표 시스템의 중장기적 개발 이행을 수립하였으며, 당해 연도 한국어 학습자 말뭉치 구축 지원을 위한 웹페이지와 도구를 개발하였다. 이와 함께 향후 개발에 참고가 될 수 있도록 국내외 공개 및 상용 프로그램을 이용한 말뭉치 활용법과 한국어 학습자 말뭉치의 구어 전사에 효과적인 도구 개발을 위한 기초 연구를 추가적으로 제시하였다.

본 연구에서는 대규모 한국어 학습자 말뭉치 자료를 효과적으로 수집, 관리, 보급, 활용하기 위한 소프트웨어의 개발과 활용을 위하여 다음과 같은 내용을 제시하였다.

우선, 한국어 학습자 말뭉치의 자료 구축과 이를 지원하기 위한 도구가 함께 개발됨을 고려하여, 목표 시스템이 한국어 학습자 말뭉치의 구축 상황에 따라 적절한 지원이 가능하도록 점진적인 소프트웨어 개발 방법론을 중심으로 한 연구 방법을 제안하였다. 이와 함께 중장기적인 한국어 학습자 말뭉치 수집, 관리, 보급, 활용을 위한 시스템 개발 방안을 연구 범위에 포함하였다.

한국어 학습자 말뭉치 통합 관리 시스템 설계 및 이행안은 한국어 학습자 말뭉치의 구축 실무진의 요구와 국내외의 한국어 교육 기관 및 교사, 강사, 학습자들의 의견을 수렴한 설문 조사, 국내외의 학습자 말뭉치 구축 사례 및 서비스 현황 분석을 토대로, 대규모 한국어 학습자 말뭉치 자료를 효과적으로 수집, 관리, 보급, 활용할 수 있는 시스템을 설계하고, 중장기적인 개발 이행을 제시하였다.

이를 토대로, 당면한 한국어 학습자 말뭉치의 구축에 우선적으로 요구되는 한국어 학습자 말뭉치 구축 지원 도구를 시험적으로 개발하였다. 본 연구에서 구현한 한국어 학습자 말뭉치 구축 지원 도구는 한국어 학습자 말뭉치의 메타데이터를 중심으로 동작하는 웹기반 소프트웨어와 한국어 학습자 말뭉치의 가공과 검수를 중심으로 하는 PC기반 소프트웨어로 구현되었다. 이 소프트웨어들은 한국어 학습자의 원자료(작문 이미지 데이터와 발화 음성 데이터)와 원자료의 메타데이터(학습자와 수집, 구축기관의 정보), 주석 정보(마크업, 형태주석, 오류주석 등)를 분리하고, 추상화한 데이터 구조에서 연결함으로써 한국어 학습자 말뭉치 데이터의 호환성과 일관성은 높이고, 수작업과 후처리로 인한 오류는 낮추도록 하였다. 특히 최근 기계 가독형 말뭉치에서 사용도가 높아지고 있는 외부 주석(Stand-off Annotation)을 적용하고, 본문 마크업, 전사 마크업, 헤더 정보, 형태주석 등 각기 다른 구조와 형태를 가진 표지들을 통합하여 일관되게 표현함으로써 다양한 사용자의 요구에 따라 말뭉치를 활용할 수 있도록 하였다.

마지막으로, 말뭉치 구축에서 가장 많은 노력이 소요되는 구어 말뭉치 구축의 부담을 줄이기 위해 한국어 학습자 말뭉치의 구어 전사 도구의 개발에 대한 기초 연구도 함께 수행하였다. 구어 전사 특유의 문제와 언어 학습자에 의한 발화라는 특성이 결합되어 발생하는 문제를 확인하였다. 한국어 학습자 말뭉치의 구어 전사 도구는 한국어 학습자의 음성적 특성을 고려해야 할 뿐만 아니라, 최소한 음절 수준에서라도 발음 그대로 인식할 수 있도록 만들어져야 함을 제안하였다. 또한 앞서 한국어 학습자 말뭉치 통합 관리 시스템에서 검토한 주석 체계를 구어 전사 표지에도 적용할 것을 함께 제안하였다.

주요어 : 학습자 말뭉치, 학습자 말뭉치 구축 지원 도구, 학습자 말뭉치 구축, 학습자 말뭉치 구축 관리, 학습자 말뭉치 구축 가공, 학습자 말뭉치 구축 배포, 학습자 말뭉치 구축 활용, 학습자 말뭉치 수집, 학습자 말뭉치 보급, 학습자 말뭉치 통합 관리 시스템

차 례

I. 연구 개요	1
1. 연구의 목적 및 필요성	1
1.1. 연구의 목적	2
1.2. 연구의 필요성	2
2. 연구의 방향	4
3. 연구 범위	5
4. 연구 방법	6
5. 연구 추진 일정	7
5.1. 연구 수행 단계	7
5.2. 연구 수행 절차	7
6. 연구 결과	9
6.1. 주요 결과물 개요	9
6.2. 주요 기능 구현 예시	10
 II. 한국어 학습자 말뭉치 통합 관리 시스템 설계 및 이행안	 16
1. 요구 분석 및 수요 조사	18
1.1. 요구 분석 개요	18
1.2. 요구 사항 수집	18
2. 유사 서비스 분석	22
2.1. 국내외 학습자 말뭉치 현황	22
2.2. 학습자 말뭉치 웹 서비스 사례	24
2.3. 말뭉치 웹 서비스 기능 및 요구 분석	59
2.3.1. 말뭉치 탐색 서비스	59
2.3.2. 어휘 및 오류 분포 탐색 서비스	59
2.3.3. 용례 검색 서비스	60
2.3.4. 공기 관계 분석 서비스	60
2.3.5. 기타 응용 서비스	61
3. 한국어 학습자 말뭉치 통합 관리 시스템 설계	62
3.1. 유스케이스 분석	62

3.2. 한국어 학습자 말뭉치 통합 관리 시스템 아키텍처	63
3.2.1. 소프트웨어 아키텍처	64
3.2.2. 한국어 학습자 말뭉치 통합 관리 시스템 구성	66
3.3. 주요 기능 목록	71
4. 한국어 학습자 말뭉치 통합 관리 시스템 개발 이행안	73
4.1. 개요	73
4.1.1. 목표	73
4.1.2. 목표 시스템 개발 추진 전략	73
4.2. 시스템 분석	74
4.2.1. 목표 시스템	74
4.2.2. 하드웨어 구성	75
4.2.3. 소프트웨어 구성	76
4.3. 기능별 분석	77
4.3.1. 수집	77
4.3.2. 관리	79
4.3.3. 가공	80
4.3.4. 분석	83
4.3.5. 배포 및 활용 서비스	84
4.4. 학습자 말뭉치 통합 관리 시스템 개발 로드맵	85

Ⅲ. 한국어 학습자 말뭉치 구축 지원 도구 개발 86

1. 한국어 학습자 말뭉치 웹페이지 개발	86
1.1. 표본 수집	87
1.1.1. 파일 업로드	87
1.1.2. 학습자 정보 검색	88
1.1.3. 파일 정보 입력	89
1.1.4. 파일명 자동 생성	91
1.1.5. 표본 파일 저장	92
1.2. 관리	94
1.2.1. 표본 관리	94
1.2.2. 주석 관리	105
1.3. 검색	111
1.4. 배포	112

1.5. 데이터 스키마	115
1.6. 클래스 다이어그램	116
2. 한국어 학습자 말뭉치 주석 편집 도구 개발	117
2.1. 작업 관리	117
2.1.1. 작업 목록 조회	117
2.1.2. 작업 유형별 작업창 할당	119
2.2. 주석 유형별 가공 작업	121
2.2.1. 오류 주석 작업 도구	121
2.2.2. 오류 주석 편집 작업	123

IV. 한국어 학습자 말뭉치 활용 방안 139

1. 공개/상용 소프트웨어를 이용한 활용 방안	139
1.1. 학습자 말뭉치의 구조	139
1.2. 엑셀을 이용한 말뭉치 활용	140
1.2.1. 정렬	140
1.2.2. 필터	143
1.2.3. 수식	145
1.2.4. 함수	147
1.2.5. 필터와 정렬의 결합 활용	149
1.2.6. 복합 기준에 의한 빈도 추출	151
1.3. 한마루 2.0을 이용한 말뭉치 활용	153
1.3.1. 한마루 사용을 위한 말뭉치 파일 코드 변환: UniConv	153
1.3.2. 말뭉치 열기	157
1.3.3. 말뭉치 통계	158
1.3.4. 용례 기본 검색	159
1.3.5. 아무개 문자 검색	160
1.3.6. 자소 검색	161
1.3.7. 범위 검색	161
1.3.8. 불리언 검색	162
1.3.9. 형태 분석 말뭉치	163
1.4. 지능형 형태소 분석기를 이용한 말뭉치 활용	166
1.5. AntConc를 이용한 말뭉치 활용	168
1.5.1. 말뭉치 파일 탑재를 위한 사전 작업: 인코딩과 파일 이름	168

1.5.2. 단순 용례 검색	170
1.5.3. 어절의 일부 검색	171
1.5.4. 검색 결과의 정렬	171
1.5.5. 아무개 문자	172
1.5.6. 정규표현	174
1.5.7. Concordance Plot	175
1.5.8. 단어 빈도 목록	175
1.5.9. 연어/공기 분석	178
1.5.10. N-Gram 분석	179
1.5.11. 키워드 분석	180
1.6. Cygwin을 이용한 말뭉치 활용	180
1.6.1. Cygwin 설치	180
1.6.2. Cygwin 사용을 위한 설정	181
1.6.3. gawk로 특정 필드 추출하기	183
1.6.4. redirection으로 출력을 파일로 저장하기	184
1.6.5. sort	184
1.6.6. uniq	184
1.6.7. sort의 옵션들	185
1.6.8. 파이프를 여러 명령어를 연결하기	187
1.6.9. 특정 필드에 조건을 지정하여 조건을 충족하는 레코드 추출하기	188
1.6.10. sort에서 기준 필드 지정하기	189
1.6.11. uniq -c를 이용한 통계 추출	190

V. 한국어 학습자 말뭉치 구어 전사 지원 도구 개발 기초 연구..... 192

1. 연구 개요	192
1.1. 연구의 목적 및 필요성	192
1.1.1. 연구의 목적	192
1.1.2. 연구의 필요성	192
1.2. 연구의 방향	194
2. 도구 개발을 위한 기초 연구	195
2.1. 음성 인식 및 전사 도구	195
2.1.1. 공개/상용 음성 인식 도구	195
2.1.2. 음성 전사 및 분석 도구	198

2.2. 국내외 음성 인식 적용 사례	200
3. 세부 연구 내용	208
3.1. 한국어 학습자의 오류 특성	208
3.1.1. 중국어권	210
3.1.2. 일본어권	212
3.1.3. 베트남어권	214
3.1.4. 스페인어권	215
3.1.5. 영어권	217
3.2. 구어 전사 도구 개발을 위한 실험	218
3.2.1. 실험 목적	218
3.2.2. 실험 방법	219
3.2.3. 실험 결과	220
3.2.4. 결론	229
4. 구어 전사 도구 개발안	230
4.1. 음성 인식을 개선 방안	231
4.2. 전사 표지 개선안	232
4.3. 결론	238

VI. 한국어 학습자 말뭉치 구축 및 지원 도구 개발 설문 조사..... 239

<부록 1> 한국어 학습자 말뭉치 구축 및 활용 도구에 관한 설문지	243
<부록 2> 한국어 학습자 음성 인식 실험지	249
<참고문헌>	250
<Abstract>	256

표 차례

<표 1> 연구의 범위 및 세부 과업 내용	5
<표 2> 주요 결과물 개요	9
<표 3> 학습자 말뭉치 표본 등록 화면 예시	10
<표 4> 회원/학습자 정보 등록 화면 예시	11
<표 5> 파일(표본) 관리 화면 예시	12
<표 6> 주석 관리 업무 흐름 예시	13
<표 7> 주석 작업 UI 화면 예시	14
<표 8> 형식 오류 검사 및 작업 UI 예시	15
<표 9> 요구 분석 명세서	18
<표 10> 사용자 정의	62
<표 11> 기능 명세서	71
<표 12> 학습자 말뭉치 통합 관리 시스템 개발 로드맵	85
<표 13> 한국어 학습자 말뭉치 웹페이지 화면 목록	86
<표 14> 국내외 음성인식용 공개/상용 도구	195
<표 15> 외국인 학습자를 대상으로 ASR 소프트웨어를 활용했을 때의 특성	206
<표 16> 한국어 학습자 말뭉치의 언어권별 수집 비율	208
<표 17> 실험 대상	219
<표 18> 음성 자료	219
<표 19> 중국어권 학습자 iOS 음성 인식 테스트 결과	221
<표 20> 일본어권 학습자 iOS 음성 인식 테스트 결과	222
<표 21> 중국어권 학습자 Android 음성 인식 테스트 결과	223
<표 22> 일본어권 학습자 Android 음성 인식 테스트 결과	224
<표 23> 모국어 화자의 음성 인식 테스트 결과	225
<표 24> 음성 인식 불가 단어 목록	228
<표 25> 오인식 단어 목록	231
<표 26> MICASE 정보	232
<표 27> MICASE 전사 기호	233
<표 28> 전사 기호 개선안	234
<표 29> 주석의 기능/역할에 따른 구조적 유형	237

그림 차례

<그림 1> 연구 개요	1
<그림 2> 연구의 방향	4
<그림 3> 연구의 방법 및 절차	6
<그림 4> 한국어 학습자 말뭉치 구축 지원 도구 개발 연구 수행 절차	7
<그림 5> 한국어 학습자 말뭉치 통합 관리 시스템 설계	16
<그림 6> 시스템 설계 및 개발 계획 수행 방법 및 절차	17
<그림 7> 영어 말뭉치 언어학 센터에서 제공하는 전 세계의 학습자 말뭉치 목록	22
<그림 8> MICASE 온라인 탐색 도구	26
<그림 9> MICASE 전사 자료 탐색 목록	27
<그림 10> MICASE 전사 자료 HTML 보기 예제	27
<그림 11> MICASE 전사 자료 XML 예시	28
<그림 12> MICASE 검색 서비스 화면	29
<그림 13> MICASE 검색 결과	30
<그림 14> MICASE 검색 결과에서 개발 용례의 “view” 결과	30
<그림 15> MICASE 검색 결과 통계 요약 정보	31
<그림 16> MICUSP 말뭉치의 전공 분야별 텍스트 개수와 글쓰기 유형별 텍스트 비율	32
<그림 17> MICASE와 MICUSP의 웹 탐색 및 검색 서비스	34
<그림 18> MICUSP 개별 텍스트	35
<그림 19> MICUSP 검색 기능	36
<그림 20> Lextutor(Compleat Lexical Tutor) 웹사이트의 첫 화면	37
<그림 21> Lextutor의 Group Lex 서비스 예시	38
<그림 22> Lextutor 사이트의 Group Lex 서비스 중 퀴즈 1단계	38
<그림 23> Lextutor 사이트의 Group Lex 서비스 중 퀴즈 2단계	39
<그림 24> Lextutor 사이트의 Corpus Grammar	40
<그림 25> Lextutor 사이트의 Concord Writer	41
<그림 26> Lextutor 사이트 Concordance 서비스에서 제공되는 말뭉치 목록	42
<그림 27> Lextutor의 문맥 색인 검색 화면	42
<그림 28> Lextutor의 문맥 색인 검색 결과	43
<그림 29> Lextutor 사이트의 Concordance 서비스 연어 후보 목록	44
<그림 30> Lextutor 사이트의 Range 서비스	45

<그림 31> Lextutor 사이트의 Range 서비스 검색 결과	45
<그림 32> Lextutor 사이트의 N-Gram 서비스	46
<그림 33> Lextutor 사이트의 N-Gram 서비스 분석 결과	47
<그림 34> Lextutor 사이트의 Frequency 서비스	48
<그림 35> Lextutor 사이트의 Text Lex Compare 서비스	48
<그림 36> Lextutor 사이트의 Text Lex Compare 분석 결과	49
<그림 37> Lextutor 사이트의 KeyWords 서비스	50
<그림 38> Lextutor 사이트의 KeyWords 서비스의 분석 결과	50
<그림 39> 체코국가말뭉치 사이트 첫 화면	51
<그림 40> 체코국가말뭉치에서 제공하는 말뭉치 목록	52
<그림 41> 체코국가말뭉치의 KonText 검색 서비스	52
<그림 42> CZESL의 용례 검색 서비스	53
<그림 43> 체코국가말뭉치 KWords 분석 결과의 예	54
<그림 44> 체코국가말뭉치 병렬 말뭉치 검색 결과	54
<그림 45> 지난 대학 학습자 말뭉치 사이트의 용례 검색 화면	56
<그림 46> Romance Phonetic Database의 검색 화면	58
<그림 47> 사용자 유형 분석	62
<그림 48> 한국어 학습자 말뭉치 통합 관리 시스템 개요도	63
<그림 49> 한국어 학습자 말뭉치 통합 관리 시스템 S/W 아키텍처	64
<그림 50> 1,2단계 한국어 학습자 말뭉치 통합 관리 시스템 H/W 구성	75
<그림 51> 정부통합전산센터 G-클라우드 구성(KLCMS 3단계 H/W 구성 시 적용)	75
<그림 52> 1,2단계 한국어 학습자 말뭉치 통합 관리 시스템 S/W 구성	76
<그림 53> 정부통합전산센터 G-클라우드 구성(KLCMS 3단계 S/W 구성 시 적용)	77
<그림 54> 학습자 말뭉치 구조	139
<그림 55> 학습자 말뭉치를 엑셀에서 열었을 때의 모습	140
<그림 56> 정렬 기준 필드 선택	141
<그림 57> 데이터-정렬	141
<그림 58> 정렬 시 팝업 메뉴	141
<그림 59> ‘오류 총위’를 기준으로 정렬한 결과	142
<그림 60> 필터 만들기	143
<그림 61> 필터 설정 1	143
<그림 62> 필터 설정 2	144
<그림 63> 필터 적용 결과	144

<그림 64> 필터 해제	145
<그림 65> 학습자 오류 현황	145
<그림 66> 수식 입력	146
<그림 67> 수식 계산 결과	146
<그림 68> 합계 함수 적용 방법	147
<그림 69> 합계 함수 적용 결과	147
<그림 70> COUNTIF 함수로 GPT의 빈도 세기	148
<그림 71> COUNTIF 함수 적용 결과	148
<그림 72> '오류 층위' 필드에 필터를 적용하여 'GPT'만 걸러낸 결과	149
<그림 73> 정렬	150
<그림 74> 정렬 기준 설정	150
<그림 75> 필드 G, 필드 D를 기준으로 정렬한 결과	151
<그림 76> COUNTIFS 함수 사용 방법	152
<그림 77> 3가지 기준을 사용한 COUNTIFS 함수 적용 결과	152
<그림 78> 언어정보나눔터 초기 화면	153
<그림 79> ls와 gawk를 이용하여 배치 파일 만들기: 원시 말뭉치	154
<그림 80> ls gawk의 출력 결과 배치 파일	155
<그림 81> 줄바꿈 문자를 Unix식(LF)에서 Dos식(CR-LF)으로 변환하기	156
<그림 82> ls와 gawk를 이용하여 배치 파일 만들기: 형태분석 말뭉치	156
<그림 83> 한마루 2.0 파일 인코딩 변환	156
<그림 84> 한마루 2.0의 말뭉치 열기 대화상자	157
<그림 85> 한마루 2.0 파일 선택 대화상자	157
<그림 86> 한마루 2.0의 통계 메뉴	158
<그림 87> 한마루 2.0의 어절·음절·자소 통계	158
<그림 88> 한마루 2.0 기본 검색	159
<그림 89> 한마루 2.0 아무개 문자 검색 결과	160
<그림 90> 한마루 2.0의 자소 검색 결과	161
<그림 91> 한마루 2.0 범위 검색 결과	162
<그림 92> 한마루 2.0 불리언 검색 결과	163
<그림 93> 한마루 2.0으로 형태분석 말뭉치에서 지정사를 검색한 결과	164
<그림 94> 한마루 2.0으로 형태분석 말뭉치에서 품사 태그 및 아무개 문자를 이용하여 검색한 결과	165
<그림 95> ls와 gawk를 이용한 코드 변환: UTF-16LE → CP949	166
<그림 96> 지능형 형태소 분석기에서 input 파일 열기	167

<그림 97> 지능형 형태소 분석기로 학습자 원시 말뭉치의 한 파일을 태깅한 결과	167
<그림 98> 파일 이름 바꾸기 배치 파일	169
<그림 99> AntConc에서 말뭉치 파일 열기	170
<그림 100> AntConc에서 ‘있다’ 검색 결과 (어절 단위)	170
<그림 101> AntConc에서 ‘-는데’ 검색 결과 (어절 조건 해제)	171
<그림 102> AntConc에서 검색 결과를 해당 어절 기준으로 소팅한 결과	172
<그림 103> AntConc에서 아무개 문자를 이용한 검색: ‘먹+’	173
<그림 104> AntConc에서 ‘먹+’ 검색 결과를 좌측 어절 기준으로 정렬한 결과	173
<그림 105> AntConc에서 정규표현 ‘고.{0,20}고’를 검색한 결과	174
<그림 106> ‘사람’의 Concordance Plot	175
<그림 107> 학습자 원시 말뭉치의 어절 빈도 목록 (빈도순)	176
<그림 108> 학습자 원시 말뭉치의 어절 빈도 목록 (자모순)	177
<그림 109> ‘많이’의 연어 분석 결과	178
<그림 110> bigram 분석 결과	179
<그림 111> Cygwin 홈페이지(http://cygwin.com) 초기 화면	181
<그림 112> 고급 시스템 설정	181
<그림 113> 윈도우 10 ‘고급 시스템 설정’ 들어가기	182
<그림 114> ‘환경 변수’, ‘시스템 변수 편집’ 팝업창	182
<그림 115> 콘솔에서 gawk 실행 방법	183
<그림 116> gawk 실행 결과를 파일로 저장	184
<그림 117> sort 실행 방법	184
<그림 118> uniq 실행 방법	185
<그림 119> uniq 실행 결과	185
<그림 120> 숫자 내림차순 sort	186
<그림 121> 숫자 내림차순 sort 결과	186
<그림 122> gawk, sort, uniq, sort를 파이프로 연결하여 실행하는 방법	187
<그림 123> 오류 현상(10째 필드)에 대한 빈도 통계를 내는 방법	187
<그림 124> 오류 현상 통계	187
<그림 125> gawk로 조건에 맞는 레코드만 추출하기	188
<그림 126> GPT인 레코드만 추출한 결과	188
<그림 127> gawk로 특정 기준을 충족하는 레코드 추출. 기준이 2개인 경우	189
<그림 128> gawk로 특정 기준을 충족하는 레코드 추출. 기준이 3개인 경우	189
<그림 129> sort 명령어에서 기준 필드 지정	189

<그림 130> uniq -c와 sort -gr을 이용한 빈도 통계 추출	190
<그림 131> 조사 오류 유형 통계	190
<그림 132> ELAN의 음성 파일 처리 기능의 일부	192
<그림 133> 음성 인식 시스템의 구성	207
<그림 134> iOS와 Android의 음성 인식을 비교	226
<그림 135> 언어권별 음성 인식을 비교	226
<그림 136> 학습자 수준별 음성 인식을 비교	227
<그림 137> 모국어 화자와 한국어 학습자의 음성 인식을 비교	228
<그림 138> 응답자들의 응답 결과의 빈도	240
<그림 139> 기타 말뭉치 활용 희망 사항	242

I. 연구 개요

1. 연구의 목적 및 필요성



<그림 1> 연구 개요

1.1. 연구의 목적

○ 본 연구는 한국어 학습자 말뭉치를 대규모로 구축하고, 효과적으로 관리, 보급, 활용하기 위한 시스템의 중장기적 설계 및 발전, 활용 방안을 제시하는 데 그 목적이 있다.

○ 이를 위해 우선, 한국어 학습자 말뭉치 구축, 관리, 배포, 활용 시스템의 설계 및 개발 이행을 수립하고, 2015년 한국어 학습자 말뭉치 기초 연구 및 구축 사업에서 한국어 학습자 말뭉치 자료를 체계적/효율적으로 수집하고 구축하기 위한 지원 도구를 개발하며, 국내외 공개 및 상용 프로그램을 이용한 말뭉치 활용법을 개발한다.

1.2. 연구의 필요성

○ 학습자 말뭉치의 특수성을 반영한 구축 지원 도구

학습자 말뭉치는 모국어 기초 말뭉치와는 다른 특수성이 다수 존재한다. 불안정한 언어 능력으로 생성된 오류 포함 데이터, 생산자 정보의 중요도가 높은 데이터, 텍스트/음성이 혼합된 비정형 데이터 등 학습자 말뭉치의 특수성을 수용할 수 있는 구축 지원 도구가 필요하다.

○ 구축부터 활용까지 학습자 말뭉치의 전주기를 고려한 관리 시스템

학습자 말뭉치를 비롯한 말뭉치 데이터는 상용적으로 널리 사용되는 데이터들과는 많은 차이가 있다. 그러나 말뭉치 데이터가 활발하게 활용되기 위해서는, 말뭉치 데이터의 특수성과 함께 상용 데이터들의 특성도 반영되어야 한다.

소프트웨어(S/W)는 수명이 다해도 데이터는 사라지지 않고 진화한다. 그러므로 구축부터 활용까지 학습자 말뭉치의 전주기를 고려한 관리 시스템이 요구된다.

○ 다양한 활용 방안을 고려한 디지털화

기존에도 학습자 말뭉치를 구축하고 이에 기반한 연구를 진행한 사례가 있었으나 개별 기관 차원, 혹은 개인 차원에서 그치고 말았다. 여러 가지 사유가 있겠으나, 수집한 말뭉치를 다른 연구자들과 공유할 수 있는 형식으로 마크업하고 체계적으로 주석하여 활용이 가능하도록 제공하기 위한 인프라가 부족했기 때문이다. 한국어교육 분야의 전문 연구자뿐만 아니라 한국어교육 현장의 교수자도 활용할 수 있도록 결과의 활용 방안을 고려하여 학습자 말뭉치 데이터를 디지털화할 필요가 있다.

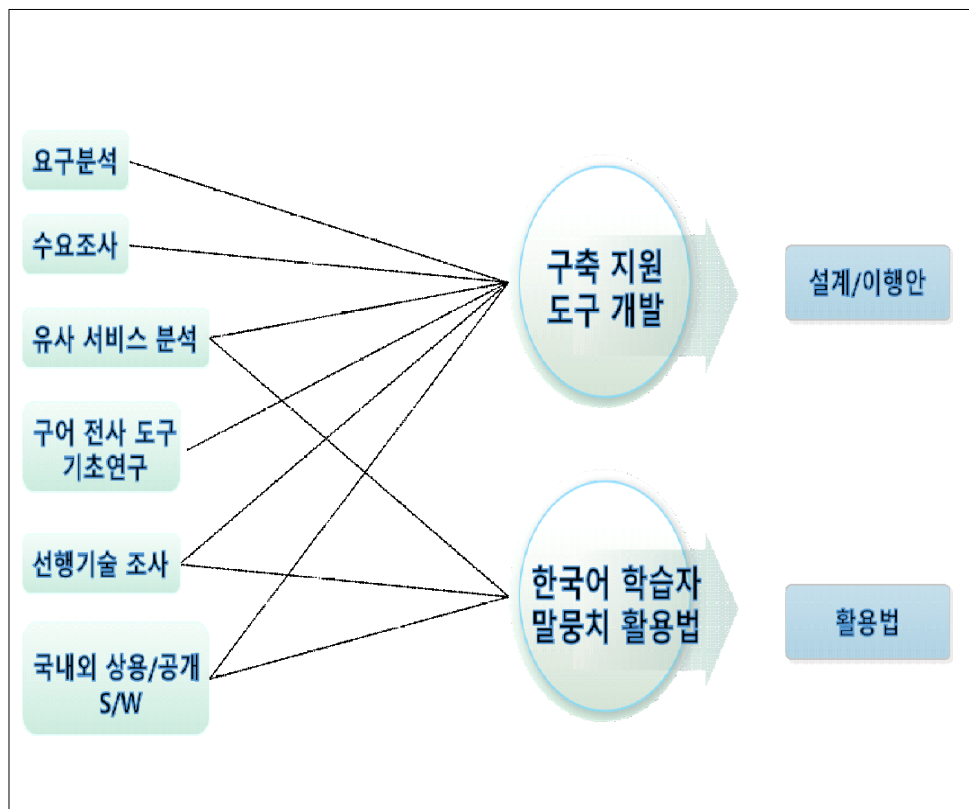
○ 개방적, 반복적, 점진적 개발 체계

장기간이 소요되는 학습자 말뭉치 구축의 특성을 고려하여, 개발 체계가 개방적이고, 구축 기간 동안 구축자의 요구를 지속적으로 반영하여 발전시킬 수 있는 시스템 개발 체계가 필요하다.

2. 연구의 방향

본 과업은 발주처와 한국어 학습자 말뭉치 구축팀의 요구 사항을 철저히 분석하고, 기존 도구 및 연구 검토를 바탕으로 한국어 학습자 말뭉치 구축 지원 도구를 선행 개발한다. 선행 개발 결과의 피드백을 통해 향후 한국어 학습자 말뭉치 구축, 관리, 배포, 활용에 필요한 전체 시스템의 설계 및 계획을 수립한다.

이와 함께 요구 분석 및 수요 조사, 국내외 유사 서비스 분석, 학습자 말뭉치 구축 지원 도구에 적용된 데이터 및 데이터 처리 기술을 바탕으로 한국어 학습자 말뭉치의 활용법을 제시한다.



<그림 2> 연구의 방향

3. 연구 범위

본 연구의 범위는 크게 ①한국어 학습자 말뭉치 구축, 관리, 배포, 활용 시스템 설계 및 계획 수립, ②구축 지원 도구 개발, ③공개/상용 프로그램을 활용한 한국어 학습자 말뭉치 활용법 개발로 구성된다. 각각의 수행 과업은 다시 다음과 같은 세부 과업을 포함한다.

핵심 과업	세부 과업
말뭉치 구축, 관리, 배포, 활용 시스템 설계 및 계획 수립	학습자 말뭉치 구축 및 이용자 수요 분석
	국내외 유사 서비스 분석
	사용자 접근성이 우수한 시스템 설계
	말뭉치 구축, 관리, 배포, 활용 시스템 설계
	한국어 학습자 말뭉치 구축, 관리, 배포, 활용 시스템 개발 이행안 수립
구축 지원 도구 개발	한국어 학습자 말뭉치 구축/지원 웹페이지 개발
	한국어 학습자 말뭉치 구축 및 관리 지원 도구
	한국어 학습자 오류 주석 말뭉치 구축 지원 도구
	한국어 학습자 원시 말뭉치 및 가공 말뭉치 형식 오류 검사 도구
한국어 학습자 말뭉치 활용법 개발	국내외 공개 및 상용 프로그램을 이용한 말뭉치 활용법 작성

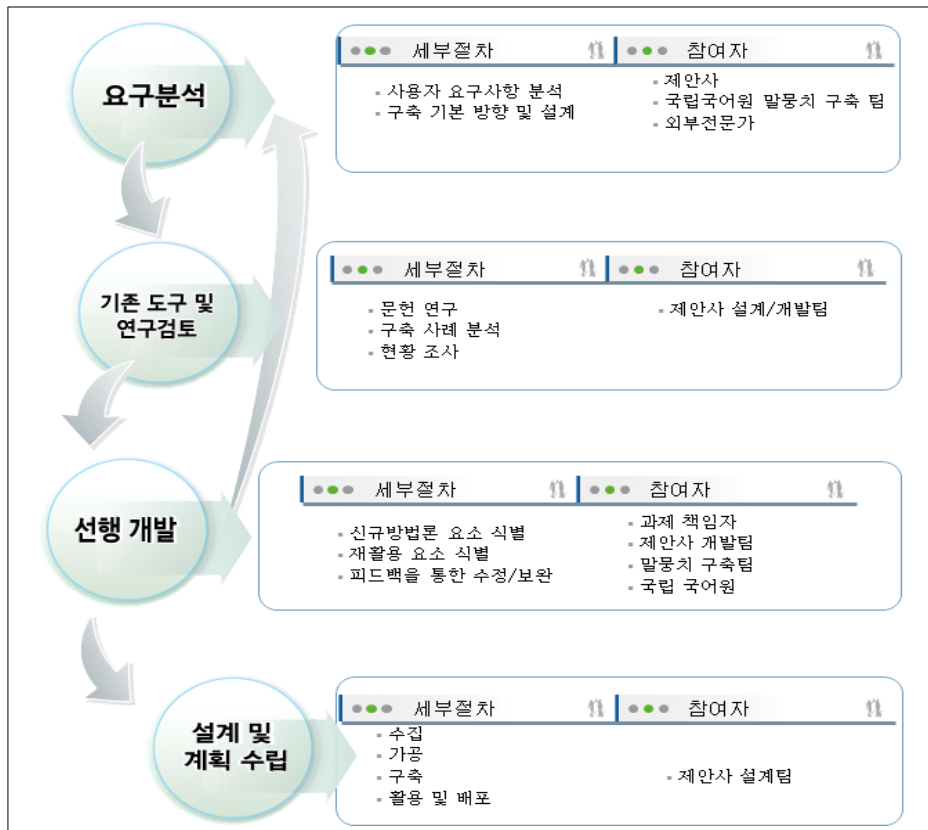
<표 1> 연구의 범위 및 세부 과업 내용

4. 연구 방법

○ 본 과업은 발주처 및 한국어 학습자 말뭉치 구축팀과 긴밀하게 협조하여, 문어/구어를 포함한 한국어 학습자 말뭉치 데이터의 통합 관리 체계(비정형 데이터 통합 관리 체계)와 이에 사용되는 주석 체계를 일관성이 있게 저장, 관리할 수 있는 통합 주석 체계를 개발한다.

○ 개발된 통합 주석과 학습자 말뭉치 데이터 관리 체계를 바탕으로 학습자 말뭉치 구축 지원 도구를 개발하여, 구축 관리자, 작업자, 사용자, 학습자에게 배포/검증한다.

○ 한국어 학습자 말뭉치의 구축 관리자, 작업자, 사용자, 학습자의 사용 경험을 바탕으로 말뭉치 구축, 관리, 배포, 활용에 필요한 전체 시스템의 설계 및 개발 계획을 수립한다.



<그림 3> 연구의 방법 및 절차

5. 연구 추진 일정

5.1. 연구 수행 단계

○ 계획(요구 수집 및 분석) 단계

시스템의 수요, 유사 서비스, 학습자 말뭉치 구축팀의 요구를 수집/정리한다.

○ 아키텍처 단계

수집/분석된 요구 사항을 수용할 수 있는 아키텍처와 컴포넌트를 도출한다.

○ 점진적 개발 단계

도출된 컴포넌트를 개발/테스트(사용자 피드백)하여 개선-보완한다.

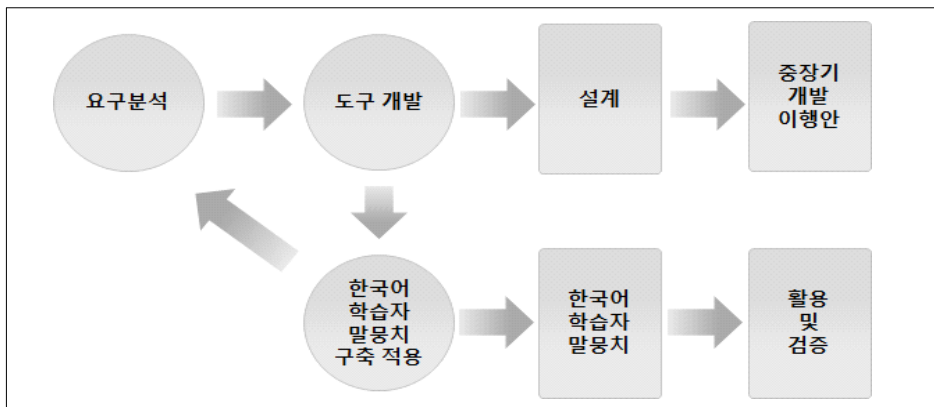
○ 반복적 점진적 개발 프로세스

컴포넌트 구현 - 테스트 - 개선/보완을 반복하여 전체 시스템의 설계 목표를 점진적으로 완성한다.

○ 인도 단계

목표 시스템 설계 결과물 및 이행을 작성하고, 개발된 컴포넌트들을 사용자에게 최종 전달한다.

5.2. 연구 수행 절차



<그림 4> 한국어 학습자 말뭉치 구축 지원 도구 개발 연구 수행 절차

○ 한국어 학습자 말뭉치 구축 지원 웹페이지

당해 연도 우선 요구 기능을 중심으로 개발하여 점진적 개선하는 전략을 취한다. 특히 학습자 말뭉치 관리 시스템의 전체 설계를 위한 요구 사항을 도출하고 적용 방법을 모색할 수 있도록 확장성과 재사용성에 중점을 두고 개발한다.

○ 한국어 학습자 말뭉치 구축 지원 도구 개발

학습자 말뭉치 관리 시스템의 전체 설계를 위한 요구 사항을 도출하고, 확장성과 재사용성에 중점을 두고 개발한다. 구축 효율을 위한 자동 주석 및 자동 주석 도구와의 연계가 용이하도록 하며, 문어/구어 데이터의 통합 저장/관리를 지원한다.

○ 한국어 학습자 오류 주석 말뭉치 구축 및 관리 지원 도구

통합적인 주석 관리 체계를 수립한 후 이를 구현한다. 단어, 구, 문장, 담화 등 다양한 층위의 주석을 일관된 인터페이스로 제공하며, 주석 작업자의 사용 편의성을 극대화하여 구축 효율을 향상시킨다.

○ 한국어 학습자 말뭉치 원시 및 가공 말뭉치 형식 오류 검사 도구

말뭉치 표본의 생성 단계에서부터 구축 및 가공 전 단계에 걸친 품질 관리를 실시한다.

6. 연구 결과

6.1. 주요 결과물 개요

구분	모듈	개발 결과물
구축	수집	학습자 말뭉치 표본 등록 모듈/화면
		회원/학습자 정보 등록 모듈/화면
		회원/학습자 정보 DB 및 그 스키마
		파일(표본) 관리 모듈/화면
		파일 정보 DB 및 그 스키마
	가공 (입력/전사/분석/오류 주석)	주석 관리 모듈/화면
		주석 작업 UI
		주석 데이터 저장/관리 모듈
		말뭉치 주석 자동 생성 모듈(또는 연동)
		통합 주석 제어 모듈
관리	열람	표본별 열람 모듈/화면
	현황	수집된 표본 통계 보기 모듈/화면
		수집된 표본의 가공 공정률 보기 모듈/화면
	설계	부분 말뭉치 설계 모듈/화면
		설계 정보에 의한 표본 제어 모듈
		구축/가공 지침 관리 모듈
	교정	오류 항목 신고 기능
		오류 항목 관리 기능
		오류 항목 교정 기능
		교정 이력 관리 기능
		교정 수행 모듈
	검사	문자코드 검사 모듈
		주석된 표본 형식 검사 모듈
		원시-가공 표본 무결성 검사 모듈
		주석 일관성 검사 모듈
		주석 무결성 검사 모듈

<표 2> 주요 결과물 개요

6.2. 주요 기능 구현 예시

기능명	학습자 말뭉치 표본 등록
주요 기능 사항	■ 한국어 학습자 말뭉치 파일 관리 기능 (정보 입력, 파일명 자동 생성 등) ■ 원본 스캔 이미지 및 음성 파일과 입력/전사 텍스트 파일 전송 ■ 파일명 자동 생성 (학습자 및 작업자, 입력 내용을 기반으로)
[화면 예시]	
추가 기능	■ 학습자 정보 입력 및 자동 부가

<표 3> 학습자 말뭉치 표본 등록 화면 예시

기능명	파일(표본) 관리
주요 기능 사항	■ 조건 검색 등을 통한 조회 및 분류 기능 ■ 말뭉치 파일 일괄 등록, 수정, 내려 받기 기능

[화면 예시]

표본조회 화면

검색조건

ID

학습환경 ☐ 해외 ☐ 국내

학습기관

학습목적

모국어

한국어등급

파일유형

주석별조회

원거주식 ☒

원머주식 ☒

구분주식 ☒

수집 기간별 조회

Date Reported: 1/1/2009 ~ 3/31/2009

3월 2009

1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31	1	2	3	4

검색조회

표본조회결과

선택	ID	파일명	학습환경	학습기관	학습목적	모국어	한국어등급	파일유형
☒	sdjfhadk	001	국내	A어학당	이주	eng	3등급	문어
☐	qowwep	002	국내	B어학당	취업	eng	1등급	구어

표본을 선택하시겠습니까?

표본삭제 화면

전체표본조회

ID

학습환경 ☐ 해외 ☐ 국내

학습기관

학습목적

모국어

한국어등급

파일유형

주석별조회

원거주식 ☒

원머주식 ☒

구분주식 ☒

수집 기간별 조회

Date Reported: 1/1/2009 ~ 3/31/2009

3월 2009

1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31	1	2	3	4

검색조회

표본조회결과

선택	ID	파일명	학습환경	학습기관	학습목적	모국어	한국어등급	파일유형
☒	sdjfhadk	001	국내	A어학당	이주	eng	3등급	문어
☐	qowwep	002	국내	B어학당	취업	eng	1등급	구어

경고창

선택한 표본을 삭제하시겠습니까?

- 추가
기능
- 파일명 자동 생성의 시스템 내 위치와 권한 설정
 - 한국어 학습자 말뭉치 자원 이력 관리

<표 5> 파일(표본) 관리 화면 예시

기능명	주석 관리
주요 기능 사항	■ 다층 주석(단어, 구, 문장, 담화 층위를 분리하여 주석) ■ 한국어 학습자 말뭉치 원시 및 가공 말뭉치 형식 오류 검사
[화면 예시]	
<pre> graph TD Manager[관리자] --> Add[주석추가/삭제/변경 이력 검토] Worker[작업자] --> Request[주석변경/추가요청] User[사용자] --> Suggest[의견제시] Add --> Main[주석체계] Request --> Main Suggest --> Main Main --> Check[주석체계 검토] Check --> DB[(주석DB)] DB --> Check Check --> Format[말뭉치 형식검사] Format --> DB DB --> Format Format --> UI[말뭉치 형식검사 화면] UI --> DB DB --> UI Main --> UI UI --> Main </pre>	
추가 기능	■ 일관성 있는 통합 주석 관리 체계 ■ 주석 관리 체계에 대한 접근 권한과 관리 정책 ■ 주석 추가, 제거, 지침/주석 체계 관리용 관리자 화면

기능명	주석 작업 UI
주요 기능 사항	■ 한국어 학습자 말뭉치 온라인 구축 기능 ■ 말뭉치 주석 자동 생성 기능 ■ 다층 주석 - 단어, 구, 문장, 담화 층위의 오류를 분리/중첩 주석 ■ 품사 주석 자동/반자동 생성 및 편집 기능

[화면 예시]

통합주석작업(예설정)

나는 친구를 부산에 갔어요.
 한계에 있는 문 중에 부분만을 택한 부분이 없어요.
 날씨가 아주 좋은 날씨가 아직 너무 주워요.
 그래서 우리는 구경하고 산책했어요.
 날씨가 굳게 공기가 얼마나 좋을까 모르겠어요.
 날씨가 맑기 위해서 에어컨을 많이 하고 물도 많이 마셨어요.
 부산을 사용했다 바쁘지 않고 진한 사람도 있어요.
 부산은 아주 멋있어서 생긴 시장에서 많이 먹었어요.
 부산 사람들은 사도리가 있으면 아깝지 않아요.
 그래서 이 때문에 사용한 날지 말고 부산에 한번 가세요.

Offset	1	2	3	4	5	6	7	8	9
원어절	나	는	친구	를	부산	에	가	갔	어요
형태주석	NP	JX	NNP	JX	NNP	JKB	VV	EP	EF
교정어절									
분석가능여부									
오류출력									
오류현상									

단어(VOW)

통합주석작업(예설정)

나는 친구를 부산에 갔어요.
 한계에 있는 문 중에 부분만을 택한 부분이 없어요.
 날씨가 아주 좋은 날씨가 아직 너무 주워요.
 그래서 우리는 구경하고 산책했어요.
 날씨가 굳게 공기가 얼마나 좋을까 모르겠어요.
 날씨가 맑기 위해서 에어컨을 많이 하고 물도 많이 마셨어요.
 부산을 사용했다 바쁘지 않고 진한 사람도 있어요.
 부산은 아주 멋있어서 생긴 시장에서 많이 먹었어요.
 부산 사람들은 사도리가 있으면 아깝지 않아요.
 그래서 이 때문에 사용한 날지 말고 부산에 한번 가세요.

Offset	1	2	3	4	5	6	7	8	9
원어절	나	는	친구	를	부산	에	가	갔	어요
형태주석	NP	JX	NNP	JX	NNP	JKB	VV	EP	EF
교정어절									
분석가능여부									
오류출력									
오류현상									

단어(VOW):음절[PS]

추가 기능	■ 헤더 정보, 마크업 자동 생성 및 편집 기능 ■ 오류 반자동 주석(반복 오류, 주석 표지 자동 완성 등) ■ 중첩 주석 표기 UI 연구
--------------	---

<표 7> 주석 작업 UI 화면 예시

기능명	형식 오류 검사 및 작업 UI
주요 기능 사항	■ 한국어 학습자 가공 말뭉치 형식 오류 검사 ■ 오류 주석 제거 기능

[화면 예시]

1

2

3

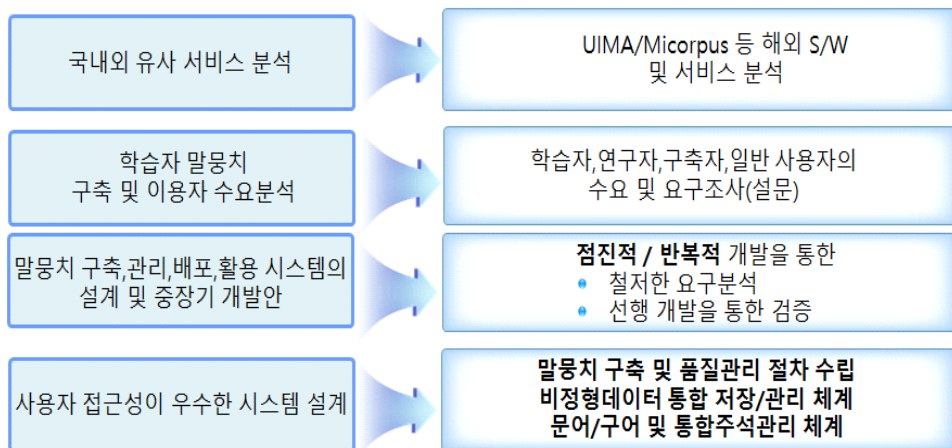
추가 기능	■ 학습자 오류 추정 기능 ■ 학습자 오류 자동 주석 기능 ■ 오류 주석 일괄 제거 기능
----------	---

<표 8> 형식 오류 검사 및 작업 UI 예시

II. 한국어 학습자 말뭉치 통합 관리 시스템 설계 및 이행안

한국어 학습자 말뭉치 통합 관리 시스템은 한국어 학습자 말뭉치를 체계적이고 효율적으로 수집하고, 대규모 한국어 학습자 말뭉치를 효과적으로 관리, 보급, 활용하는 데 그 목적이 있다. 본 연구의 목표는 이러한 한국어 학습자 말뭉치의 통합 관리 시스템을 설계하고, 중장기적인 이행을 수립하는 것이다.

이를 위해 먼저, 한국어 학습자 말뭉치의 구축자와 이용자의 수요를 분석하여 시스템의 규모의 적정성과 요구 기능들을 도출한다. 한국어 학습자 말뭉치 통합 관리 시스템 아키텍처는 다양한 사용자에게 접근성이 우수한 인터페이스와 말뭉치 구축 작업에 효율적인 시스템 구성 요소를 배치하였다.



최신 기술과 Know-how를 접목한 차세대 말뭉치 통합 서비스 시스템 설계

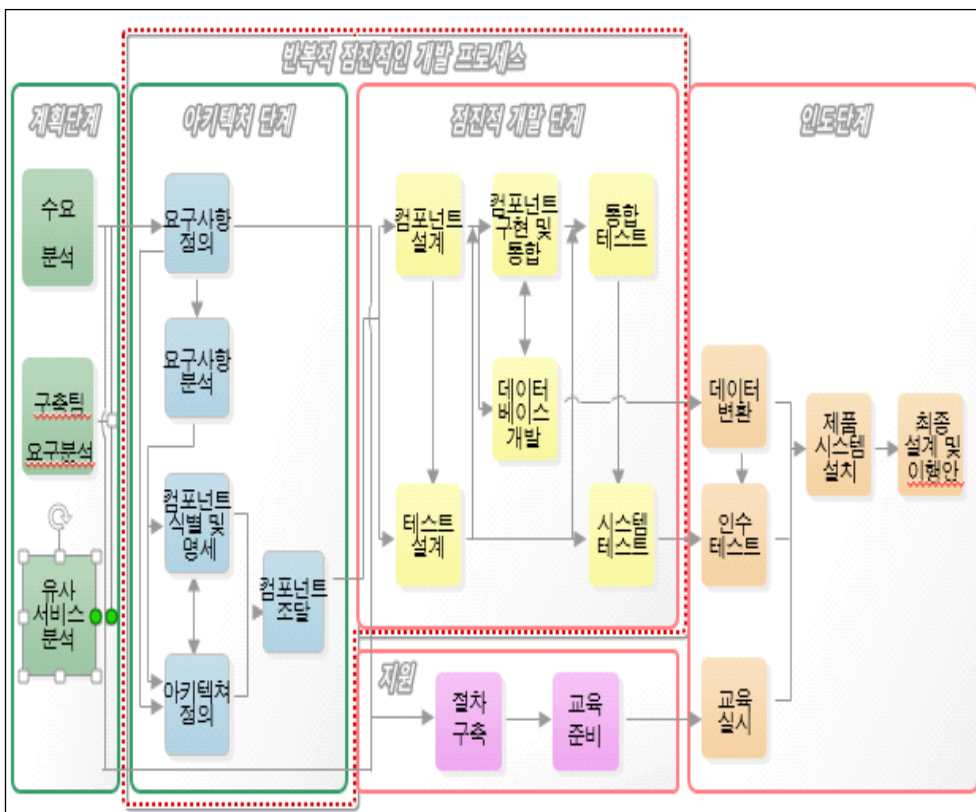
<그림 5> 한국어 학습자 말뭉치 통합 관리 시스템 설계

한국어 학습자 말뭉치 구축 및 이용자의 요구, 수요 분석은 구축 연구/실무팀의 면담, 선행 개발된 도구 시연 및 테스트, 선행 연구 결과의 분석을 통해 수행하였다. 또한, 수집된 요구 사항을 정리/분석하여 설문 조사 계획을 수립함으로써 보다 다양한 사용자의 요구와 의견을 수집할 수 있도록 하였다.

수집/정리된 요구 분석은 국내외 유사 서비스 및 선행 연구 검토를 통해 한국

어 학습자 말뭉치 통합 관리 시스템의 요소 기술을 결정하고 요구 분석의 타당성을 검증하는 데 적용되었다. 또한 작업 공정, 데이터별 품질 관리(주석(메타데이터 포함), 말뭉치 유형, 매체(텍스트, 음성, 이미지, 비디오)) 등과 관련한 시스템의 품질 요소를 선정하는 데에도 적용되었다.

이와 함께 사용자 접근성이 우수한 시스템 개발을 위해, 4차례의 구축 지원 도구 시제품 개발과 시험을 통해 사용자의 요구 사항을 구체화하는 한편, 'UI/UX'적 요소와 사용자의 다양한 요구 환경(목적, 언어학적 견해, 활용 도구 등에 따른 차이)을 충족할 수 있는 고가용성을 확보할 수 있도록 시스템 구성 요소들을 조정하였다.



<그림 6> 시스템 설계 및 개발 계획 수행 방법 및 절차

금번 연구는 한국어 학습자 말뭉치 통합 관리 시스템의 전체 구성, 요구 분석을 통한 유스케이스와 아키텍처, 기능 목록 도출을 중점적으로 수행하였으며, 구

축 지원 도구의 사용자 테스트와 추가 요구 사항 수용 등의 과정을 거쳐 보완하였다. 개별 기능에 대한 상세한 설명은 III. 한국어 학습자 말뭉치 구축 지원 도구 개발에서 제시하기로 한다.

1. 요구 분석 및 수요 조사

1.1. 요구 분석 개요

한국어 학습자 말뭉치 구축 및 이용자의 요구 및 수요 분석은 구축 연구/실무팀의 면담, 선행 개발된 도구 시연 및 테스트, 선행 연구 결과의 분석을 통해 수행하였다. 또한, 수집된 요구 사항을 정리 분석하고, 설문 조사를 실시하여 다양한 사용자의 요구와 의견을 확인하였다. 설문 조사와 관련한 사항은 VI장에서 기술하였다.

1.2. 요구 사항 수집

구분	요구 사항	설명 내용	출처/목차
수집	말뭉치 구축 규모 산출	전체 말뭉치 현황을 알 수 있어야 한다.	내용 구축팀 제안서 /IV.2.4.1/3.1.1
수집	표본의 크기 산출	학습자의 실제 작문이나 발화의 산출 량을 기준으로 산정해야 한다.	내용 구축팀 제안서 /IV.2.4.1
수집	대상 특성별 표본 식별	교육 환경, 학습 목적, 학습 대상별 언 어 특성을 살필 수 있어야 한다.	내용 구축팀 제안서 /IV.2.4.1
수집	숙달도 단계별 표본 식별	단계별 언어 발달과 습득 과정에 대한 종적 연구가 가능해야 한다.	내용 구축팀 제안서 /IV.2.4.1
수집	학습자 유형별 표본 식별	국내 학습자(1단계), 이주민(2단계), 국 외 학습자(3단계)로 나누어 한국어 학 습자 말뭉치를 구축해야 한다.	내용 구축팀 제안서 /IV.3.1.1

수집	수집기관별 데이터 식별	일정 규모를 갖춘 다양한 기관으로부 터 수집해야한다.	내용 구축팀 제안서 /IV.3.1.1
수집	원격지 표본 수집	온라인 자료 수집 시스템을 활용한 수집	내용 구축팀 제안서 /IV.3.1.1
수집	학습자 개인 정보의 활용과 자료 사용에 관련 동의서 수집	수집된 학습 자료는 학습자의 동의 증 빙과 연동되어야 한다.	내용 구축팀 제안서 /IV.3.1.1
가공	말뭉치 주석 틀(태그 세트) 체계 마련	학습자 언어의 특성을 반영하되, 한국 어교육 연구, 교육, 학습 분야에서 효율 적으로 활용도가 높은 내용과 체계로 개발되어야 한다.	내용 구축팀 제안서 /IV.3.1.1
가공	1차 가공 주석 틀 체계 개발	중간언어 자료로서 학습자 언어를 살 필 수 있는 품사 주석 외에 유창성, 복 잡성, 다양성의 측면에서 학습자 언어 를 관찰할 수 있는 T-단위 주석이 필 요하다.	내용 구축팀 제안서 /IV.3.1.1
가공	2차 가공 주석 틀 체계 개발	오류에 초점을 둔 2차 가공 주석인 오 류 주석이 필요하다.	내용 구축팀 제안서 /IV.3.1.1
가공	개방형 오류 주석 체계	다목적으로 광범위하게 사용 가능한 공통의 오류 주석을 제공하고 사용자 의 목적에 따라 제공되는 오류 주석을 수정·보완하여 사용하거나 추가 주석 작업을 할 수 있어야 한다.	내용 구축팀 제안서 /IV.3.1.1
가공	주석 체계 호환성	국제적 수준에서 전 세계의 학습자 말 뭉치와 21세기 세종 한국어 균형 말뭉 치와의 호환성을 고려해야 한다.	내용 구축팀 제안서 /IV.2.4.3.
가공	입력 지침과 전사 지침 개발	말뭉치의 입력 지침과 전사 지침의 기 본 원칙을 따르되, 학습자 말뭉치의 특 징인 중간언어 체계를 수용하기 위한 원칙도 정해야한다.	내용 구축팀 제안서 /IV.3.1.1

가공	형태 주석 교정	반자동 방식으로 1차 형태소 분석을 진행한 후 2차 수작업을 통해 오분석 어절을 수정할 수 있어야 한다.	인터뷰
관리	파일명 부여 체계 개발	파일명은 자료의 유형, 교육환경, 언어권, 숙달도 단계 등의 내용을 담고 있는 정보를 연상할 수 있도록 체계화하여 부여해야 한다.	내용 구축팀 제안서 /IV.2.4.3. 내용 구축팀 제안서 /IV.3.1.1.(2)
관리	헤더 주석 체계 개발	21세기 세종 한국어 균형 말뭉치의 헤더 체계를 기초로 파일에 대한 정보, 입력자에 대한 정보, 발화자 정보 등을 포함해야 한다.	내용 구축팀 제안서 /IV.2.4.3. 내용 구축팀 제안서 /IV.3.1.1.(2)
관리	다양한 메타데이터와 연동된 표본 관리	다량의 자료를 체계적이고 효율적으로 관리할 수 있도록 파일의 세부 정보를 활용한 기능을 갖춘 프로그램을 개발해야 한다.	내용 구축팀 제안서 /IV.3.1.1.
관리	호환성, 가용성이 높은 말뭉치 관리	웹 관련 기술과 Open Annotation 기술을 결합한 동적인 말뭉치 관리 시스템을 개발해야 한다.	서상규(2010)/IV.1.5
분석	개발 결과물의 재활용성	집단별 활용 목적에 부합하는 특화된 도구를 개별적으로 개발해야 한다. 1)연구자 집단 / 2)교수자 집단 / 3)학습자집단	서상규(2010)/IV.1.3.4
분석	다양한 검색 / 색인지원 / Web 기반 구현	용례 검색 등의 정적 기능 별도의 소프트웨어 설치 및 갱신이 필요 없는 웹 기반 도구로 개발해야 한다.	서상규(2010)/IV.1.4.2
분석	구축된 학습자 말뭉치 응용 도구 지원 (학습 등)	작문과정에서 활용할 수 있는 오류 검출 및 교정 도구를 온라인으로 접속 가능하도록 구현하여 구축한 학습자 말뭉치 자료를 학습자별 맞춤 자료로 제공함으로써 더 효과적인 오류의 검출과 교정 기능을 제공해야 한다.	서상규(2010)/IV.1.4.3

배포	통계 분석	다양한 추가 정보의 주석과 이에 대한 검색 및 통계 정보 추출이 가능해야 한다.	서상규(2010)/ IV.1.4.1
배포	표본의 다양한 정보 종합 제공	학술적 활용이 가능하고 새로운 원자료의 추가 및 연구 노트에 가까운 자유 형식의 주석을 제공해야 한다.	
배포	협업 가능한 주석 편집 도구	네트워크를 통해 이미지 및 텍스트 형태로 공유되는 원자료에 대한 다양한 형태의 주석을 협업 작업에 의해 수행할 수 있는 환경을 제공해야 한다.	
가공	사용자 입력 오류를 배제한 주석 편집 도구	모든 주석 작업에서 사용자의 입력에 의한 오류를 배제해야 한다.	인터뷰
가공	주석 편집 이력 제공	각 주석에 변경 사항이 있는 경우, 그 변형이력을 조회할 수 있어야 한다.	인터뷰
가공	작업 현황 제공	작업자별, 표본별 작업자의 작업 현황을 조회 할 수 있어야 한다.	인터뷰
관리	작업 파일 수정 및 삭제 권한 제한	작업 파일의 수정 및 삭제 권한을 작업자 또는 관리자로 제한하여 파일이 임의로 수정되거나 삭제되지 않도록 한다.	인터뷰
관리	주석 편집기의 주석 사전	오류 주석과 마찬가지로 형태 주석 태그 세트를 참조하여 수정할 수 있어야 한다.	인터뷰
관리	작업자 로그 정보 표시	작업 파일 관리의 효율을 위해 작업 시간, 파일 최종 저장 시간 등을 볼 수 있도록 한다.	인터뷰

<표 9> 요구 분석 명세서

2. 유사 서비스 분석

2.1. 국내외 학습자 말뭉치 현황

루뱅 가톨릭 대학교의 영어 말뭉치 언어학 센터에서 제공하는 목록에는 전 세계적으로 140여 개의 학습자 말뭉치가 소개되어 있다.

(<http://www.uclouvain.be/en-cecl-lcworld.html>) <2015년 12월 현재>

The screenshot shows the website of the Centre for English Corpus Linguistics (CECL) at the University of Louvain. The page is titled "Learner corpora around the world" and contains a table listing various learner corpora. The table has columns for Corpus, Target language, First language, Medium, Text type/task type, Proficiency level, Size in words, Project director, and Availability.

Corpus	Target language	First language	Medium	Text type/task type	Proficiency level	Size in words	Project director	Availability
The Arabic Learner Corpus (AUC)	Arabic	66 languages	written and spoken	narrative and discussion	intermediate and advanced	written: c. 263,000 audio: c. 3e30	Abdullah Alfaraj & Eric Atwell	available
The First Arabic Learner Corpus	Arabic	English	written	narrative	intermediate and advanced	c. 9,000	Ghada Abuhakima, Reem Faraj, Anna Feilman, Elwan Fikselink, Marianne Sime University, USA	
The Urban Chinese Learner Corpus (UCLC)	Chinese	50 languages	written	exams and assignments	beginners, intermediate and advanced	c. 6 million Chinese characters	Naoh Wang, Shoulin Malenisi, Mingquan Huang	freely available
The AKCEC/CECL corpus (Acquisition corpus of Czech/English)	Czech	various	written and spoken	student essays and interviews	various	2 m	Karel Sebesta, Charles University in Prague, Technical University in Liberec, Czech Republic	available

<그림 7> 영어 말뭉치 언어학 센터에서 제공하는 전 세계의 학습자 말뭉치 목록

국내에서 구축된 것으로는 다음과 같이 주로 영어 학습자 말뭉치가 주를 이루고 있다. 국내의 한국어 학습자 말뭉치는 자세히 소개되어 있지 않다.

- The EFL Teacher Corpus (ETC): 한국인 영어 교사의 수업 중 발화 123,000 단어
- The Gachon Learner Corpus: 한국인 영어 학습자의 작문 250만 단어
- The Neungyule Interlanguage Corpus of Korean Learners of English (NICKLE): 한국인 영어 학습자의 에세이, 인터뷰, 구두 말하기 시험 전사 약 100만 단어

- The Seoul National University Korean-speaking English Learner Corpus (SKELC): 한국인 영어 학습자의 에세이 90만 단어
- The Yonsei English Learner Corpus (YELC): 한국인 영어 학습자의 작문 약 100만 단어
- The Korean learner corpus (연세대): 한국어 학습자 작문 1만 어절

○ 학습 대상 언어별 통계

대상 언어는 모두 22가지로 영어, 프랑스어, 독일어, 스페인어, 이탈리아어 등 주요 유럽어가 대부분을 차지하고 있다. 대상 언어가 여럿인 다언어 학습자 말뭉치도 있다. 유럽권 이외에는 아랍어, 한국어, 중국어 정도만 존재한다. 대상 언어 목록과 말뭉치의 수를 요약하면 다음과 같다.

대상 언어	말뭉치 수	대상 언어	말뭉치 수	대상 언어	말뭉치 수
English	92	Swedish	3	Slovene	1
French	15	Finnish	3	Russian	1
German	14	Finnish	2	Romanian	1
Spanish	13	Dutch	2	Portuguese	1
Multilingual	8	Czech	2	Norwegian	1
Italian	8	Arabic	2	Korean	1
Hungarian	1	Gaelic	1	Estonian	1
Chinese	1	Catalan	1		

○ 학습자 모어별 통계

학습자 모어는 다양하게 섞여 있는 경우가 35여 건으로 가장 많다. 개별 언어 별로 살펴보면 중국어가 모어인 경우가 가장 많으며, 영어와 독일어가 모어인 학습자의 말뭉치가 같은 수준으로 많은 수를 차지한다. 스페인어, 일본어, 스웨덴어, 한국어 등이 뒤를 잇고 있다. 중국어, 일본어, 한국어의 경우에는 학습 대상 언어로서 말뭉치는 매우 적지만 학습자의 모어로서는 많은 비중을 차지하고 있음을 알 수 있다.

대상 언어	말뭉치 수	대상 언어	말뭉치 수	대상 언어	말뭉치 수
Chinese	20	Malay	6	Lithuanian	2
English	18	Greek	5	Icelandic	2
German	17	Polish	4	Hungarian	2
Spanish	9	Finnish	4	Cypriot	2
Japanese	9	Dutch	4	Brazilian	2
Swedish	8	Catalan	4	Turkish	1
Korean	7	Portuguese	3	Tswana	1
Italian	7	Arabic	3	Slovak	1
French	7	Ukrainian	2	Romanian	1
Russian	6	Norwegian	2	Québec	1
Qatar	1	Komi	1	Estonian	1
Punjabi	1	Jamaican	1	Czech	1
Mandarin	1	Indonesian	1	Creole	1
Latvian	1	Indian	1	Cantonese	1
Lappish	1	Hebrew	1	Belorussian	1
Asian	1	American	1		

2.2. 학습자 말뭉치 웹 서비스 사례

(1) MICASE & MICUSP (<http://quod.lib.umich.edu/m/micase/>)

○ MICASE 말뭉치

MICASE(The Michigan Corpus of Academic Spoken English)와 MICUSP(The Michigan Corpus of Upper-level Student Papers)는 미국 미시건 대학에서 구축한 영어 말뭉치이다.

MICASE는 미시건 대학에서 이루어진 강연, 수업, 토론, 세미나, 발표, 인터뷰 등 다양한 학술적인 상황의 발화를 녹음하고 전사한 영어 구어 말뭉치이다. 1997-2002년 사이에 구축되었으며, 약 200시간 분량의 음성 자료를 전사한 152개 스크립트로 이루어졌으며, 모두 합쳐 약 185만 단어 규모이다. 인문, 사회, 자연, 공학 등 다양한 학과에서 자료를 수집하였고, 발화 참여자도 학부생, 대학원생,

교수, 직원 등을 포괄하고 있다. 발화의 유형도 독백부터 대화까지 다양하게 수집되었다.

MICASE는 기본 성격은 학술적인 상황에서 사용되는 구어 말뭉치이지만, 많은 유학생들이 대화에 참여하고 있으므로 학습자 말뭉치의 성격을 가지게 된다. 말뭉치에 포함된 발화 참여자의 모어는 모두 38가지(아랍어, 아르메니아어, 광둥어, 중국어, 체코어, 네덜란드어, 에스토니아어, 페르시아어, 불어, 독일어, 구자라트어, 헤브루어, 힌디어, 헝가리어, 이보어, 인니어, 이탈리아어, 일본어, 칸나다어, 한국어, 마케도니아어, 마라티어, 폴란드어, 포르투갈어, 러시아어, 남아공어, 슬로바키아어, 스페인어, 스와힐리어, 스웨덴어, 타갈로그어, 텔루구어, 태국어, 터키어, 영국 영어, 우크라이나어, 우르두어, 베트남어)이다. 전사된 모든 발화에는 발화자 정보가 표시되어 있으며 발화자의 역할, 나이, 성별, 모어 정보가 제공된다. MICASE 말뭉치의 XML의 본문은 다음과 같다.

```
<U1 WHO="S1" NSS="NS" ROLE="SF" SEX="M" AGE="4" RESTRICT="NONE">
  okey-doke, uh well Kim Sook was gonna do another, very brief summary of what
  he's up to. uh to bring it all up on our screens...
</U1>
<U1 WHO="S2" NSS="NNS" ROLE="SG" SEX="M" AGE="2" RESTRICT="NONE" FL
ANG="KOR">
  alright, um... first of all i'd like to thank all of you, for agreeing to be on the
  committee, reading the draft, and coming to the defense, being, with me at, my last
  moment of, graduate school
</U1>
<U1 WHO="S3" NSS="NS" ROLE="SF" SEX="M" AGE="4" RESTRICT="NONE">
  <EVENT DESC="LAUGH"/>such optimism<EVENT WHO="SS" DESC="LAUGH"/>
</U1>
<U1 WHO="S2" NSS="NNS" ROLE="SG" SEX="M" AGE="2" RESTRICT="NONE" FL
ANG="KOR">
  <OVERLAP1>um</OVERLAP1></U1>
```

○ MICASE 탐색 서비스

브라우저 서비스에서는 화자의 "Academic Position", "Native Speaker" 여부, L1 언어를 지정하고 "Speech Event"의 형태(컬로퀴엄, 인터뷰, 강의 등)와 단과대학, 학과, 참여 수준, 상호작용 수준 등을 지정하여 해당 말뭉치를 찾아볼 수 있다. 원문을 HTML 형태로 보기 좋게 볼 수 있으며, XML로 다운로드 받을 수도 있다.

온라인 탐색 도구에서는 다음과 같은 메뉴를 제공하고 있다. 크게는 화자 속성(Speaker Attributes)과 전사 속성(Transcript Attributes)으로 나뉘어 있다. 가질 수 있는 속성값은 모두 나열하지 않았다.

●Speaker Attributes

▷Academic Position/Role:

All, Junior Faculty, Junior Graduate Student, ...

▷Native speaker statues:

All, Non-native speaker, Near-native speaker, ...

▷First language:

All, Arabic, Armenian, Cantonese, Mandarin, Czech, ...

●Transcript Attributes

▷Speech Event Type:

All, Advising Session, Colloquium, Dissertation Defense, ...

▷Academic Division:

All, Biological and Health Science, Humanities and Arts, ...

▷Academic Displine:

All, Afroamerican and African Studies, American Culture, ...

▷Participant Level:

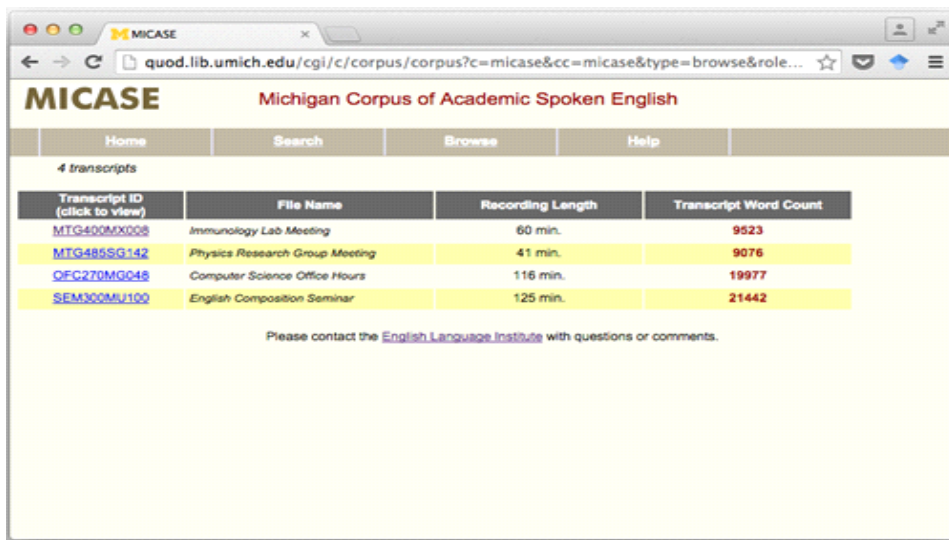
All, Junior Faculty, Junior Graduate Studentds, ...

▷Interactivity Rating:

All, Highly interactive, Highly monologic, Mostly interactive, ...

<그림 8> MICASE 온라인 탐색 도구

다음은 "Native speaker status: Non-native speaker", "First language: Korean", "Interactive Rating: Highly interactive"로 설정하고 나머지 필드는 All로 지정한 후 "Submit Browse" 버튼을 눌렀을 때 얻어지는 결과이다.



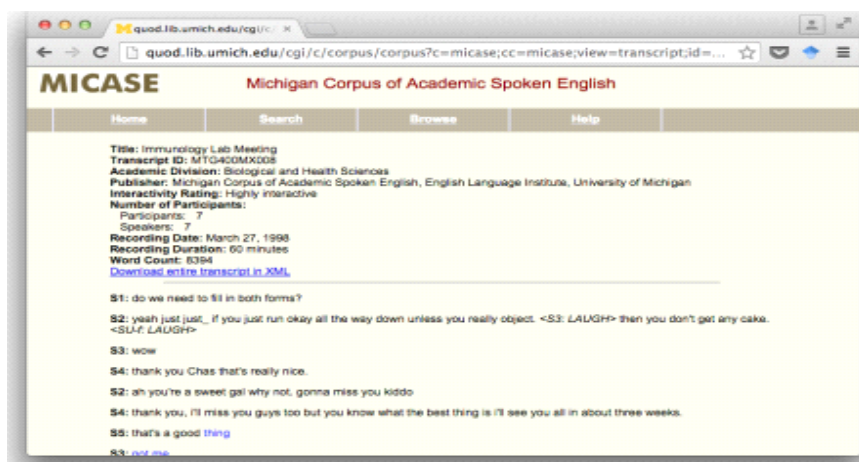
The screenshot shows the MICASE (Michigan Corpus of Academic Spoken English) website. The browser address bar shows the URL: quod.lib.umich.edu/cgi/c/corpus/corpus?c=micase&cc=micase&type=browse&role... The page title is "MICASE Michigan Corpus of Academic Spoken English". Below the title is a navigation bar with links: Home, Search, Browse, Help. The main content area is titled "4 transcripts" and contains a table with the following data:

Transcript ID (click to view)	File Name	Recording Length	Transcript Word Count
MTG400MX008	Immunology Lab Meeting	60 min.	9523
MTG485SG142	Physics Research Group Meeting	41 min.	9076
QEG270MG048	Computer Science Office Hours	116 min.	19977
SEM300MU100	English Composition Seminar	125 min.	21442

Below the table, it says: "Please contact the [English Language Institute](#) with questions or comments."

<그림 9> MICASE 전사 자료 탐색 목록

4개의 전사 텍스트의 목록이 고유번호, 제목, 녹음 분량, 전사 단어 수에 대한 정보로 제시되었다. 텍스트의 고유번호를 클릭하면 원문을 볼 수 있다. 다음은 첫 번째로 제시된 전사 자료의 고유번호를 클릭하였을 때 볼 수 있는 결과이다.



The screenshot shows the MICASE website displaying the HTML view of a transcript. The browser address bar shows the URL: quod.lib.umich.edu/cgi/c/corpus/corpus?c=micase;cc=micase;view=transcript;id=... The page title is "MICASE Michigan Corpus of Academic Spoken English". Below the title is a navigation bar with links: Home, Search, Browse, Help. The main content area displays the following information:

Title: Immunology Lab Meeting
 Transcript ID: MTG400MX008
 Academic Division: Biological and Health Sciences
 Publisher: Michigan Corpus of Academic Spoken English, English Language Institute, University of Michigan
 Interactivity Rating: Highly interactive
 Number of Participants:
 Participants: 7
 Speakers: 7
 Recording Date: March 27, 1998
 Recording Duration: 60 minutes
 Word Count: 9594
[Download entire transcript in XML](#)

The transcript text is displayed below, with speaker labels (S1, S2, S3, S4, S5) and their corresponding utterances. The text is as follows:

S1: do we need to fill in both forms?
 S2: yeah just just... if you just run okay all the way down unless you really object. <S3: LAUGH> then you don't get any cake. <S4: LAUGH>
 S3: wow
 S4: thank you Chas that's really nice.
 S2: oh you're a sweet gal why not, gonna miss you kiddo
 S4: thank you, I'll miss you guys too but you know what the best thing is I'll see you all in about three weeks.
 S5: that's a good thing
 S3: cool, yes

<그림 10> MICASE 전사 자료 HTML 보기 예제

말뭉치 내용을 HTML 형태로 보기 좋게 제시하고 있다. XML 파일의 헤더 정보의 일부가 상단에 제시되고, 이어서 전사 텍스트 본문이 발화자 정보와 함께 제시되고 있다. 예시된 자료의 경우, 7명의 발화자가 참여한 60분간의 음성 자료에서 8394개 단어로 이루어진 전사 자료를 얻은 것임을 알 수 있다. XML 파일 안에는 각 발화자에 대한 상세 정보를 포함하여 해당 전사 자료에 대한 자세한 정보가 포함되어 있다. “Download entire transcript in XML” 링크를 클릭하여 XML 형식의 말뭉치를 내려 받을 수도 있다. XML 파일을 보면 다음과 같다.

```
<TEI.2 ID="mtg400mx008">
  <TEIHEADER>
    <FILEDESC>
      <TITLESTMT>
      <WORDCOUNT 8394 /WORDCOUNT>
      <PUBLICATIONSTMT>
      <SOURCEDESC>
        <RECORDINGSTMT>
          <RECORDING DUR="60" />
        </RECORDINGSTMT>
      </SOURCEDESC>
    </FILEDESC>
    <PROFILEDESC>
      <KEYWORDS>
      <LANGUAGE>
      <PARTICDESC>
        <PERSONGRP SIZE="7" ROLE="Participants"/>
        <PERSONGRP SIZE="7" ROLE="Speakers"/>
        <PERSON ID="S1" ROLE="JF, Second Presenter" AGE="3" RESTRICT="NONE" SEX="M" LANG="NNS">
          <FIRSTLANG> CHI </FIRSTLANG>
        </PERSON>
        <PERSON ID="S2" ROLE="SF, Research Group Chair" AGE="4" RESTRICT="NONE" SEX="M" LANG="NS"/>
        <PERSON ID="S3" ROLE="JF, Group Member" AGE="3" RESTRICT="NONE" SEX="F" LANG="NNS">
          <FIRSTLANG> KOR </FIRSTLANG>
        </PERSON>
        <PERSON ID="S4" ROLE="SG, Main Presenter" AGE="2" RESTRICT="NONE" SEX="F" LANG="NS"/>
        <PERSON ID="S5" ROLE="ST, Technician" AGE="3" RESTRICT="NONE" SEX="M" LANG="NS"/>
        <PERSON ID="S6" ROLE="JG, Group Member" AGE="2" RESTRICT="NONE" SEX="M" LANG="NS"/>
        <PERSON ID="R1" ROLE="SF, MICASE Researcher" AGE="4" RESTRICT="NONE" SEX="M" LANG="NSO"/>
        <PERSON ID="SU-5" ROLE="UN, Unknown" AGE="0" RESTRICT="NONE" SEX="M" LANG="NS"/>
        <PERSON ID="SU-6" ROLE="UN, Unknown" AGE="0" RESTRICT="NONE" SEX="M" LANG="NS"/>
        <PERSON ID="SU-7" ROLE="UN, Unknown" AGE="0" RESTRICT="NONE" SEX="U" LANG="NS"/>
        <PERSON ID="SU-f" ROLE="UN, Unknown" AGE="0" RESTRICT="NONE" SEX="F" LANG="NS"/>
        <PERSON ID="SU-m" ROLE="UN, Unknown" AGE="0" RESTRICT="NONE" SEX="M" LANG="NS"/>
      </PARTICDESC>
    </PROFILEDESC>
  </TEIHEADER>
  <TEXT>
    <P>
      <S>
        <S>
          <S>
            <S>
              <S>
                <S>
                  <S>
                    <S>
                      <S>
                        <S>
                          <S>
                        </S>
                      </S>
                    </S>
                  </S>
                </S>
              </S>
            </S>
          </S>
        </S>
      </S>
    </P>
  </TEXT>
</TEI.2>
```

<그림 11> MICASE 전사 자료 XML 예시

위에서 볼 수 있듯이 MICASE 전사 자료는 TEI 형식으로 되어있다. 헤더에 발화자에 관한 정보와 발화 상황에 관한 정보가 기록되어 있음을 확인할 수 있다.

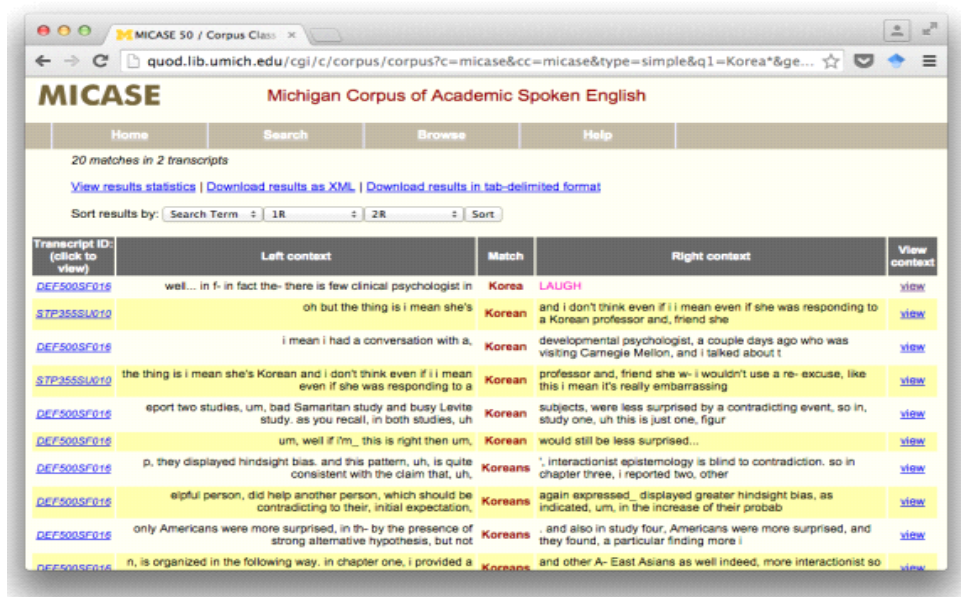
○ MICASE 검색 서비스

검색 서비스에서도 발화자 속성(Speaker Attributes)과 전사 텍스트 속성(Transcript Attributes)을 선택한 후 검색어를 입력하여 문맥 색인 결과를 얻을 수 있다. 검색에서는 아무개(wildcard) 문자를 사용할 수 있다. 예를 들어, “walk*”라고 검색어를 지정하면 “walk, walks, walked, walking” 등이 모두 검색된다. 전체 말뭉치를 대상으로 검색하려고 할 때에는 발화자 속성과 전사 텍스트 속성을 따로 선택하지 않고 기본값으로 두면 된다.

<그림 12> MICASE 검색 서비스 화면

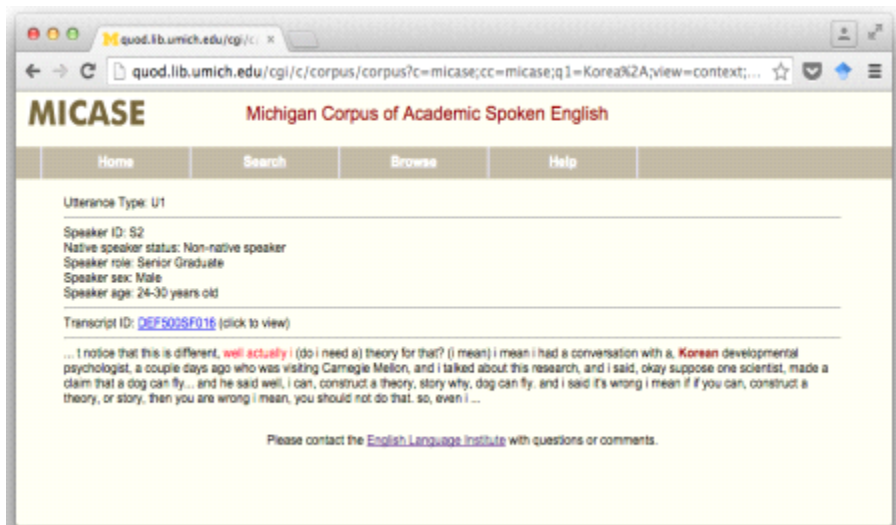
검색어 지정 후 “Submit Search” 버튼을 누르면 검색 결과가 문맥 색인의 형태로 제시된다.

- 검색어가 나타난 총 횟수와 검색어가 사용된 텍스트의 개수가 출력된다. “Korean*”을 검색어로 하였을 때에는 20개의 용례가 발견되었으며 2개의 텍스트가 발견되었다.
- 검색 결과로 얻어진 문맥 색인은 “Download results as XML” 버튼을 이용하여 XML 형식으로 다운로드 받을 수 있으며 “Download results in tab-delimited format” 버튼을 이용하여 탭으로 구분된 텍스트 파일 형식으로 다운로드 받을 수 있다.
- 문맥 색인의 정렬 기능도 제공한다. 검색어와 함께 좌우 문맥 각 5단어를 기준으로 정렬할 수 있다.



<그림 13> MICASE 검색 결과

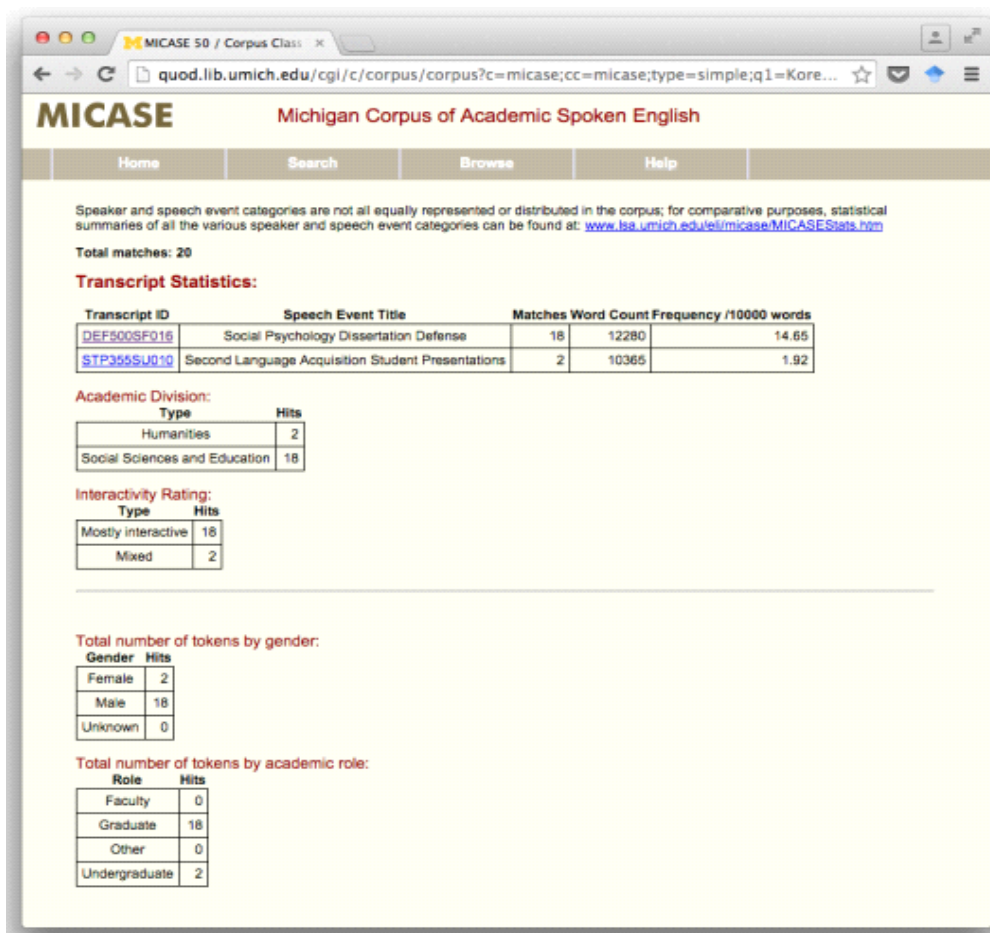
결과에서 각 항목의 “view” 버튼을 클릭하여 해당 용례의 속성을 볼 수 있으며 문맥을 좀 더 넓혀 살펴볼 수 있다. 예를 들어, 아래와 같이 특정 용례에 대해 발화자 고유번호가 “S2”이며 “Non-native speaker”이며 대학원 고학년, 남성, 24-30세라는 정보가 제공된다. 아래쪽에는 조금 더 확장된 문맥이 노출된다.



<그림 14> MICASE 검색 결과에서 개별 용례의 “view” 결과

검색 결과 화면에서 상단의 “View results statistics” 링크를 클릭하면 검색어에 대한 상세한 통계 요약 정보가 제공된다.

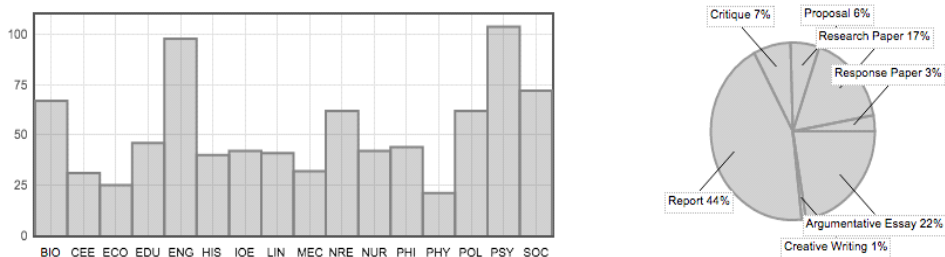
- 전사텍스트별로 발견된 용례의 개수가 제시된다. 아래 예의 경우에는 2개의 텍스트에서 “Korean*”이 발견되었으며 한 텍스트에서 18개, 다른 텍스트에서 2개가 발견되었다. 각 텍스트의 단어 규모와 함께 1만 단어당 검색어의 발견 비율도 제시된다.
- 학문 분과별로 용례 발견 횟수가 제시된다.
- 상호작용 수준에 따른 용례 발견 횟수가 제시된다.
- 발화자와 관련하여서는 성별에 따라, 역할에 따라 용례가 몇 개가 발견되었는지 제시된다.



<그림 15> MICASE 검색 결과 통계 요약 정보

○ MICUSP

MICUSP는 고학년 학부생과 대학원생이 작성한 논술, 비평, 제안서, 보고서, 연구 논문 등의 글쓰기 자료를 수집한 것이다. 약 800여개의 텍스트로 이루어졌으며 160만 단어 규모이다. A학점을 받은 것 또는 제안서와 같은 경우에는 통과된 것들을 수집하였다. 작성자는 철학, 언어학, 심리학, 사회학, 경제학, 물리학, 토목공학, 산업공학, 간호학, 생물학 전공자들이다. 하나의 텍스트는 500~10,000 단어로 이루어져 있다. 전공 분야별 텍스트의 개수와 글 유형별 비율을 보면 다음과 같다.



<그림 16> MICUSP 말뭉치의 전공 분야별 텍스트 개수와 글쓰기 유형별 텍스트 비율

각 텍스트는 학년, 영어 모어 여부, 텍스트 특성, 글의 유형, 전공 분야에 따라 분류되어 있다. 각 세부 항목을 선택하여 체크 표시를 하면 조건에 맞는 텍스트에 대한 통계가 전공별 막대그래프와 글 유형별 원그래프에 즉시 반영된다. 세부 항목을 보면 다음과 같다.

Student Levels:

- ☐ Senior Undergrad (G0)
- ☐ 1st Year Graduate (G1)
- ☐ 2nd Year Graduate (G2)
- ☐ 3rd Year Graduate (G3)

Nativeness:

- ☐ Native English Speaker
- ☐ Non-native English Speaker

Text Features:

- ☐ Abstract
- ☐ Definitions
- ☐ Discussion of results
- ☐ Literature review
- ☐ Methodology section
- ☐ Problem-solution pattern
- ☐ Reference to sources
- ☐ Tables, graphs or figures

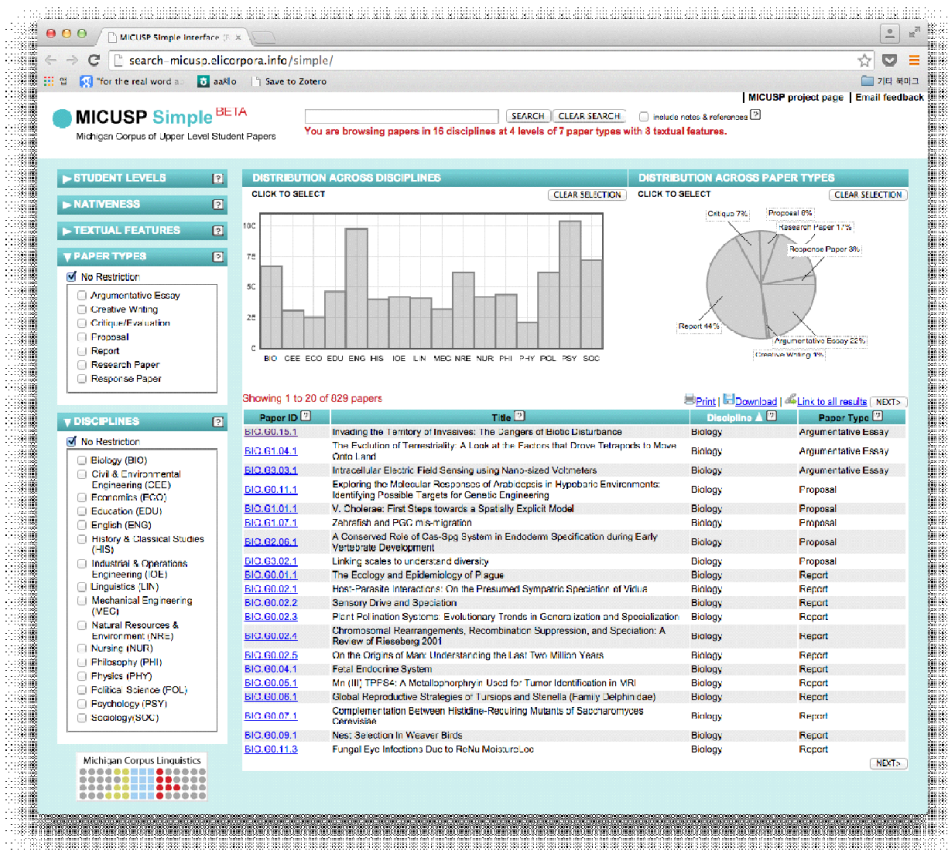
Paper Types:

- ☐ Argumentative Essay
- ☐ Creative Writing
- ☐ Critique/Evaluation
- ☐ Proposal
- ☐ Report
- ☐ Research Paper
- ☐ Response Paper

Disciplines:

- ☐ Biology (BIO)
- ☐ Civil &Environmental Engineering (CEE)
- ☐ Economics (ECO)
- ☐ Education (EDU)
- ☐ English (ENG)
- ☐ History &Classical Studies (HIS)
- ☐ Industrial &Operations Engineering (IOE)
- ☐ Linguistics (LIN)
- ☐ Mechanical Engineering (MEC)
- ☐ Natural Resources &Environment (NRE)
- ☐ Nursing (NUR)
- ☐ Philosophy (PHI)
- ☐ Physics (PHY)
- ☐ Political Science (POL)
- ☐ Psychology (PSY)
- ☐ Sociology(SOC)

화면 왼쪽에 텍스트 세부 정보의 목록이 제시되며 각 항목을 클릭하여 선택하거나 선택을 해제할 수 있다. 주 화면의 상단에 전공별 통계와 글 유형별 통계가 제시되며 그 아래에 조건에 부합하는 목록이 제시된다. 왼쪽의 텍스트 세부 항목의 조건을 변경하면 즉시 통계 그래프와 텍스트 목록도 변경된다. 목록은 엑셀 파일로 다운로드 받을 수 있다.



<그림 17> MICASE와 MICUSP의 웹 탐색 및 검색 서비스

조건에 부합하는 텍스트 목록에는 각 항목의 고유번호(Paper ID)와 함께 제목, 전공 분야, 글 유형이 제시된다. 여기서 “Paper ID” 링크를 클릭하면 텍스트 본문을 볼 수 있다.

- “View original paper (PDF)” 링크를 클릭하면 수집된 원래 글을 볼 수 있다. 원래 글의 PDF 파일은 그림과 표도 포함하고 있으며 다양한 형식으로 편집되어 있을 수 있다.
- 상단 우측에는 해당 글에 대한 워드 클라우드(word cloud)가 나타난다. 워드 클라우드는 텍스트에 사용된 단어의 빈도를 이용하여 주요 키워드를 시각화하는 방법이다. 어떤 단어가 많이 사용되고 있는지 쉽게 파악할 수 있다. 일반적으로 워드 클라우드는 기능어를 제외하고 보여주는 데에 비해 여기에서는 전치사, 관사, 대명사 등 문법적 요소들을 모두 포함하여 보여주고 있다.
- 상단 좌측에는 상세 정보가 제시된다. 다음 예에서는 전공 분야는 언어학, 학부 마지막 학년, 여성, 영어 모어 화자가 작성한 글임을 알 수 있고 에세이 유형의 글이며 참고문헌 목록이 포함되어 있음을 알 수 있다. 총 3829 단어로 작성된 글임을 알 수 있다.

LIN.G0.12.1

National Identity and Language Education Policy

View original paper (PDF)

Discipline: Linguistics

Student Level: Final Year Undergraduate

Sex: Female

Native speaker status: NS

Paper type: Essay

Paper contains following features: Reference to sources

Word count: (3829 including notes and references)

June 2006 is http www

state school reform

programs

language minority students

lep students

immigrant children

education

the federal

federal

bilingual education

inofenglish

schools

language education

legislation

the

education policy

language

to

and

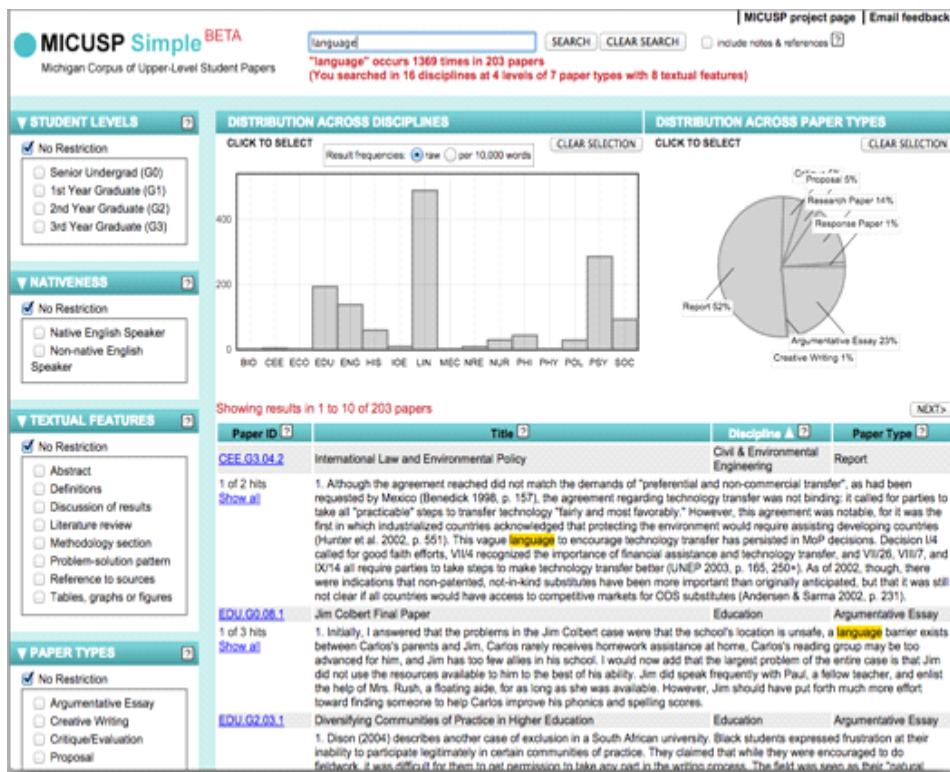
that

National Identity and Language Education Policy

Regardless of ideological leanings, most Americans agree that education is a hugely important factor that impacts not only democratic civic participation, but also social mobility and economic growth. Education is taken seriously as an integral part of the American Dream, and public education policy has always been a hotly contentious issue. People feel strongly about education and the effects of educational policy on their kids and their country; thus, education is an area that has traditionally been as closely linked as possible with the people, governed in large part by state and local governments. The education of immigrant children is an area of education policy that has been even more controversial than most, with its various contending programs and arguments over discrimination, inclusion, language rights, identity politics, assimilation and civil rights. It is a debate both at the federal level and at the individual state and local

<그림 18> MICUSP 개별 텍스트

검색 서비스도 제공된다. 검색어의 분포를 막대그래프와 원그래프로 시각화하여 제공한다. 검색어를 포함하고 있는 텍스트의 목록이 나열되며 원문과 함께 검색어를 강조하여 보여준다. 다음 예는 “language”로 검색한 결과를 보여준다. 203개 텍스트에서 총 1369번 사용되었다. 언어학, 심리학, 교육, 영어영문학 분야의 텍스트가 많다는 것을 확인할 수 있다. 글 유형별로는 보고서와 논술이 주를 차지하고 있음을 알 수 있다.

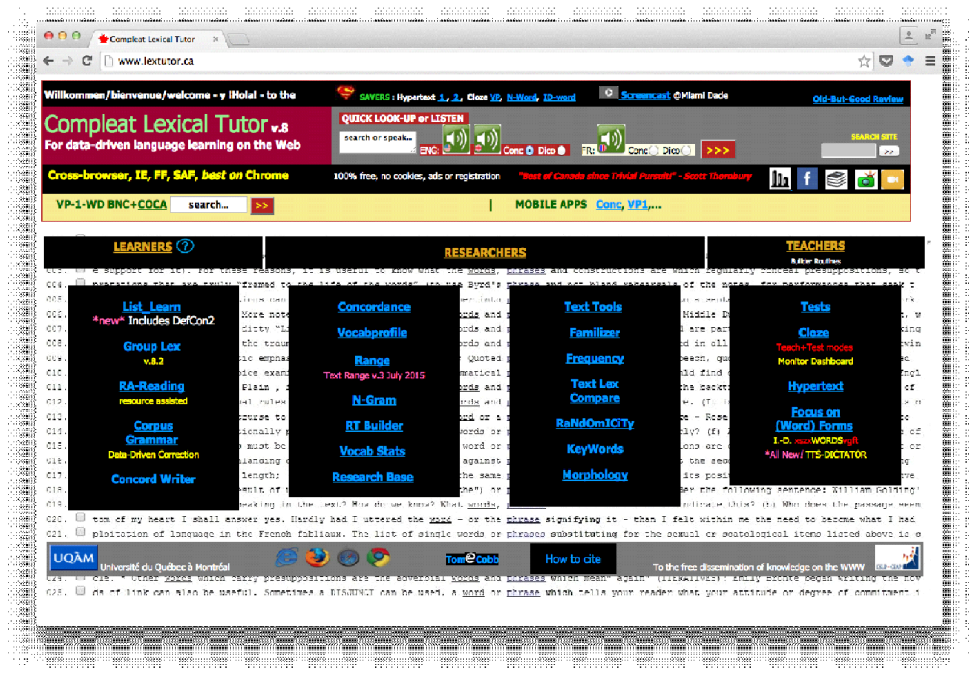


<그림 19> MICUSP 검색 기능

(2) Compleat Lexical Tutor (<http://www.lextutor.ca/>)

캐나다 몬트리올 퀘벡 대학교의 언어교육학과의 Thomas Michael Cobb 교수 주도로 운영하고 있는 Compleat Lexical Tutor 사이트는 ‘웹상에서 자료 주도 언어 학습을 위하여(For data-driven language learning on the Web)’라는 기치 아래 학습자, 연구자, 교사를 위한 온라인 말뭉치 도구를 제공하고 있다.

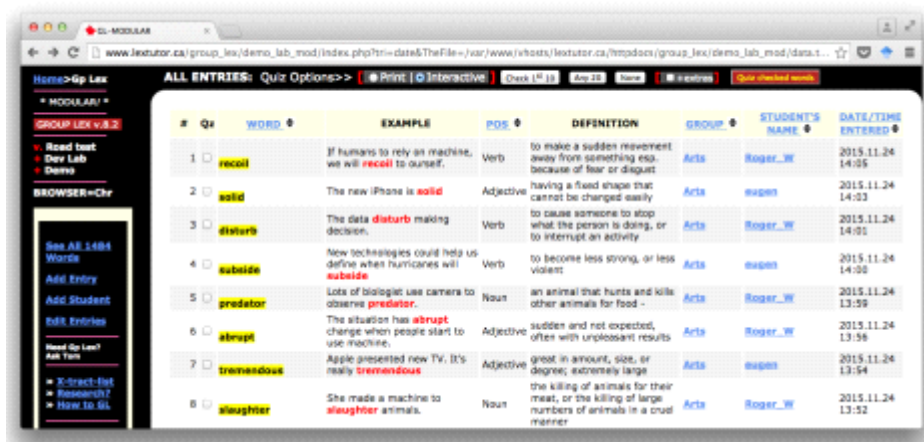
크게는 학습자(Leaners)를 위한 서비스, 연구자(Researchers)를 위한 서비스, 교사(Teachers)를 위한 서비스로 구분되어 있다. 학습자들은 말뭉치를 참조하여 작문을 하거나 온라인 테스트를 해볼 수 있으며 교사들은 주어진 텍스트에서 시험문제를 생성하는 기능을 이용할 수 있다. 연구자들에게는 빈도와 용례를 중심으로 하는 다양한 도구들이 제공된다.



<그림 20> Lextutor(Compleat Lexical Tutor) 웹사이트의 첫 화면

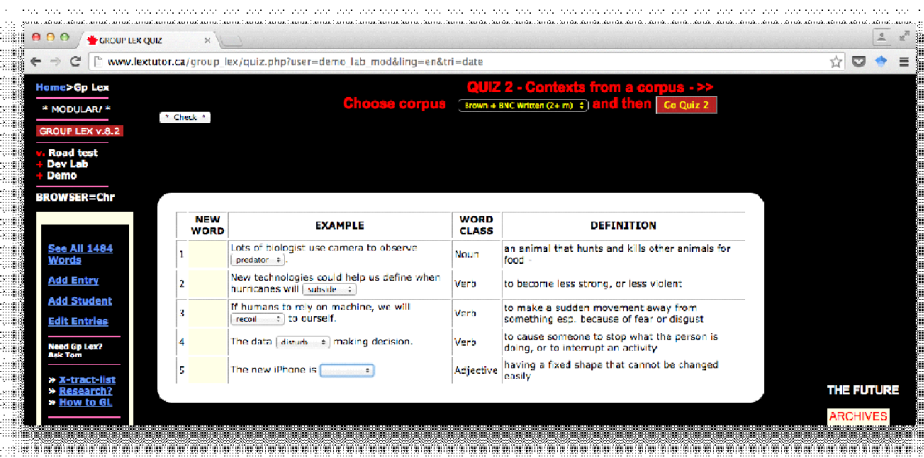
○ Group Lex

학습자용 서비스의 하나인 “Group Lex”는 일종의 집단 지성을 활용한 어휘집 구축 도구 또는 문맥 색인 생성 도구라 할 수 있다. 학습자들은 자신의 계정을 생성한 뒤에 자유롭게 어휘 항목을 추가할 수 있다. 표제어의 형태, 용례, 품사, 뜻풀이, 어휘 의미 부류를 등록할 수 있다. <그림 21>은 현재 등록되어 있는 항목의 예를 보여주고 있다. 이러한 시스템의 시작은 2010년경부터 시도된 것으로 확인되나 본격적으로 항목이 등록되기 시작한 것은 2015년에 들어서이다. 현재 등록된 항목은 1500여개에 불과하나 최근 많은 항목이 지속적으로 추가되고 있다.



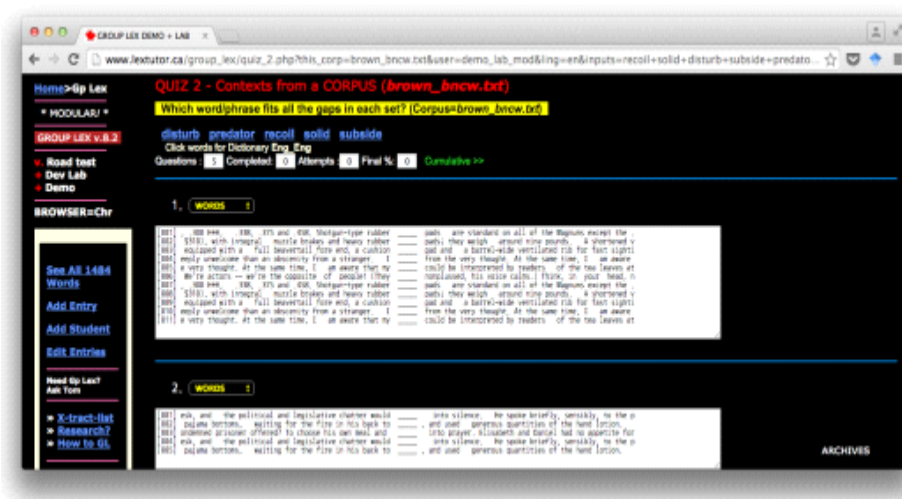
<그림 21> Lextutor의 Group Lex 서비스 예시

각 필드별로 정렬하여 보거나 특정 부류만 선별적으로 목록을 생성하는 등의 기능도 제공된다. 서비스 중 주목되는 것은 퀴즈 기능이다. 학습자들은 스스로 퀴즈를 생성할 수 있으며 즉시 테스트해 볼 수 있다. 교사들이 어휘 시험 문제를 만들고자 할 때에도 이용할 수 있다. 테스트에 포함되기를 원하는 항목을 체크한 후에 퀴즈를 생성할 수 있으며 “+extras”를 체크하면 선택된 것 이외에 어휘가 시험 문제에 섞여서 출제된다. <그림 22>는 1단계 퀴즈의 예시이다. 상단 왼쪽의 “Check” 버튼을 클릭하면 정답과 오답을 알려준다.



<그림 22> Lextutor 사이트의 Group Lex 서비스 중 퀴즈 1단계

학생들이 스스로 작성한 용례와 뜻풀이를 그대로 이용하여 생성되는 퀴즈1에 이어 말뭉치에서 추출한 용례로 문제를 생성하는 퀴즈2도 진행할 수 있다. 퀴즈1에서 오른쪽 상단의 “QUIZ 2 - Context from a corpus” 링크를 클릭하여 이동할 수 있다. 이때에 용례를 추출할 말뭉치를 선택할 수 있다. <그림 23>은 퀴즈2의 예시이다. 키워드가 비어 있는 문맥 색인이 주어지면 공통으로 들어갈 어휘를 맞히는 문제가 출제된다.



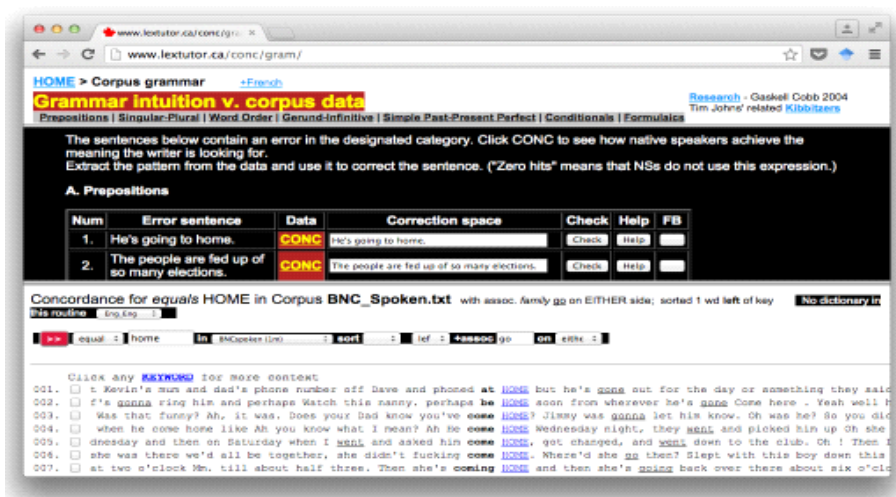
<그림 23> Lextutor 사이트의 Group Lex 서비스 중 퀴즈 2단계

○ Corpus Grammar

학습자용 서비스 중 “Corpus Grammar”는 문법적 직관(Grammar intuition) 대 말뭉치 자료(corpus data)라는 관점에서 만들어진 서비스이다. 오류가 있는 문장을 수정하는 문제 형태로 제시되며 말뭉치 자료를 참조할 수 있도록 구성되어 있다.

오류가 있는 문장의 목록은 고정되어 있다. 전치사, 단수-복수, 어순, 동명사-부정사, 시제, 조건법, 정형표현 등의 주제로 나누어 예제가 등록되어 있다. 현재 많은 예가 등록되어 있지는 않으며 새로운 예제를 사용자가 자유롭게 등록할 수 있는 기능을 제공하지는 않고 있다. 영어 학습자에게서 흔히 발견되는 오류를 수집하여 목록을 만든 것으로 보인다. <그림 24>는 “He’s going to home”이라는 문장의 경우를 예시한 것이다. 전치사 없이 “He’s going home”으로 고쳐야 올바른 문장이 된다. 말뭉치를 통해 ‘come’, ‘go’ 등의 이동동사 뒤에 전치사 없이

‘home’이 쓰이는 것을 확인할 수 있다.



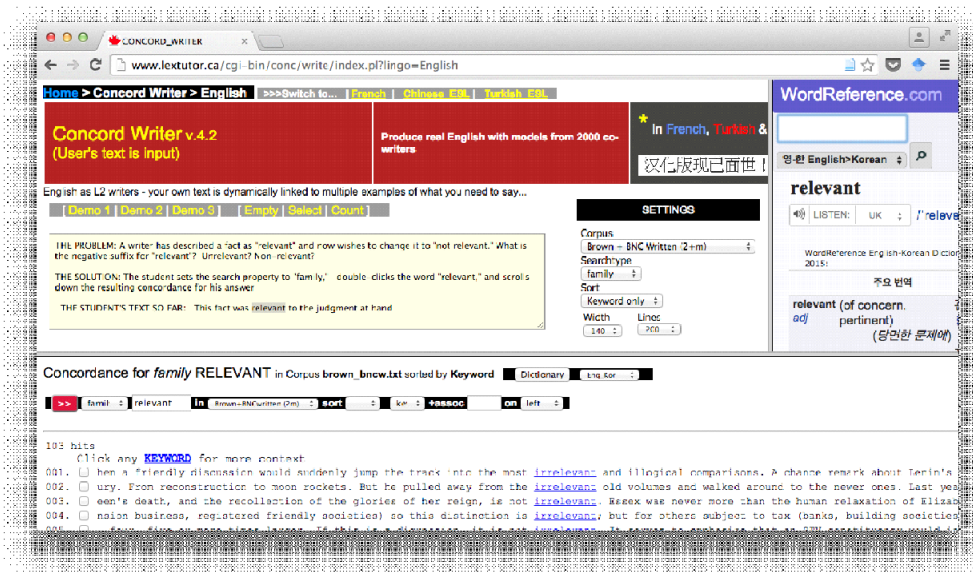
<그림 24> Lextutor 사이트의 Corpus Grammar

○ Concord Writer

학습자용 서비스 중 “Concord Writer”는 학습자들이 작문을 할 때 문맥 색인의 도움을 받을 수 있도록 해주는 서비스이다. 사용자는 텍스트 영역에 작문을 할 수 있으며 사전과 문맥 색인의 도움을 받을 수 있다.

사전을 사용자가 선택할 수 있으며 영어-영어, 영어-프랑스어, 영어-스페인어, 영어-독일어, 영어-터키어, 영어-중국어, 영어-일본어, 영어-한국어, 영어-아랍어 사전이 제공된다. 문맥 색인을 생성하는 데에 사용할 말뭉치도 사용자가 지정할 수 있다. 브라운 코퍼스, BNC(British National Corpus) 문어, 구어, 의학, 인문학, 법학, 사회과학 등과 대통령 연설문, 학술 논문 초록 등 고급 학습자가 작문할 때 참조할 수 있는 말뭉치들을 제공한다.

검색 방식을 완전 일치, 시작 일치, 끝 일치 등으로 지정할 수 있으며 검색 결과를 정렬하는 기준으로 키워드, 왼쪽 문맥, 오른쪽 문맥으로 지정할 수 있다. 작문 중인 텍스트에서 원하는 단어를 더블클릭하면 해당 단어에 대한 문맥 색인이 아래쪽에 출력된다. 두 단어 이상을 키워드로 지정할 수도 있으며 자동으로 마지막 1~4 단어를 키워드로 지정할 수도 있다.



<그림 25> Lextutor 사이트의 Concord Writer

○ Concordance

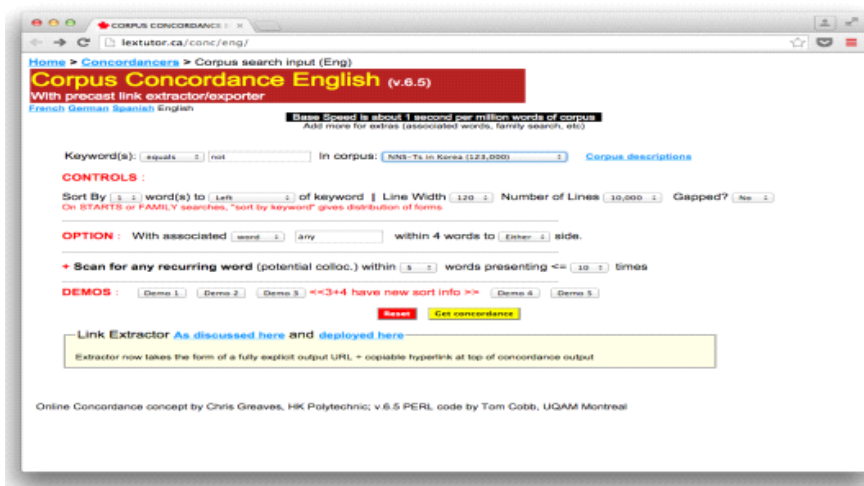
연구자용 서비스인 “Concordance”는 영어, 프랑스어, 독일어, 스페인어 말뭉치를 탑재하여 문맥 색인을 생성해 주는 서비스이다. 영어의 경우 30여 종의 말뭉치의 검색 서비스를 제공하고 있다. “Mobile Concordance”를 통해 모바일 기기에서 문맥 색인 서비스도 제공하고 있다.

영어 말뭉치 중에는 학습자 말뭉치도 일부 포함되어 있다. Brown Corpus, BNC, COCA가 주요 말뭉치로 제공되고 있으며 JPU Learner Corpus, TC Learner (Student/Teacher) Corpus, Korean EFL teachers Corpus 등 영어 학습자 말뭉치를 일부 제공하고 있다. 여러 학술 분야의 말뭉치, TV 방송 말뭉치 등 분야별 말뭉치도 제공하고 있다.



<그림 26> Lextutor 사이트 Concordance 서비스에서 제공되는 말뭉치 목록

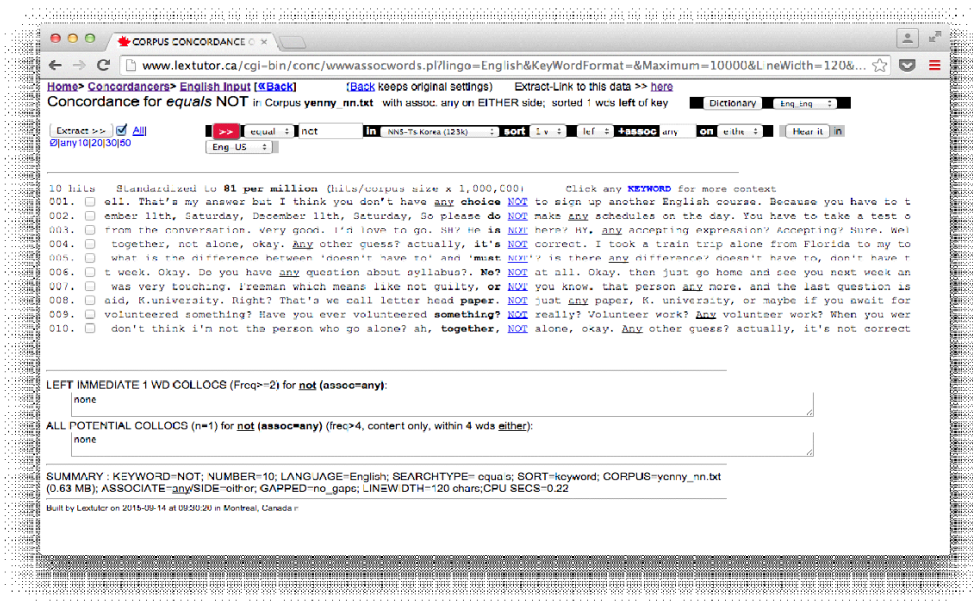
기본적으로 KWIC(Keyword In Context) 형태의 문맥 색인 검색 기능을 제공하며 여러 가지 추가 옵션으로 상세한 설정을 할 수 있다. 연관된 단어를 지정하여 검색어의 왼쪽, 오른쪽, 또는 양쪽으로 4단어 이내에 나타나는 용례로 결과를 제한할 수 있다. 불완전하기는 하나 잠재적 결합어(연어)를 스캔하는 기능을 제공한다.



<그림 27> Lextutor의 문맥 색인 검색 화면

검색 결과는 <그림 28>과 같은 형태로 제공된다. 검색 결과에도 여러 가지 옵션을 제공하여 정렬을 하거나 사전에서 추가 정보를 찾아보는 등의 작업을 할 수 있다.

- 검색 결과에서 키워드를 클릭하면 키워드를 포함한 원문 텍스트의 일부분을 팝업 창에서 확인할 수 있다. 이때 해당 키워드가 포함된 문장이 표시된다.
- 'Dictionary' 버튼을 클릭하면 검색된 단어에 대해서 영어-영어, 영어-프랑스어, 영어-독일어, 영어-이탈리아어, 영어-러시아어, 영어-중국어, 영어-일본어, 영어-한국어, 영어-아랍어 등의 사전을 볼 수 있다.
- 'Hear it' 버튼을 클릭하면 키워드로 입력된 단어의 미국식, 영국식, 호주식 원어민 발음을 선택하여 들을 수 있다.
- 검색 결과 중 원하는 것을 선별하여 체크한 후 'Extract' 버튼을 눌러 선택 항목으로만 구성된 목록을 추출할 수 있다.
- 단순 텍스트 형태로 결과물을 제공한다.



<그림 28> Lextutor의 문맥 색인 검색 결과

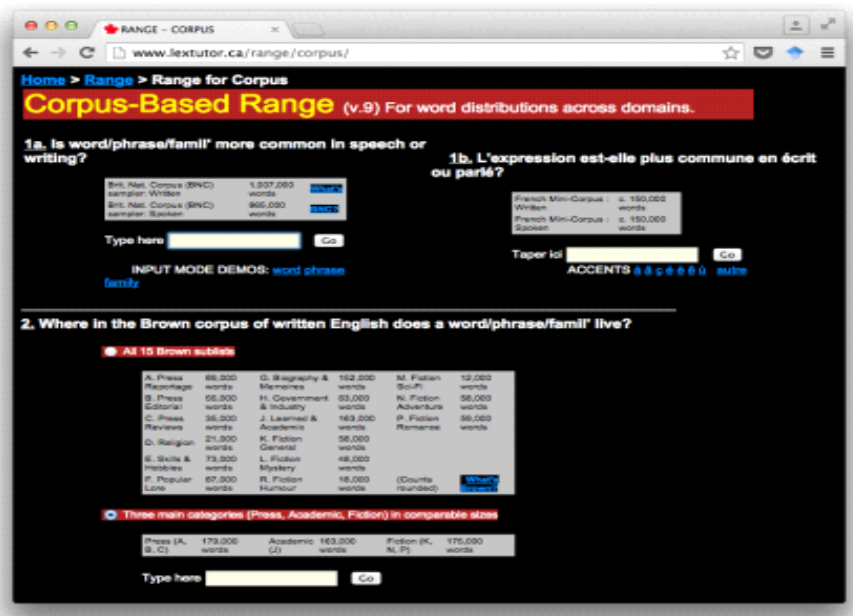
검색 결과 중 하단에는 “RIGHT IMMEDIATE 1 WD COLLOCS”라는 창에서 해당 키워드의 바로 오른쪽에 인접한 단어와 그 빈도의 목록을 보여주며 “ALL POTENTIAL COLLOCS” 창에서는 잠재적 연어로서 좌우 5단어 문맥 이내에서 빈도가 5이상인 단어들의 목록을 보여준다. 오른쪽 문맥을 볼지 왼쪽 문맥을 볼지는 사용자가 지정할 수 있다. <그림 29>는 “make”로 검색하였을 때 연어 후보의 목록을 보여주는 결과이다.



<그림 29> Lextutor 사이트의 Concordance 서비스 연어 후보 목록

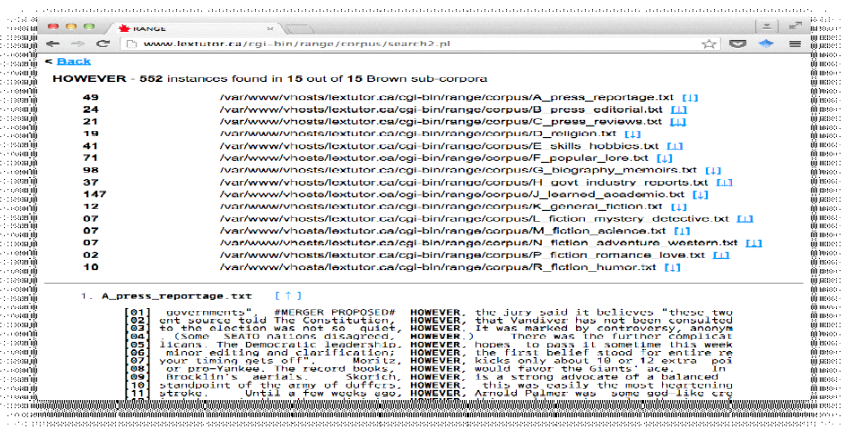
○ Range

연구자용 서비스의 하나인 “Range”는 단어를 비롯한 어휘 단위 표현이 둘 이상의 텍스트에서 나타나는 분포를 보여주는 서비스이다. Lextutor에서 제공하고 있는 말뭉치에서 분포를 살펴볼 수 있으며 사용자가 텍스트를 업로드하여 분석할 수도 있다. 기본적으로 제공하는 말뭉치를 이용하여 분석하는 경우 간단하게는 구어와 문어에서 차이를 볼 수 있다. 더 상세하게는 브라운 코퍼스의 구어 텍스트를 신문 보도, 비평, 종교, 취미, 전기, 정부, 산업, 학술, 문학, 미스터리, 과학소설 등으로 구분하여 비교할 수도 있다.



<그림 30> Lextutor 사이트의 Range 서비스

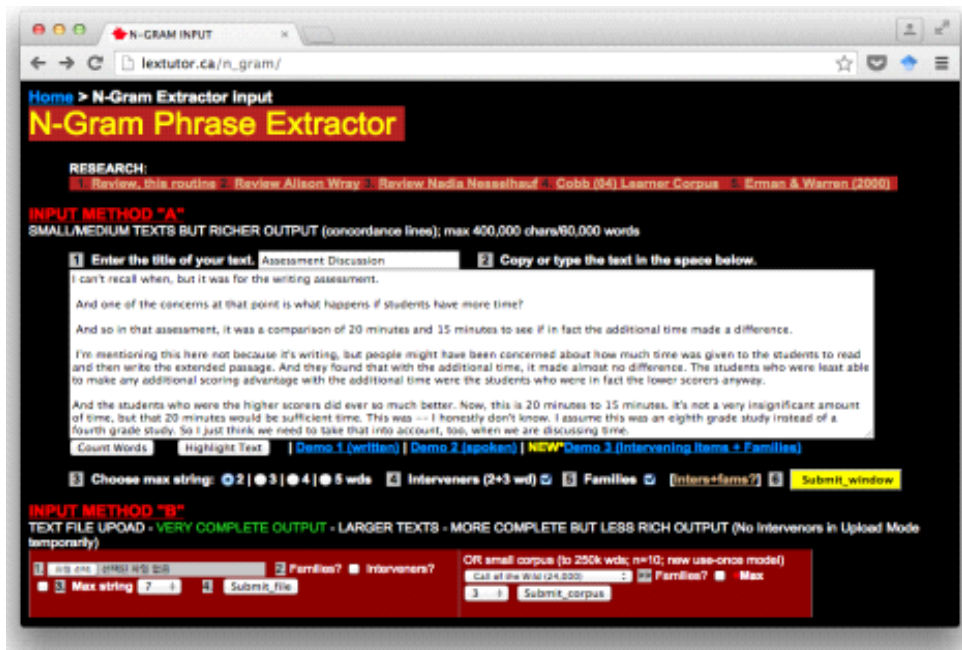
검색하고자 하는 단어 또는 구문을 입력하고 “Go” 버튼을 누르면 결과가 출력된다. 결과 페이지의 상단에는 각 텍스트별로 출현 빈도가 요약된다. 그 아래에 이어서 각 출현 용례가 나열된다. <그림 31>은 “however”로 검색한 결과의 예시이다. 학술적인 텍스트에 많이 나타나며 소설에는 드물게 나타나는 것을 확인할 수 있다. 물론 각 텍스트의 크기가 다르므로 단순히 절대 빈도를 비교해서는 안 된다.



<그림 31> Lextutor 사이트의 Range 서비스 검색 결과

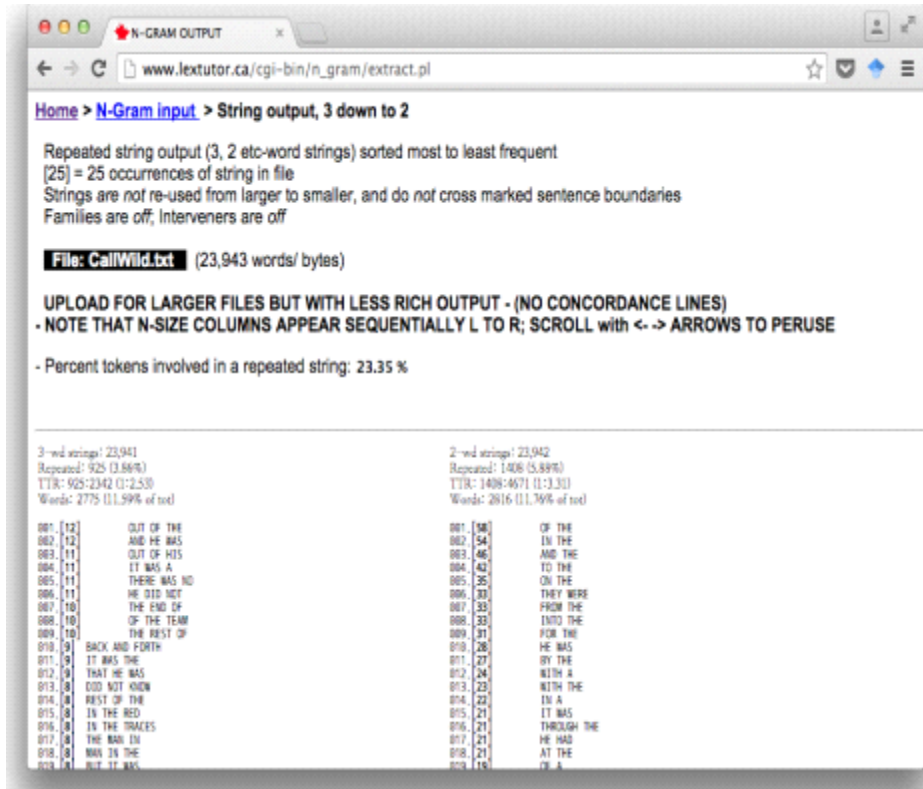
○ N-Gram

연구자용 서비스의 하나인 “N-Gram”은 N개의 어절로 이루어진 구문을 추출해 주는 기능을 제공한다. 사용자는 텍스트 창에 분석을 원하는 텍스트를 복사하여 붙여 넣거나 파일로 업로드한 후 분석할 수 있다. 텍스트를 붙여 넣어 분석할 때에는 최대 40만자/6만 어절까지로 제한이 있다.



<그림 32> Lextutor 사이트의 N-Gram 서비스

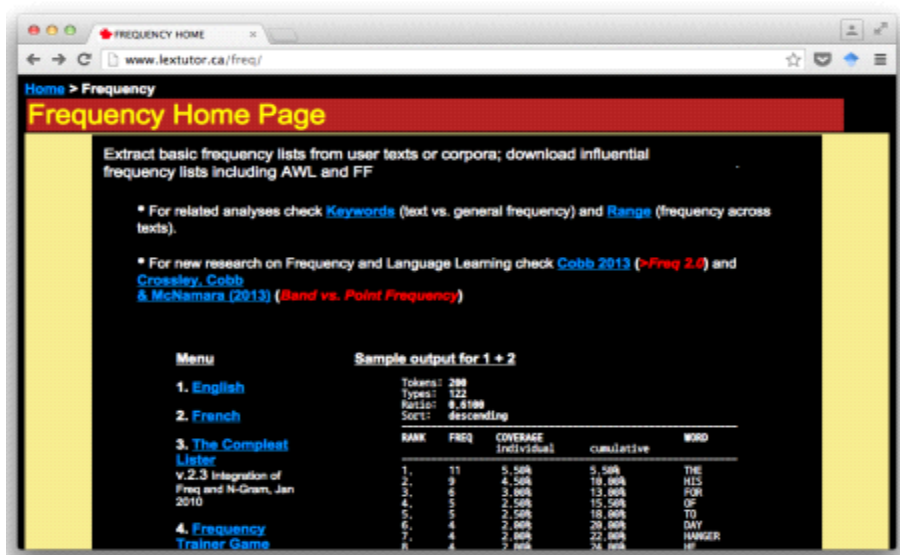
텍스트 창에 붙여 넣기를 통해 분석하는 경우 2-Gram에서 5-Gram까지 분석할 수 있다. 파일을 업로드하는 경우에는 10-Gram까지 분석할 수 있다. 분석 결과는 다음과 같은 형식으로 출력된다. 약 2만 3천 어절의 ‘Call Wild’라는 텍스트를 대상으로 하였으며 최대 3-Gram으로 지정하였다. 따라서 3-Gram과 2-Gram의 목록을 생성하며 각각은 빈도순으로 제시된다.



<그림 33> Lextutor 사이트의 N-Gram 서비스 분석 결과

○ Frequency

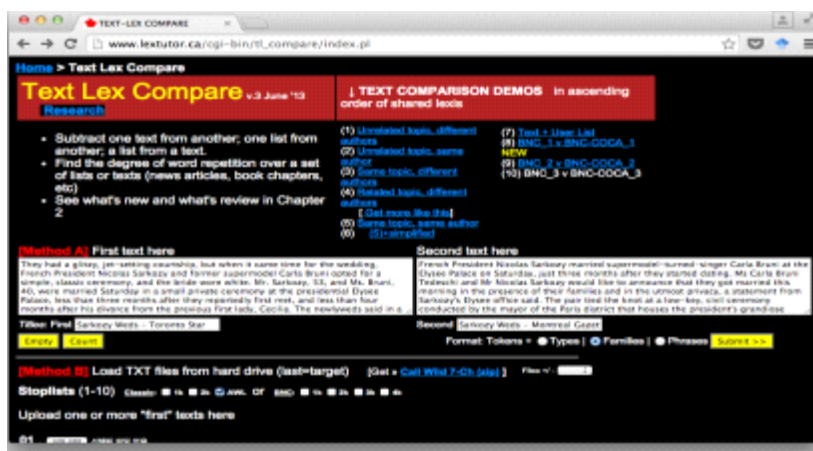
연구자용 서비스 중 “Frequency”는 어휘 빈도 목록을 만드는 기능을 제공한다. 사용자가 자신의 텍스트 파일을 업로드하여 빈도표를 작성할 수도 있다. 결과는 빈도순으로 정렬되어 제공된다. 순위(RANK)와 빈도(FREQ)와 함께 개별 어휘의 점유율(COVERAGE)과 누적 점유율도 출력된다.



<그림 34> Lextutor 사이트의 Frequency 서비스

○ Text Lex Compare

연구자용 서비스 중 “Text Lex Compare”는 두 텍스트에 나타나는 어휘를 비교하는 기능을 제공한다. 두 개의 텍스트를 복사하여 붙여 넣거나 각각의 파일을 업로드한 후 어휘를 비교할 수 있다. 파일을 업로드하여 분석하는 경우에 어느 텍스트에나 일반적으로 나타나는 기능어들을 “Stoplists” 옵션을 이용하여 제외하고 분석할 수 있다.



<그림 35> Lextutor 사이트의 Text Lex Compare 서비스

텍스트 어휘 비교 분석 결과는 <그림 36>과 같은 형식으로 출력된다. 세 개의 칼럼으로 어휘 목록이 출력된다. 왼쪽에는 첫 번째 텍스트에만 출현하는 어휘가 빈도순으로 제시되고 오른쪽에는 두 번째 텍스트에만 출현하는 어휘가 빈도순으로 제시된다. 가운데 칸에는 두 텍스트에 공통으로 나타나는 어휘의 목록이 빈도순으로 출력된다.

The screenshot shows a web browser window titled "TEXT LEX COMPARE output" with the URL "www.lexutor.ca/cgi-bin/tl_compare/new_items.pl?ups=3". The page displays the results of a word frequency comparison between two texts. The first text is "Sarkozy Weds - Toronto Star (366 families)" and the second text is "Sarkozy Weds - Montreal Gazette (306 families)".

FAMILY ANALYSIS
Using the family as unit of comparison means that if cat is in Text 1 and cats in Text 2 then this is considered a repetition of the cat family, i.e. these are equivalent tokens.
Note also that in this routine, "unfamiliar" items revert to classification by type (e.g., repeated proper nouns).

TOKEN Recycling Index: (479 repeated tokens : 658 tokens in new text) = 72.80%
FAMILIES Recycling Index: (149 repeated families : 306 families in new text) = 48.69%
(Token recycling will normally be the most interesting measure of e.g. text comprehensibility, as it is with VP's.)

Unique to first 236 tokens 217 families	Shared 479 tokens 149 families	Unique to second 179 tokens 157 families
001. report 6	001. the 38	001. 3 5
002. love 4	002. and 23	002. last 5
003. before 3	003. be 23	003. start 3
004. courtship 3	004. a 20	004. follow 2
005. young 3	005. he 15	005. get 2
006. apparent 2	006. sarcozy 13	006. just 2
007. chad 2	007. in 11	007. office 2
008. do 2	008. marry 11	008. pair 2
009. during 2	009. number 11	009. press 2
010. girlfriend 2	010. of 10	010. reside 2
011. glitz 2	011. say 10	011. there 2
012. high 2	012. bruns 9	012. two 2
013. italy 2	013. they 9	013. very 2
014. jet 2	014. to 9	014. wood 2
015. jess 2	015. at 8	015. would 2
016. meet 2	016. as 7	016. across 1
017. model 2	017. france 7	017. add 1
018. newsweds 2	018. from 7	
019. news 2	019. have 7	

VP novel items

<그림 36> Lextutor 사이트의 Text Lex Compare 분석 결과

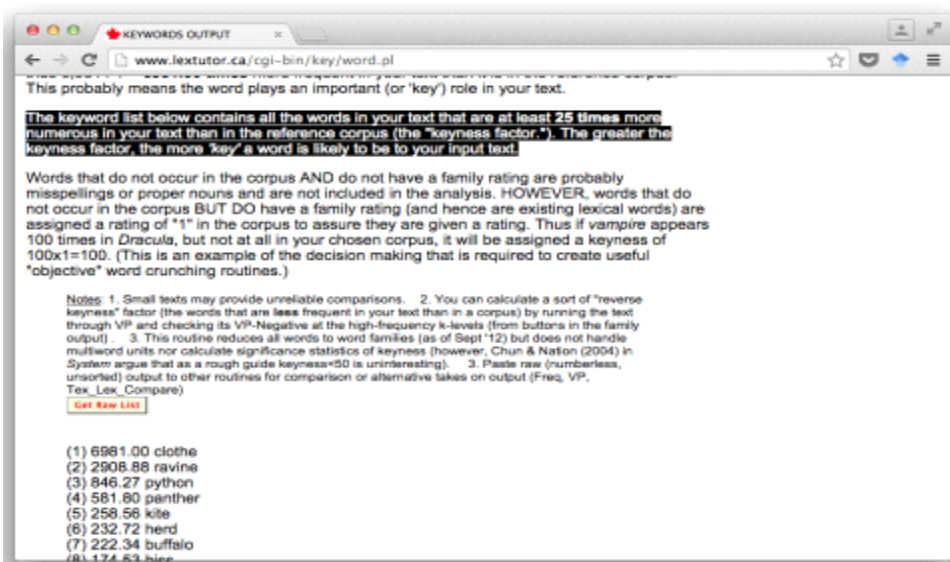
○ KeyWords

연구자용 서비스의 하나인 “KeyWords”는 전문적인 텍스트에서 키워드를 추출해 주는 서비스이다. 해당 텍스트에 나타나는 단어의 빈도를 보편적인 참조 말뭉치에서의 빈도와 비교하여 키워드를 추출하는 기법을 사용하고 있다. 분석을 원하는 텍스트를 화면에 붙여 넣기 하거나 업로드한 후 참조 말뭉치를 지정하고 분석을 진행할 수 있다.



<그림 37> Lextutor 사이트의 KeyWords 서비스

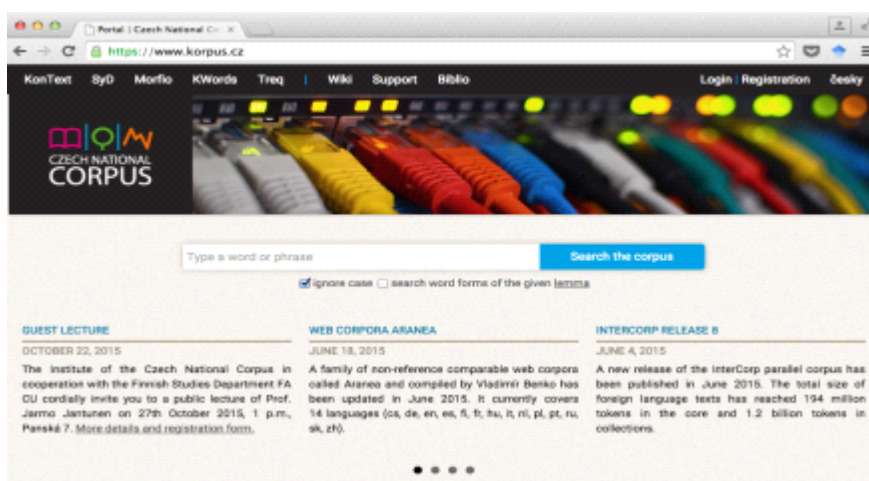
<그림 38>은 예제로 주어진 'Jungle Book' 텍스트를 대상으로 영국 영어, 미국 영어, 문어, 구어가 혼합되어 있는 말뭉치를 기준으로 삼아 키워드를 추출한 결과이다. 기본적으로 출력되는 단어의 목록은 참조 말뭉치에서보다 25배 이상 많이 출현한 단어들의 목록이다. 단어 옆의 숫자가 참조 말뭉치에 비해 몇 배나 많이 출현하고 있는지를 보여준다.



<그림 38> Lextutor 사이트의 KeyWords 서비스의 분석 결과

(3) CZESL (<https://www.korpus.cz/>)

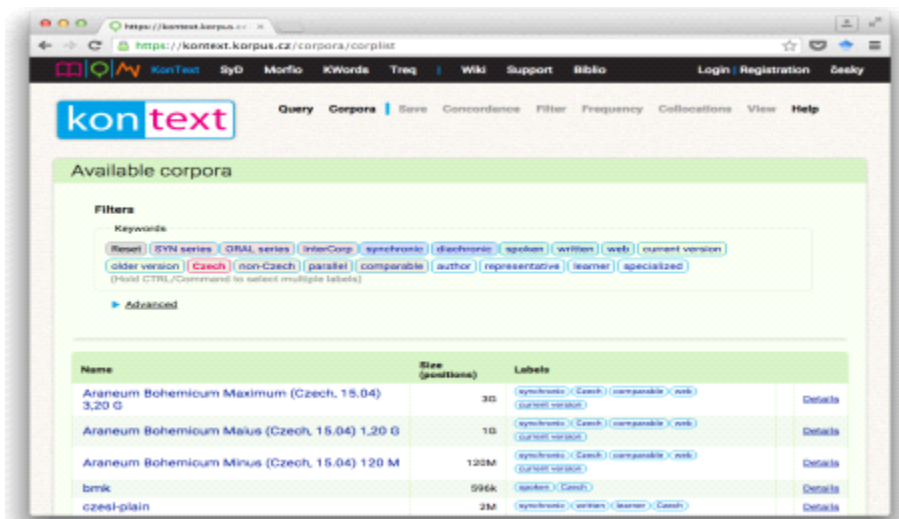
제2언어로서의 체코어(Czech as a Second Language) 교육을 위해 약 230만 단어 규모의 체코어 학습자 말뭉치인 CzeSL-plain이 구축되었으며 체코 국가 말뭉치(Czech National Corpus) 연구소에서 관리되고 있다. 체코 국가 말뭉치는 프라하 카렐 대학교의 체코 국가 말뭉치 연구소에서 진행하는 프로젝트이며 웹상의 온라인 서비스를 제공하고 있다. 공시적 문어 말뭉치, 공시적 구어 말뭉치, 통시적 말뭉치, 외국어 말뭉치, 병렬 말뭉치 등의 검색 서비스를 제공하고 있다. 체코어 국가 말뭉치 온라인 서비스는 외국인도 사용할 수 있도록 영어 인터페이스를 제공하고 있다. 그러나 상세 설명은 대부분 체코어로만 되어 있다.



<그림 39> 체코국가말뭉치 사이트 첫 화면

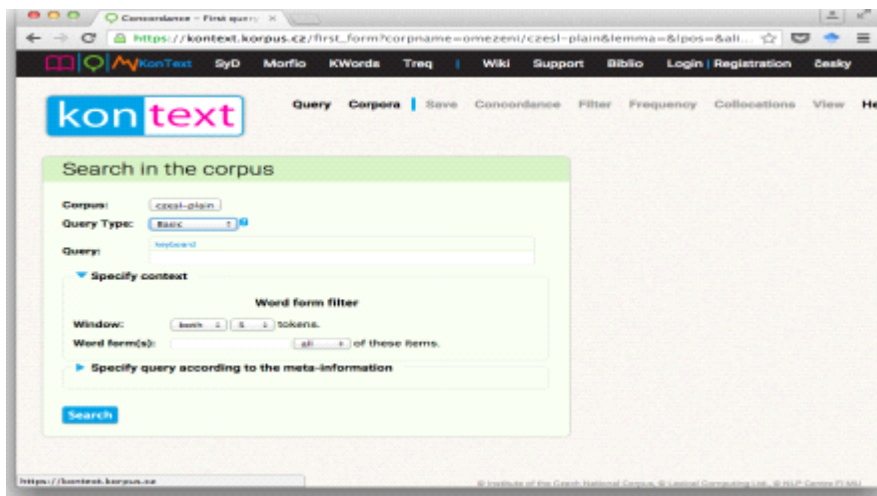
○ KonText

체코국가말뭉치 사이트의 가장 기본적인 기능은 KonText 메뉴를 통해 제공되는 문맥 색인 검색 서비스이다. 매우 다양한 말뭉치를 제공하고 있다. Vladimír Benko가 구축한 Aranea 시리즈는 체코어, 독일어, 영어, 스페인어, 핀란드어, 프랑스어, 헝가리어, 이탈리아어, 네덜란드어, 폴란드어, 포르투갈어, 러시아어 말뭉치를 제공한다. 체코어 말뭉치도 30가지 이상을 제공한다. <그림 40>에서 볼 수 있듯이 각 말뭉치에는 태그가 달려있으며 “Filter” 메뉴에서 태그를 클릭하여 해당 조건에 부합하는 말뭉치를 찾을 수 있다.



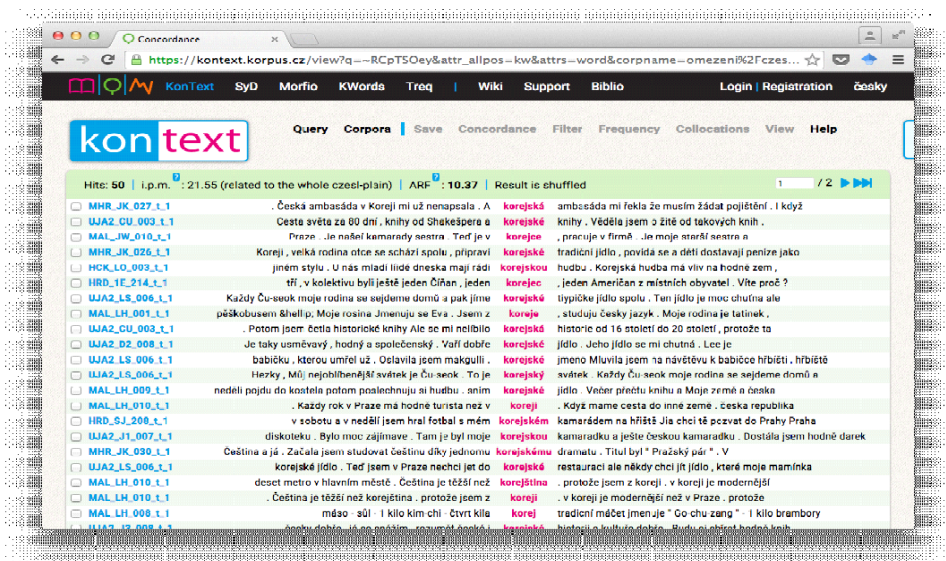
<그림 40> 체코국가말뭉치에서 제공하는 말뭉치 목록

검색 서비스에서는 “Corpus”에서 말뭉치를 지정하고 “Query Type”에서 검색 방법을 Basic, Lemma, Phrase, Word Form 등으로 지정하고 진행할 수 있다. 체코어는 형태론이 발달한 언어이므로 기본형을 이용해 검색하는 것과 어형이 정확하게 일치하는 경우만 검색하는 것에 상당한 차이가 있다. 추가로 문맥을 지정할 수 있다. 검색 범위를 오른쪽, 왼쪽, 양쪽으로 1~15 토큰까지 지정할 수 있다. <그림 41>에서는 체코어 학습자 말뭉치 CzeSL-plain으로 말뭉치를 지정한 것을 보여준다.



<그림 41> 체코국가말뭉치의 KonText 검색 서비스

<그림 42>는 검색 결과를 예시한 것이다. 한국을 지칭하는 단어가 사용된 용례를 검색하여 모두 50개의 용례가 추출되었다. 왼쪽에는 텍스트 고유번호가 출력되고 가운데 붉은색 키워드를 중심으로 왼쪽 문맥과 오른쪽 문맥이 5~8 단어로 출력된다. 각 검색 결과를 클릭하면 아래쪽에 좀 더 긴 문맥이 팝업 창에 나타난다.

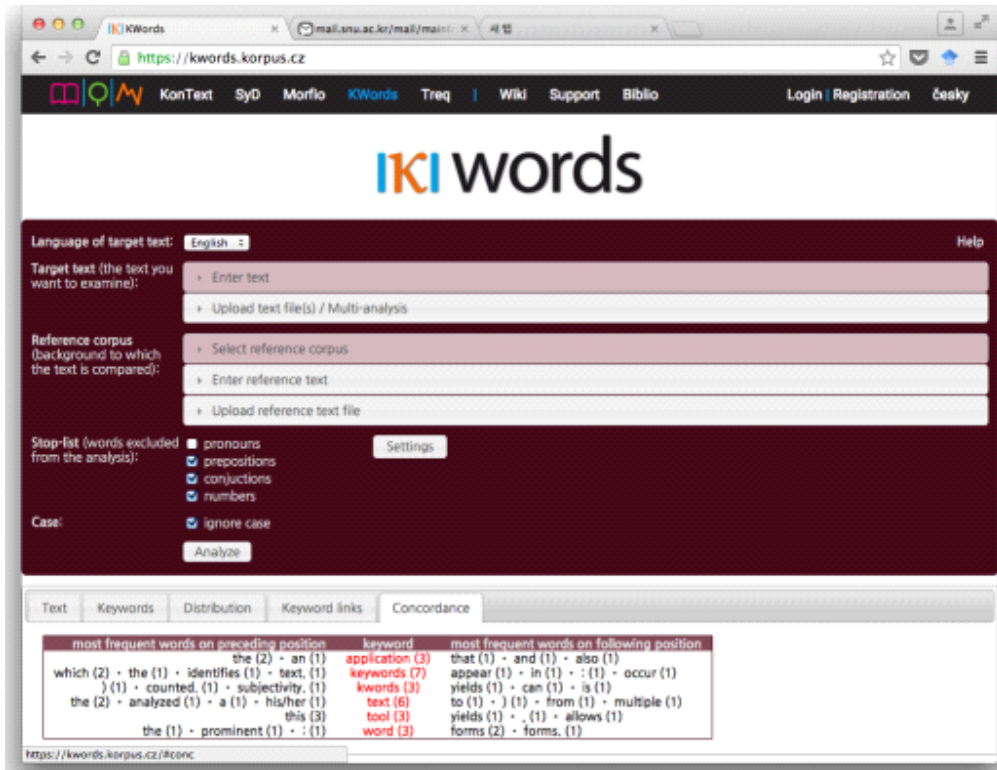


<그림 42> CZESL의 용례 검색 서비스

사이트에 가입하고 로그인을 하면 더 많은 기능을 이용할 수 있다. 검색 결과를 저장할 수 있으며 문맥 색인 결과를 정렬하거나 필터링하는 기능도 제공된다. 추가적인 빈도 분석과 다양한 점수를 이용한 연어 후보 추출도 가능하다.

○ KWords

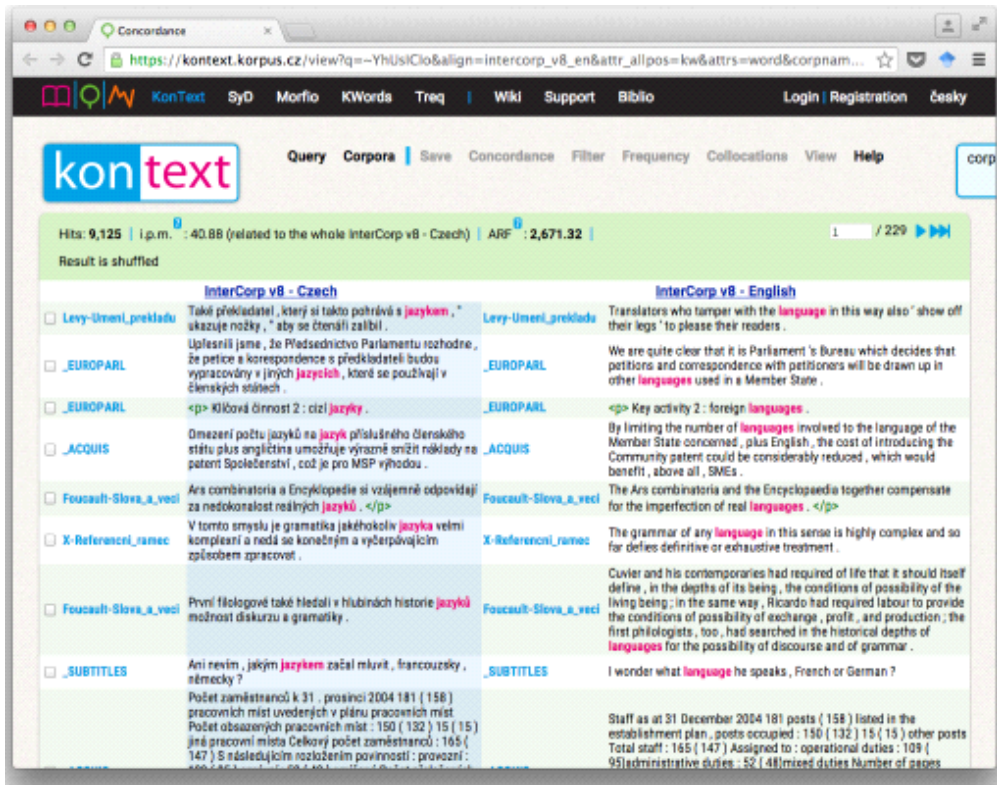
KWords 서비스는 키워드를 추출하는 기능을 제공한다. 분석 대상 텍스트를 올리고 참조 말뭉치를 지정하여 키워드를 추출할 수 있다. 체코어와 영어를 분석할 수 있다. 단순히 키워드를 추출할 뿐만 아니라 키워드가 텍스트에서 어떻게 분포하는지, 키워드들 사이의 관계는 어떠한지도 보여준다. 키워드를 중심으로 문맥에 공기는 어휘들도 보여준다.



<그림 43> 체코국가말뭉치 KWords 분석 결과의 예

○ Treq

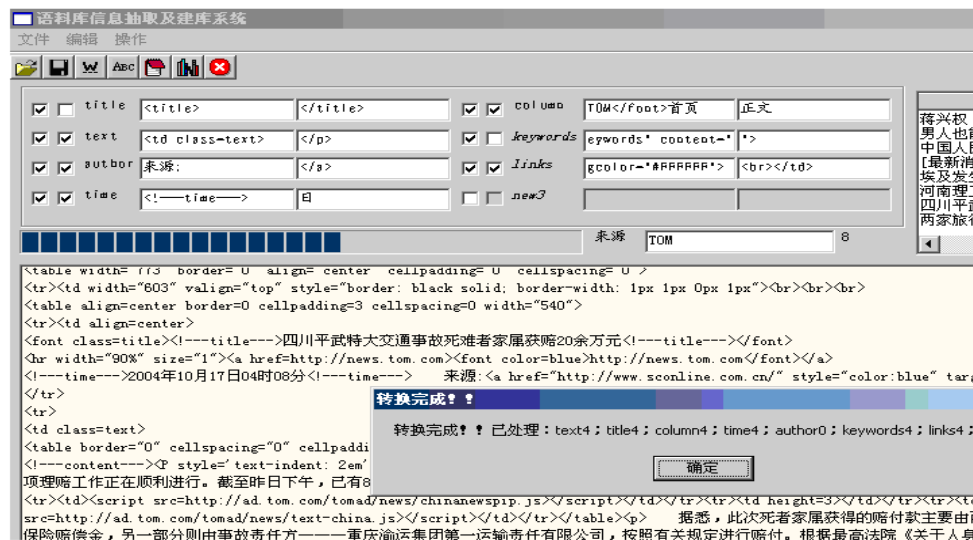
Treq는 “translation equivalent” 데이터베이스를 검색해 주는 서비스이다. 언어를 설정하고 검색어를 지정하여 어떻게 번역되는지 확인할 수 있다. 체코어와 영어를 제1언어로 서비스하고 있으며, 번역 결과인 제2언어는 독일어, 프랑스어, 네덜란드어, 핀란드어, 에스토니아어, 이탈리아어, 일본어 등 수십 개 언어를 서비스하고 있다. 검색 결과를 클릭하면 <그림 44>에서 볼 수 있는 병렬 말뭉치 검색 결과가 제시된다.



<그림 44> 체코국가말뭉치 병렬 말뭉치 검색 결과

(4) 중국 지난 대학 학습자 말뭉치 (<http://www.globalhuayu.com/>)

중국 지난 대학에 의해 관리되는 중국어 학습자 말뭉치로 문어 말뭉치와 구어 말뭉치가 구축되어 있다. 말뭉치 구축 도구 및 중국어 자연 언어 처리 도구를 소개하고 있으며 온라인 검색이 가능하다.



<그림 45> 지난 대학 학습자 말뭉치 사이트의 용례 검색 화면

(5) Romance Phonetic Database (<http://rpd.chass.utoronto.ca/>)

토론토 대학의 로망스어 음성 데이터베이스는 프랑스어, 이탈리아어, 포르투갈어, 루마니아어, 스페인어를 포함하는 로망어의 분절음 및 프로소디 측면의 주석된 음성 파일을 포함하는 온라인 연구와 언어 교육용 도구를 제공한다. 다양한 정보를 제공한다. 세 가지 측면에서 세밀한 정보를 제약하여 검색이 가능하다.

- Language
- Linguistic variables
- Speaker variables

검색 결과는 즉시 들어볼 수 있으며 wave 파일을 다운로드할 수도 있다. SAMPA로 음성 전사가 되어 있다. 발화자의 정보도 팝업 창으로 확인 가능하다.

RPD University of Toronto
Romance Phonetics Database

■ Query Words ■ Advanced query ■

In order to query the database, select values for the following variables; not all variables need to be specified. When finished, click on 'Query Database'.
(Tip: if you don't specify any variable and hit 'Query Database' all the analyses will be retrieved from the database!).

Language

Study

English Language Survey
French Stop-liquid Clusters (Steele 2001 and 2004)
French Word-final Clusters (Steele 1999)
L2a of OL Clusters (Colantoni and Steele 2004-05)
L3 Phonology (Steele 2004-08)
L3 Stress (Sunkara 2008)
Linguistic Atlas of Argentina (Colantoni 1994-97)
OL Clusters (Colantoni and Steele 2003-04)
Romance Language Survey

Target Language **Language Type**

English Native
French Non-native
Italian
Portuguese
Romanian
Spanish

Linguistic variables

Target Phoneme **Target Grapheme**

Highlight: None Highlight: None

#	Phoneme	#	Grapheme
0		0	
1		1	
2		2	
3		3	
4		4	
5		5	
6		6	
7		7	
8		8	
9		9	
10		10	
11		11	
12		12	
13		13	
14		14	
15		15	
16		16	
17		17	
18		18	
19		19	
20		20	
21		21	
22		22	
23		23	
24		24	
25		25	
26		26	
27		27	
28		28	
29		29	
30		30	
31		31	
32		32	
33		33	
34		34	
35		35	
36		36	
37		37	
38		38	
39		39	
40		40	
41		41	
42		42	
43		43	
44		44	
45		45	
46		46	
47		47	
48		48	
49		49	
50		50	
51		51	
52		52	
53		53	
54		54	
55		55	
56		56	
57		57	
58		58	
59		59	
60		60	
61		61	
62		62	
63		63	
64		64	
65		65	
66		66	
67		67	
68		68	
69		69	
70		70	
71		71	
72		72	
73		73	
74		74	
75		75	
76		76	
77		77	
78		78	
79		79	
80		80	
81		81	
82		82	
83		83	
84		84	
85		85	
86		86	
87		87	
88		88	
89		89	
90		90	
91		91	
92		92	
93		93	
94		94	
95		95	
96		96	
97		97	
98		98	
99		99	

Romance Phonetics Database

rpdb.chass.utoronto.ca/public/query.php

Ante Pre-tonic
Pre-tonic
Tonic
Post-tonic
Post Post-tonic

Initial
Medial
Final

Speaker variables

Sex **Age** (-number-) **Education** **First Language Variety** **Multiple L1s only**

Female
Male

min age max age

None
Elementary
Secondary
University

Afrikaans
Albanian
Arabic
Arabic
Armenian
Basque
Belarusian

☐

L2 **L2 Proficiency** **Multiple L2s only**

Afrikaans
Albanian
Arabic
Arabic
Armenian
Basque
Belarusian

None
Beginner
Intermediate
Advanced
Near-native

☐

Type of Acquisition **Age of Onset of Acquisition** (-number-)

Instructed
Naturalistic
Both

min age max age

L2 Immersion Place **L2 Immersion Length (months)** (-number-)

Argentina
Belgium
Belize
Bolivia
Brazil
Canada
Chile

min length max length

Birth Country **Birth Province** **Birth City**

Afghanistan
Albania
Algeria
Argentina
Armenia
Austria
Azerbaijan

Buenos Aires
Ciudad de Buenos Aires
Córdoba
Corrientes
Entre Ríos
Misiones
Quebec

9 de Julio
Albarradón
Alvear
Angaco
Apóstoles
Bella Vista
Berrón de Astrada

<그림 46> Romance Phonetic Database의 검색 화면

2.3. 말뭉치 웹 서비스 기능 및 요구 분석

지금까지 살펴본 국내외 학습자 말뭉치 웹서비스들을 기반으로, 한국어 학습자 말뭉치에서 제공될 필요가 있는 주요 서비스 기능을 정리하였다. 주요 서비스는 말뭉치 표본과 메타데이터를 이용한 파일 수준의 탐색 서비스, 말뭉치 표본의 내·외적 자료 현황을 제공하는 통계 중심 서비스, 말뭉치의 언어학적 자료를 제공하는 용례 검색 서비스, 공기 관계 분석 서비스 등으로 구분하였다.

2.3.1. 말뭉치 탐색 서비스

- 개별 텍스트 단위를 탐색하여 특정 조건에 맞는 자료의 목록을 확보하고 내용을 읽어 볼 수 있는 서비스.
- 각 텍스트의 메타정보 또는 태그를 기준으로 탐색할 수 있는 서비스.
- 텍스트 작성자 또는 음성 전사 자료의 경우에는 발화자에 해당하는 학습자에 대한 다양한 정보에 따른 탐색 기능이 필요. 성별, 나이, 모어, 숙달도, 국외/국내 학습자, 기관, 학습 목적 등.
- 텍스트의 유형과 속성에 따른 탐색 기능이 필요. 작문, 말하기, 시험, 백일장, 발표, 대화 등.
- 음성 전사 자료의 경우 원 음성 파일을 들어볼 수 있는 기능 필요.
- 탐색 결과를 파일로 내려받을 수 있는 기능 필요. 개별 텍스트별로 plain text, rich text format, XML, JSON 등 적절한 형식을 선택하여 다운로드하는 기능.

2.3.2. 어휘 및 오류 분포 탐색 서비스

- 키워드를 입력하여 그에 일치하는 텍스트를 탐색할 수 있는 서비스.
- 학습자 속성 또는 텍스트 유형에 따라 말뭉치를 크게 몇 개의 집단으로 나눌 수 있음. 어떻게 나눌지를 서비스 사용자가 선택할 수 있음. 예를 들어, ‘모어’에 체크 표시를 하면 학습자의 모어에 따라 말뭉치가 하위 그룹으로 나뉨.
- 분석 대상은 1차적으로는 어휘 또는 구문이 될 수 있으며 오류 주석이 되어 있는 자료에 대해서는 오류 유형을 분석 대상으로 할 수 있음.
- 사용자가 지정한 분석 대상이 말뭉치 하위 그룹별로 출현하는 빈도를 계산하여 보여줌. 예를 들어, ‘때문에’를 검색어로 지정하면 모어에 따른 ‘때문에’의 사용 빈도를 보여줌. 분석 대상을 ‘조사 오류’로 지정하면 모어에 따라 조사

오류의 빈도를 보여줌.

- 절대 빈도와 함께 상대 빈도를 보여줄 필요 있음.
- 막대그래프, 원그래프 등의 시각화를 할 수 있음.
- 검색어 또는 분석 대상을 포함하고 있는 텍스트의 목록이 제시되고 해당 텍스트와 생산자에 관한 상세한 내용을 볼 수 있음.

2.3.3. 용례 검색 서비스

- 어휘, 어형, 구문 등을 검색어로 지정하여 문맥 색인(concordance)을 생성하는 서비스.
- 아무개(wildcard) 문자 또는 정규 표현식(regular expression)을 이용한 고급 검색을 지원할 필요 있음.
- 한글의 특성상 자소 검색을 지원할 필요 있음.
- 한국어가 형태론적으로 복잡한 언어라는 특성을 감안하여 기본형 검색과 어형 검색을 구분하여 서비스할 필요 있음. 형태소 분석이 이루어진 말뭉치의 경우에 가능. 예를 들어, 기본형 검색으로 “가다”를 검색하면 “간”, “갈”, “감”, “갔다” 등이 모두 검색이 됨. 어형 검색으로 “가다”를 검색하면 정확히 “가다”인 경우만 검색됨.
- 문맥을 제한하는 기능 필요. 키워드를 중심으로 좌우 일정 길이 이내에 특정 구문이 나타나는 경우로 검색 범위를 제한할 수 있음. 예를 들어, “주다”를 키워드로 하여 “에게/한테”와 “을/를” 목적어가 동시에 나타나는 경우만 검색할 수 있음.
- 검색 결과를 정렬하는 기능 필요. 키워드 중심의 정렬과 오른쪽 문맥과 왼쪽 문맥을 기준으로 정렬이 필요. 문맥으로 정렬할 때에는 키워드로부터 떨어진 거리에 따라 정렬 기준에 포함할 수 있음. 한국어의 특성을 고려하여 한 어절의 뒤쪽 음소부터 기준으로 정렬하는 역순 정렬 기능을 제공할 필요 있음.
- 검색 결과를 다양한 형식으로 내보낼 수 있는 기능 필요.

2.3.4. 공기 관계 분석 서비스

- 관용구, 속어, 연어, 정형표현 등으로 불리는 서로 밀접하게 함께 출현하는 언어 요소들을 분석하는 서비스.
- 문맥 색인 검색 서비스를 보조할 수 있는 분석 서비스임.
- 특정 키워드를 중심으로 좌우에 인접한 어휘들을 출현 빈도에 따라 제시함.

- 특정 키워드를 중심으로 좌우 일정 거리 이내에 사용된 어휘들을 출현 빈도에 따라 제시함.
- 여러 가지 언어 점수에 따른 순위를 제시함.
- N-Gram 분석으로 확장할 수 있음.

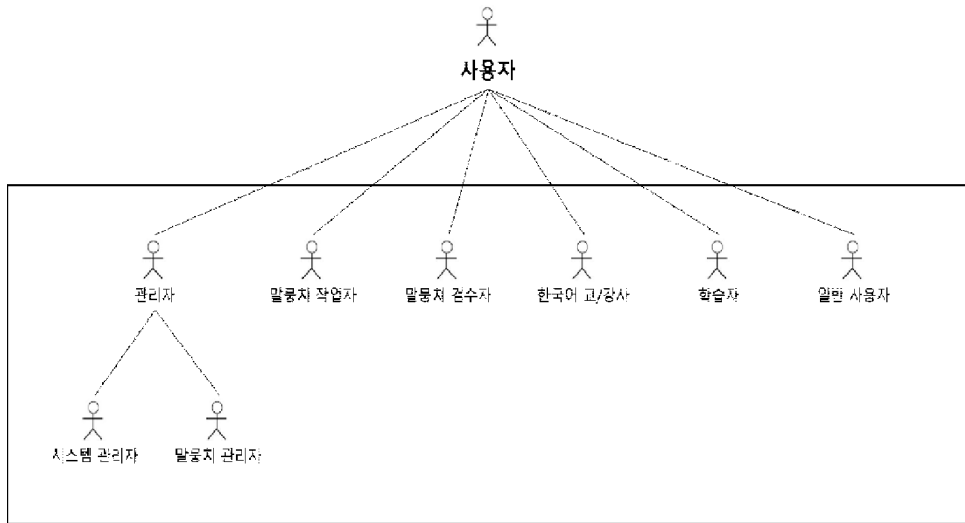
2.3.5. 기타 응용 서비스

- 학습자, 연구자, 교사를 위한 응용 서비스들을 개발할 수 있음.
- 학습자와 교사를 위한 서비스에서는 학습자 생산 말뭉치와 함께 사전 및 참조 말뭉치가 제공될 필요가 있음.
- 학습자는 참조 말뭉치의 문맥 색인과 사전이 연계되어 제공되는 시스템을 통해 어휘 및 문법을 학습하거나 작문에 도움을 받을 수 있음.
- 학습자는 참조 말뭉치에서 추출된 용례를 이용해 자동으로 생성된 빈칸 채우기 문제 등을 통해 온라인 자가 테스트를 할 수 있음.
- 학습자 말뭉치에 나타나는 전형적인 오류를 담고 있는 문장과 그에 대응하는 용례를 참조 말뭉치에서 찾은 결과를 비교하여 보여줌으로써 문법 학습이 가능함.
- 교사에게는 참조 말뭉치에서 용례를 추출하고 시험 문제를 자동으로 생성하는 기능을 제공할 수 있음.
- 연구자들에게는 다양한 추가적인 분석 기능을 제공할 수 있음.

3. 한국어 학습자 말뭉치 통합 관리 시스템 설계

3.1. 유스케이스 분석

(1) 사용자 유형 분석



<그림 47> 사용자 유형 분석

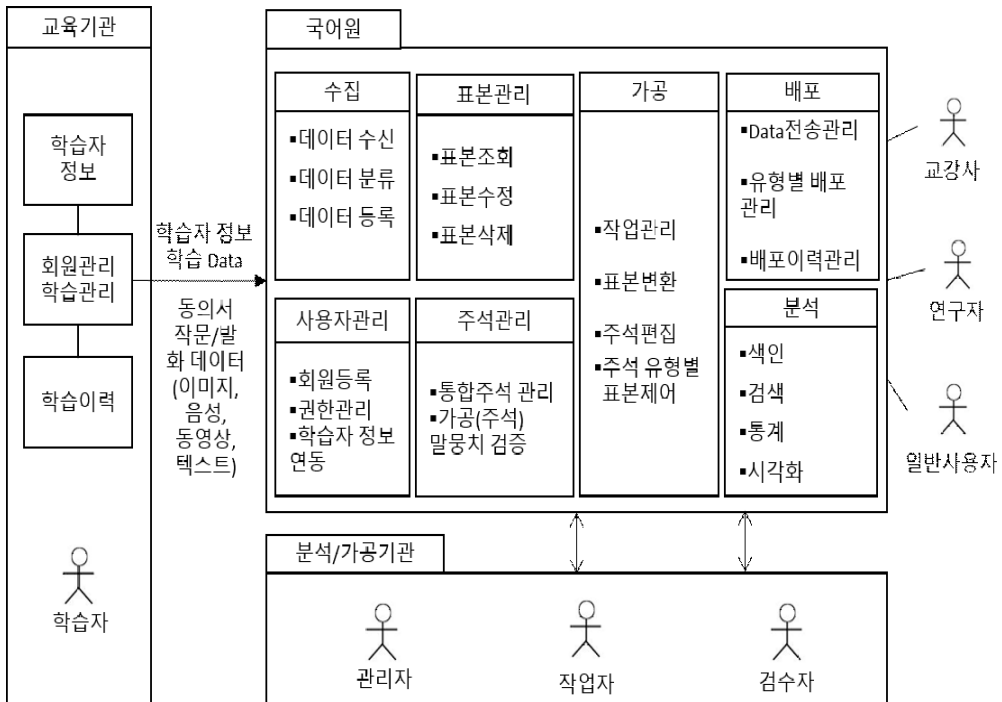
(2) 사용자 정의

유형	설명
시스템 관리자	한국어 학습자 말뭉치 통합 관리 시스템의 하드웨어(H/W), 소프트웨어(S/W)를 관리, 제어하는 권한을 가진 사용자
말뭉치 관리자	한국어 학습자 말뭉치 통합 관리 시스템을 이용한 학습자 말뭉치의 수집, 가공 작업에 필요한 사용 권한을 획득한 관리자
말뭉치 작업자	한국어 학습자 말뭉치 통합 관리 시스템에서 등록된 표본에 대한 표본 수정 및 가공 작업을 처리하는 사용자

말뭉치 검수자	한국어 학습자 말뭉치 통합 관리 시스템에서 표본에 대한 주석 및 오류에 대한 가공 작업을 검수하는 사용자
한국어 교/강사	- 교육 기관 소속의 한국어학습 강사 및 교사 - 한국어 학습자 말뭉치 통합 관리 시스템에서 표본을 수집하는 권한을 가진 사용자
학습자	- 교육 기관 소속의 한국어 학습을 하는 사용자 - 자신의 한국어 말뭉치 표본을 제공하는 사용자
일반 사용자	한국어 학습자 말뭉치 통합 관리 시스템을 활용하고자 하는 사용자

<표 10> 사용자 정의

3.2. 한국어 학습자 말뭉치 통합 관리 시스템 아키텍처



<그림 48> 한국어 학습자 말뭉치 통합 관리 시스템 개요도

한국어 학습자 말뭉치 통합 관리 시스템의 외적 요인은 주관 기관이자 시스템의 서비스 제공자인 국립국어원과 학습자 말뭉치의 원시 데이터를 공급하는 한국어 교육기관, 학습자 말뭉치를 설계, 가공, 분석하는 연구기관(한국어 학습자 말뭉치 구축팀)으로 구분된다.

한국어 학습자 말뭉치 통합 관리 시스템의 사용자는 유스케이스와 요구 분석을 통해 식별하였다. 한국어 학습자 말뭉치 통합 관리 시스템의 사용자는 외적 요인인 기관의 역할과 서비스에 따라 요구 기능과 사용 시나리오가 달라진다. 사용자별 기능 분석은 상세 유스케이스를 통해 도출하였으며, 여기서는 외적 요인(기관), 역할, 관련 서비스의 관점에 대해서만 도식하였다.

한국어 학습자 말뭉치 통합 관리 시스템의 구성 요소는 수집부, 관리부, 가공부, 분석부, 배포부로 구분된다. 이 중에서 관리부는 사용자관리, 표본관리, 주석관리로 세분된다. 시스템 특성상 관리 부분의 영향 요인이 크므로, 향후 상세 설계 과정을 통해 모듈의 응집성과 독립성에 대한 균형 유지에 주의하도록 한다.

여기서 도식되지 않은 통합 주석 체계의 구조, 자료형의 검증, 데이터의 교환 등은 상세 설계 및 III. 한국어 학습자 말뭉치 구축 지원 도구 개발의 데이터 스키마, 활동 다이어그램, 클래스 다이어그램 등을 통해 제시한다.

3.2.1. 소프트웨어 아키텍처



<그림 49> 한국어 학습자 말뭉치 통합 관리 시스템 S/W 아키텍처

한국어 학습자 말뭉치 통합 관리 시스템은 말뭉치 관리 시스템의 요구 기능과 함께 학습자 말뭉치의 특수성을 수용할 수 있어야 한다. 학습자 말뭉치는 원시 말뭉치 수준에서 학습자에 의한 다양한 오류를 포함하고 있으며, 문어와 구어가 함께 처리될 수 있어야 한다. 또한 가공 단계인 주석 과정에서는 형태소 분석, 구문 분석과 같은 일관된 문법 단위의 분석뿐만 아니라, 문법 단위가 혼재된 주석이 요구되기도 한다. 그러므로 한국어 학습자 말뭉치 구축 시스템은 각각의 소프트웨어 계층에서 이러한 요구 사항을 수용할 수 있는 기능적 집합과 데이터 통신 규약을 확립해야 한다.

본 연구에서는 한국어 학습자 말뭉치 통합 관리 시스템의 소프트웨어 계층을 저장, 데이터 처리, 데이터 관리, 데이터 조작, 데이터 출력의 개념으로 구분하였다.

저장은 표본, 주석, 메타데이터를 저장 매체로 입력하고 출력하는 기능을 담당한다. DBMS를 중심으로 활용하며, 필요에 따라서는 운영 체제에 의한 파일 시스템의 이용한다.

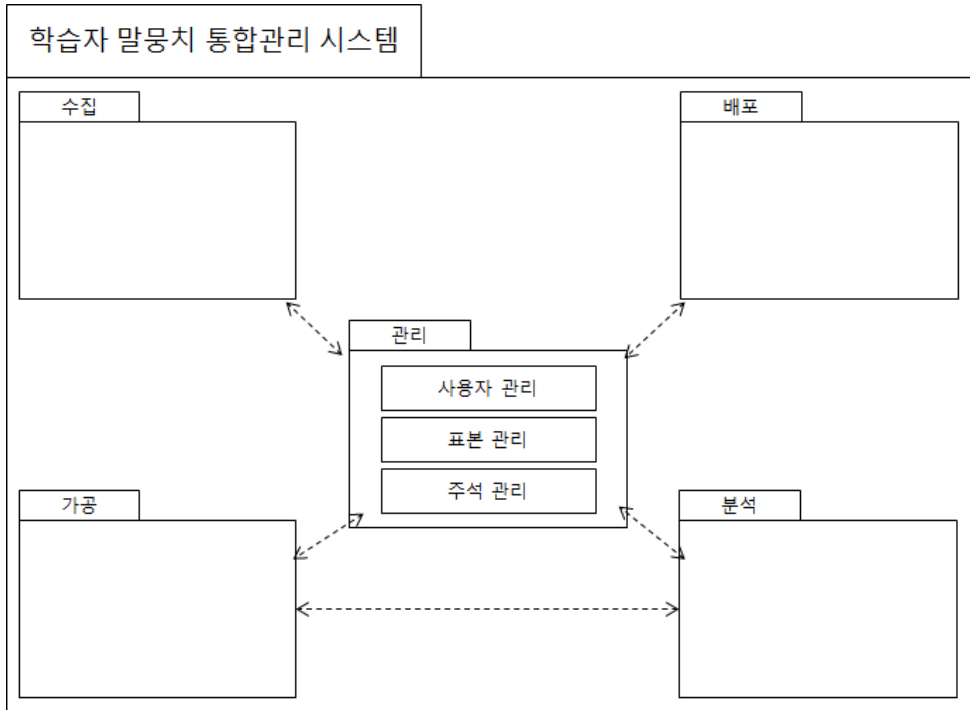
데이터 처리는 비정형 데이터를 통합적으로 관리하는 역할을 한다. 텍스트, 음성과 같은 비정형 데이터와 이에 부가되는 주석 및 메타데이터를 일관성 있게 처리한다. Apache UIMA의 CAS(Common Analysis Structure)와 Type System Model에서처럼 Stand-off 방식의 주석 처리를 전제로 설계하였다. 처리된 데이터는 주로 DBMS에 의해 최종 변경/저장된다.

데이터 관리는 주석, 메타데이터, 사용자 요청에 따라 각각의 데이터의 이동, 변경, 삭제, 수정을 담당한다. 데이터 관리는 사용자에게는 주석별, 매체별, 기능별로 필요한 관점의 정보를 제공하지만, 데이터의 처리는 비정형 데이터 통합 관리 체계에 위임한다.

데이터 조작과 데이터 출력은 모두 사용자에게 제공되는 데이터 표현 개념이다. 데이터 조작은 사용자가 데이터의 갱신을 목적으로 한 처리를 지원하는 UI 및 기능을 제공하고, 데이터 출력은 사용자가 데이터에 영향을 미칠 수 없는 단순 보기(View), 내려받기 기능 등을 제공한다.

3.2.2. 한국어 학습자 말뭉치 통합 관리 시스템 구성

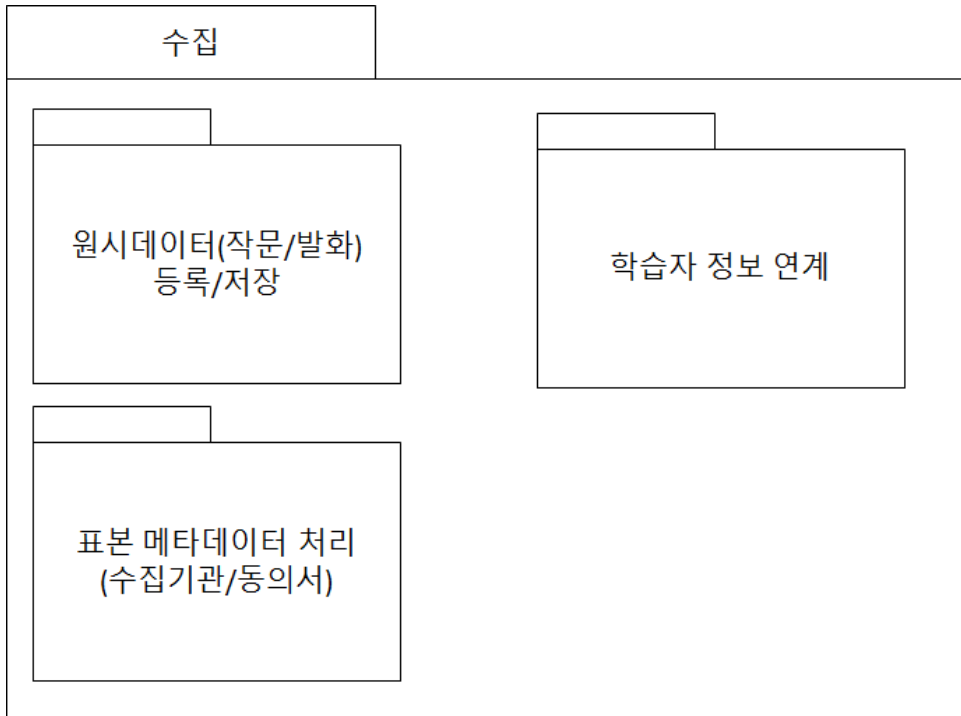
(1) 한국어 학습자 말뭉치 통합 관리 시스템 모듈 구성



말뭉치는 가독성에 있어서 2가지의 상충된 요구 사항을 충족시켜야 한다. 기계 가독(Machine Readable)하면서도, 인간 가독(Human Readable)해야 한다는 점이다. 한국어 학습자 말뭉치 통합 관리 시스템에서도 말뭉치 가독성의 2가지 요구 사항을 융합하는 것이 가장 중요한 설계 조건이 된다. 인간 가독(Human Readable)의 조건은 가시적인 인터페이스에 의해 사용자가 데이터를 사용하는 시점 즉, 시스템 내부에서의 데이터 처리가 끝나고 출력되는 시점이다. 결국 사용자는 시스템에 저장된 데이터를 보는 것이 아니라, 사용자 인터페이스에 노출된 메타 데이터와 시스템에 의해 처리된 결과물을 보는 것이다.

그러므로 한국어 학습자 말뭉치 통합 관리 시스템은 핵심 데이터의 내부적 처리를 담당하는 관리 모듈을 중심으로 사용자와 직접 소통하는 수집, 가공, 분석, 배포 모듈로 구성된다. 사용자에게 제공되는 인간 가독(Human Readable)한 출력 결과와 내부 시스템이 관리하는 기계 가독(Machine Readable) 데이터가 일관된 연결성을 유지하도록 하는 것이 핵심이다.

(2) 수집부 모듈 구성

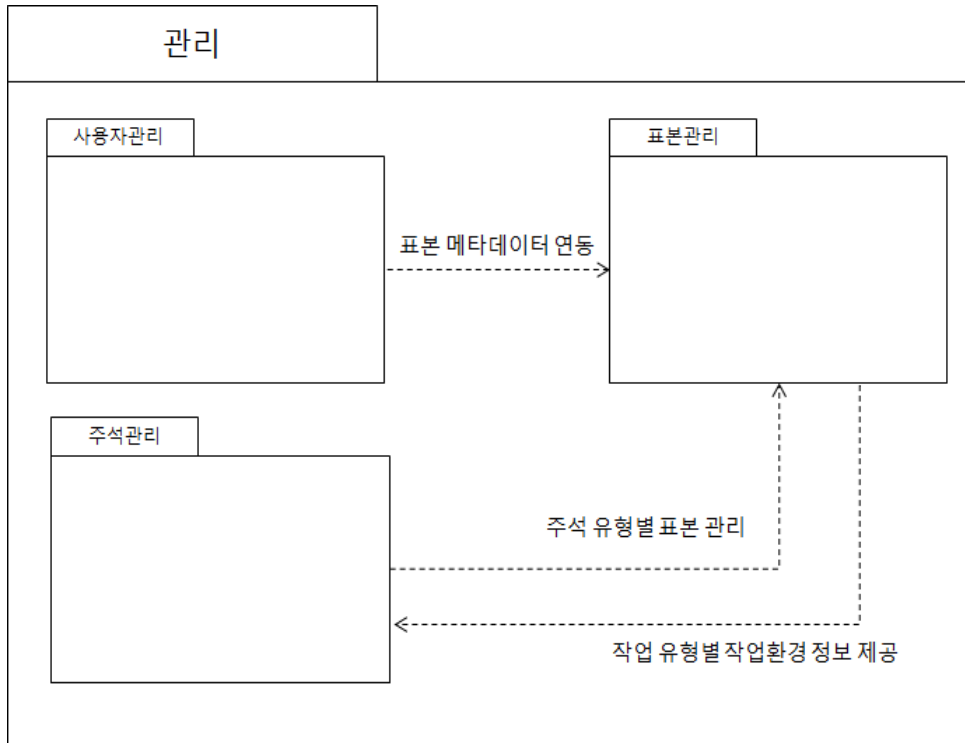


수집부는 각 교육기관의 학습자, 교/강사 또는 작업자로부터 학습자의 발화/작문과 관련한 데이터들을 획득하는 역할을 한다. 현재는 발화/작문의 원시 데이터, 학습자의 학습 관련 정보, 수집 기관(교육 기관) 정보, 학습자의 동의 증빙 등이 요구된다. 수집 대상 정보 및 데이터의 명세는 수집 환경, 정책에 따라 변동될 수 있음을 고려한다.

수집부 설계에 있어서 핵심적인 요구 사항은 수집되는 발화/작문의 원시 데이터와 메타데이터(학습자, 수집 기관 정보 등)의 연관 관계를 확고히 하는 것이다. 이러한 부분은 수집 데이터를 저장하는 관리 모듈뿐만 아니라, 수집 모듈의 UI에도 반영되어야 한다.

본 연구에서 등록 파일 선택 시 동의서 및 학습자 정보 유무를 검증하는 것도 원시 데이터와 메타 데이터의 연동을 명확히 하기 위해 사용자에게 제공하는 도움 장치이다.

(3) 관리부 모듈 구성

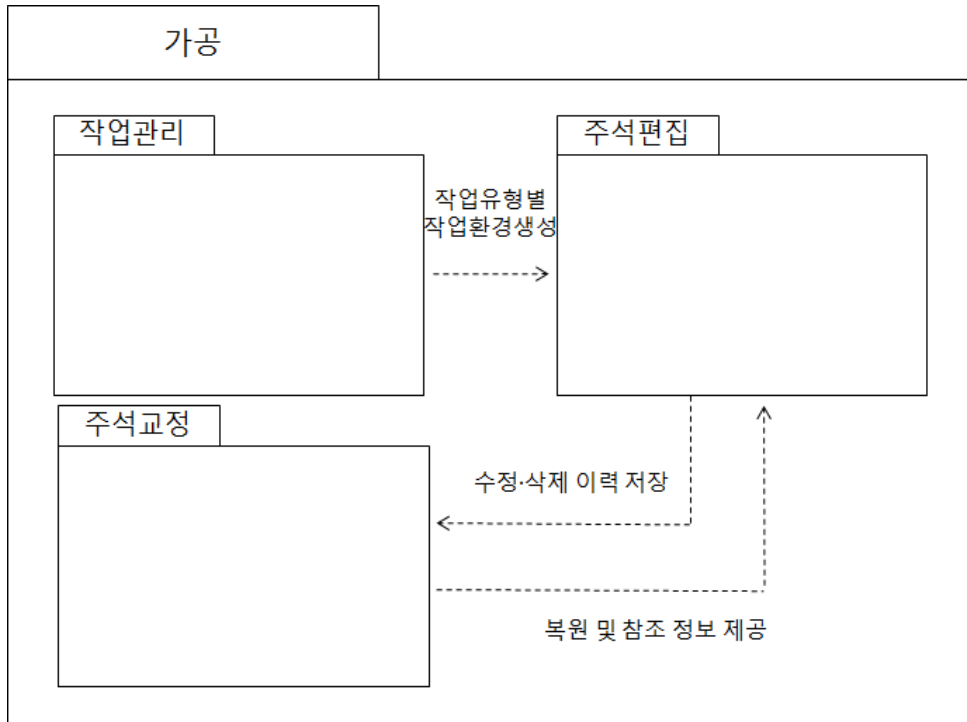


관리 모듈의 주요 관리 대상은 사용자, 표본 파일, 주석이다. 그러나 앞서 살펴본 바와 같이 관리 모듈은 한국어 학습자 말뭉치 통합 관리 시스템의 모든 모듈과 상관 관계를 갖는다. 예를 들어 가공 모듈의 작업 관리는 실제로 사용자와 표본 관리에 의해 이루어진다. 또한 배포 모듈의 배포 기능 역시 관리 모듈을 통해 수행된다.

무엇보다 관리 모듈은 한국어 학습자 말뭉치 통합 관리 시스템이 사용하는 거의 대부분의 데이터를 관리한다. 그러므로 관리 모듈은 앞서 살펴본 바와 같이 관리 대상 데이터를 기계 가독 데이터와 인간 가독 데이터로 구분하고 연관 관계를 처리할 수 있도록 중간적 데이터와 함께 보관해야 한다.

관리 모듈의 핵심 기능은 각 데이터들 간의 관계와 제약에 따라 일관성과 무결성, 동기화를 유지할 수 있도록 제어하는 것이다.

(4) 가공부 모듈 구성

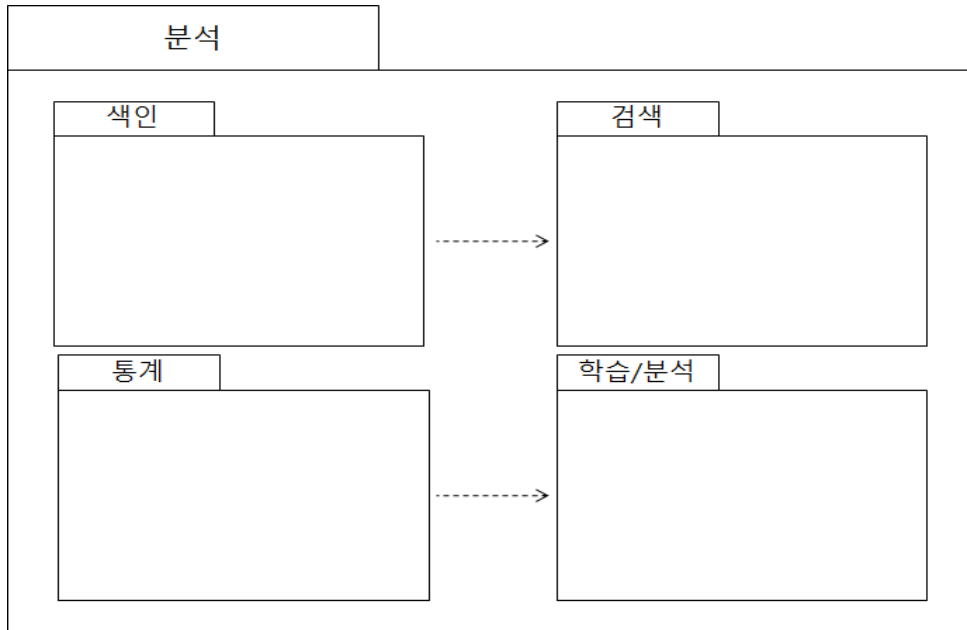


가공부는 작업 관리, 주석 편집, 주석 교정으로 구성되지만, 실질적인 역할은 주석 유형별로 적절한 작업 환경을 제공하고, 그 결과를 관리 모듈에 저장하는 것이다.

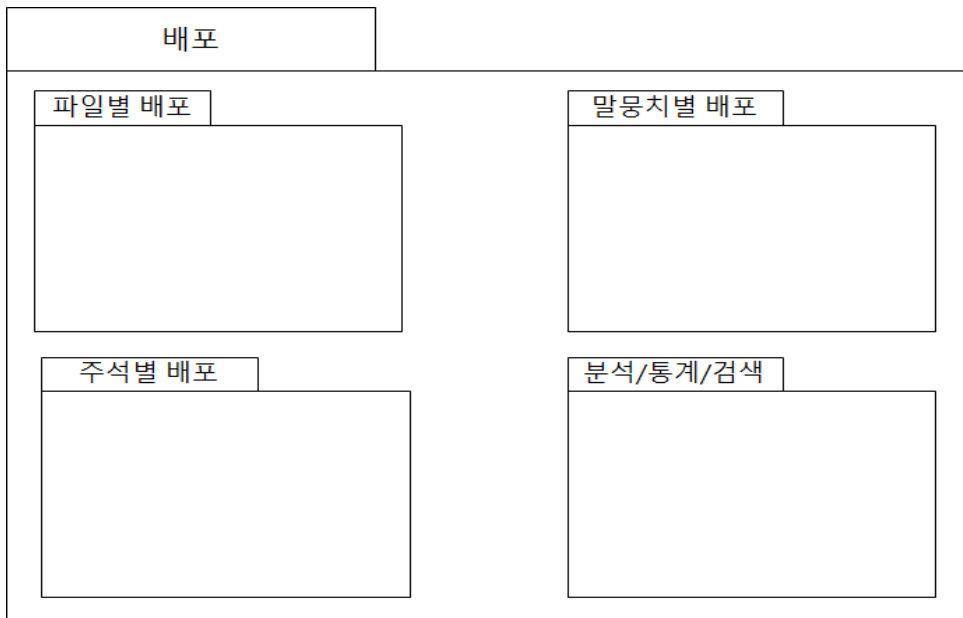
주석 편집은 작업 유형별로 주석 작업자에게 편리한 작업 절차와 적절한 관련 정보를 제공하는 사용자 인터페이스(UI : User Interface)에 있다. 또한 제공된 사용자 인터페이스가 사용자에게 의한 임의의 데이터 입력을 방지하도록 해야 한다. 주석 작업자는 오직 주석이 부가될 대상과 그 값의 적절성에만 집중할 수 있도록 한다.

주석 교정은 주석 작업자 중 검수자나 사용자 신고에 의해 이루어진다. 주석 교정은 이미 주석된 값에 대한 수정/삭제 기능이 중심이 되지만, 숙고된 교정 결과가 수행될 수 있도록 수정/삭제 이력, 복원 기능, 참조 정보 등을 제공한다.

(5) 분석부 모듈 구성



(6) 배포부 모듈 구성



3.3. 주요 기능 목록

기능 명세서				
프로젝트명		한국어 학습자 말뭉치 통합 관리 시스템		
ID			요구 사항	
1. 수집	1.1 표본 수집	1.1.1 파일등록		사용자는 말뭉치 표본 파일을 등록한다.
		1.1.2 학습자 검색		사용자는 학습자 정보를 검색한다.
		1.1.3 학습자정보 등록		사용자는 학습자정보를 등록한다.
		1.1.4 동의서등록		사용자는 학습자의 동의서를 등록한다.
		1.1.5 파일정보 등록		사용자는 파일정보를 등록한다.
		1.1.6 파일명생성		사용자는 파일명을 자동생성 한다.
		1.1.7 표본저장		사용자는 생성된 파일명으로 표본을 저장한다.
		1.1.8 수집현황 조회		사용자는 표본에 대한 수집현황을 조회한다.
2. 관리	2.1 사용자 관리	2.1.1 로그인	2.1.1.1	사용자는 웹페이지 로그인을 한다.
			2.1.1.2	사용자는 웹페이지 로그아웃을 한다.
		2.1.2 회원가입		사용자는 웹페이지 회원가입을 한다.
		2.1.3 변경	2.1.3.1	사용자는 개인정보를 변경한다.
			2.1.3.2	사용자는 회원의 권한을 변경한다.
		2.1.4 조회	2.1.4.1	사용자는 개인정보 조회를 한다.
			2.1.4.2	사용자는 회원 정보를 조회한다.
		2.1.5 삭제	2.1.5.1	사용자는 회원탈퇴를 한다.
			2.1.5.2	사용자는 회원정보를 삭제한다.
	2.2 표본 관리	2.2.1 표본조회	2.2.1.1	사용자는 표본을 주석유형별로 조회한다.
			2.2.1.2	사용자는 표본을 파일 정보로 조회한다.
			2.2.1.3	사용자는 표본을 열람한다.
			2.2.1.4	사용자는 표본의 가공율 통계를 조회한다.
		2.2.2 표본수정	2.2.2.1	사용자는 말뭉치 표본 파일을 수정(변경)한다.
			2.2.2.2	사용자는 파일 정보를 수정(변경)한다.
			2.2.2.3	사용자는 동의서 파일을 수정(변경)한다.
		2.2.3 표본삭제		사용자는 원본 표본 파일을 삭제한다.
	2.3 주석 관리	2.3.1 주석등록	2.3.1.1	사용자는 사용할 주석의 체계를 등록한다.
			2.3.1.2	사용자는 등록된 주석 체계의 지침을 등록한다.
		2.3.2 주석수정		사용자는 주석 메뉴 시스템의 주석을 수정한다.
		2.3.3 주석삭제		사용자는 주석 메뉴 시스템의 주석을 삭제한다.

		2.3.4 주석검증		시스템은 주석 지침에 따라 주석 말뭉치를 검증한다.
3. 가공	3.1 표본 가공	3.1.1 작업관리	3.1.1.1	사용자는 가공할 표본 목록을 조회 한다.
			3.1.1.2	사용자는 표본목록의 작업 유형에 따라 작업한다.
		3.1.2 주석부가	3.1.2.1	사용자는 표본을 문장 단위로 편집한다.
			3.1.2.2	사용자는 표본의 어절을 교정한다.
			3.1.2.3	사용자는 주석을 부가한다.
			3.1.2.4	사용자는 주석을 다층위로 부가한다.
			3.1.2.5	사용자는 부가된 주석을 삭제한다.
			3.1.2.6	사용자는 주석을 수정한다.
			3.1.2.7	사용자는 변경된 주석 상태를 확인한다.
		3.1.3 주석교정	3.1.3.1	사용자는 형태주석을 수정한다.
			3.1.3.2	사용자는 오류주석을 수정한다.
			3.1.3.3	사용자는 원시말뭉치 입력/전사를 수정한다.
			3.1.3.4	사용자는 구문주석을 수정한다.
			3.1.3.5	사용자는 의미주석을 수정한다.
			3.1.3.6	사용자는 표본에 부가된 주석의 오류를 신고한다.
			3.1.3.7	사용자는 주석오류 신고를 조회한다.
4. 배포	4.1 배포	4.1.1 파일별 배포		사용자는 표본파일을 내려받는다.
		4.1.2 말뭉치 배포		사용자는 말뭉치파일을 내려받는다.
		4.1.3 주석별 배포		사용자는 주석가공파일을 내려받는다.
		4.1.4 분석/통계/검색 색 배포		사용자는 분석/통계/검색 자료를 내려받는다.
5. 분석	5.1 색인			사용자는 학습자 말뭉치를 검색할 수 있도록 색인한다.
	5.2 검색	5.2.1 주석별 검색		사용자는 주석별 검색을 한다.
		5.2.2 키워드 기반 검색		사용자는 키워드로 검색을 한다.
		5.2.3 주석 패턴 검색		사용자는 주석으로 패턴 검색을 한다.
	5.3 통계	5.3.1 주석별 통계		사용자는 주석별 통계를 낸다.
		5.3.2 표본별 통계		사용자는 표본별 통계를 낸다.
		5.3.3 메타 데이터 기반 통계		사용자는 학습자 정보, 수집기관 등의 메타데이터를 기반으로 통계를 낸다.
	5.4 학습 및 분석	연관분석		하나 이상의 변인(요소)간의 상관관계를 분석/저장한다.

<표 11> 기능 명세서

4. 한국어 학습자 말뭉치 통합 관리 시스템 개발 이행안

4.1. 개요

4.1.1. 목표

한국어 학습자 말뭉치 대규모 자료를 효과적으로 관리, 보급, 활용하기 위한 시스템의 설계 및 중장기적 개발 이행안을 도출한다. 목표 시스템은 용어상 일관성을 위해 “한국어 학습자 말뭉치 통합 관리 시스템(Korean Learner’s Corpus Management System : KLCMS)”이라고 부르기로 한다.

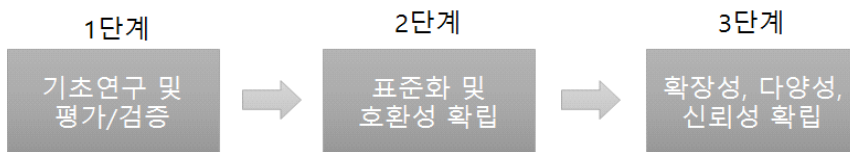
4.1.2. 목표 시스템 개발 추진 전략

4.1.2.1. 핵심 원칙

- 1) 모든 주석은 주석 직전 상태로 복원될 수 있어야 한다.
- 2) 모든 표본은 메타 데이터에 의해 조작될 수 있어야 한다.
- 3) 모든 자료는 원시 데이터(스캔, 음성 데이터)를 제외하고 변경될 수 있다.

4.1.2.2. 추진 전략

- 1) 2015~2020까지 단계별 개발
- 2) 한국어 학습자 말뭉치의 설계, 구축 공정, 활용방안, 평가/검증을 위한 다각적 연구
- 3) 한국어 학습자 말뭉치의 활용성을 극대화하기 위한 수집, 가공, 배포 관련 표준화 확립
- 4) 한국어 학습자 말뭉치 데이터의 품질 신뢰성, 시스템 확장성, 서비스 다양성 및 관리 효율성 확보

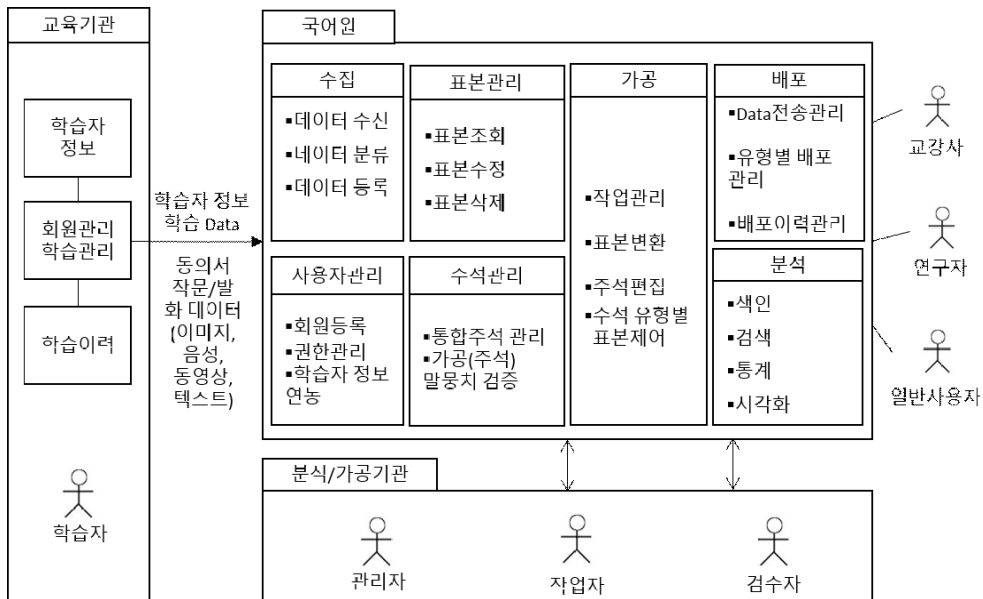


각 단계별 핵심 과제와 중점사항은 다음과 같다.

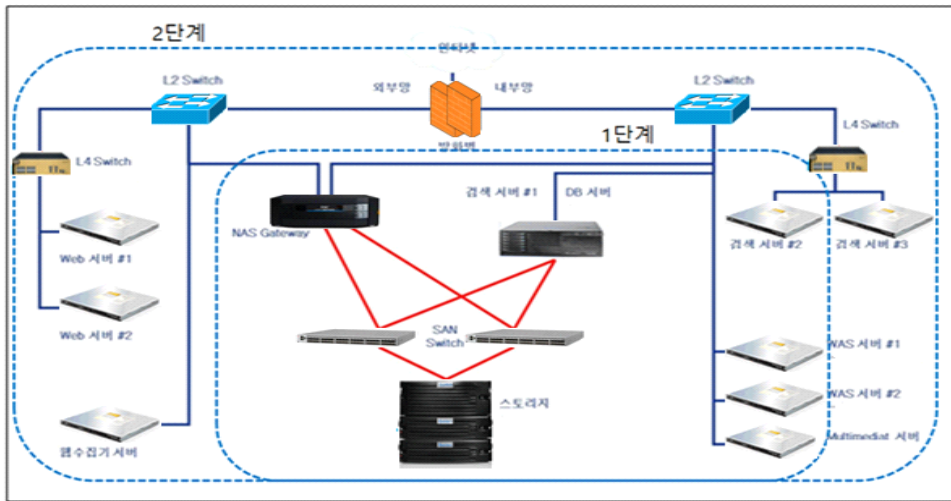
단계	핵심 과제	중점사항
1단계	- 다양한 유형의 학습자 말뭉치 수집 및 구축 - 학습자 말뭉치 수집, 구축, 활용간의 다양한 문제 도출 및 발전 방안 제시 - 호환성, 일관성이 보장되는 학습자 말뭉치 데이터 및 메타 데이터의 처리 방안과 SW	- 기초 연구 및 검증을 위한 초기 데이터 및 시스템 구축 - 수작업으로 인한 오류 배제 및 검증을 위한 도구 및 절차
2단계	- 다양한 수집원(기관)과의 온라인상에서 자료 교환이 가능하도록 데이터 및 데이터 처리에 대한 표준 또는 지침 수립 - 표준 또는 지침의 활용성 확대를 위한 호환성 지원 도구의 개발 및 배포	- 중앙 집중식 탈피 - 다양한 학습자 말뭉치 데이터의 호환성 확보를 통한 확장 기반 확보
3단계	- 학습자 말뭉치 이외의 언어 자원과의 연계 - 학습자 말뭉치의 데이터 품질 인증 - 다양한 학습자 말뭉치 활용 서비스 발굴	- 신뢰성, 편의성, 다양성을 갖춘 학습자 말뭉치 서비스 제공

4.2. 시스템 분석

4.2.1. 목표 시스템

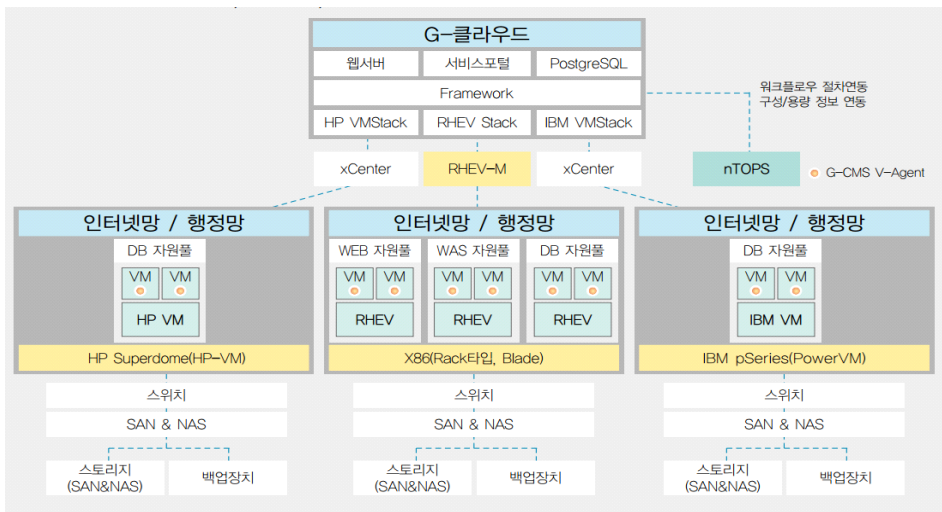


4.2.2. 하드웨어 구성¹⁾



<그림 50> 1,2단계 한국어 학습자 말뭉치 통합 관리 시스템 H/W 구성

3단계 : 정부통합전산센터 G-클라우드로의 통합²⁾

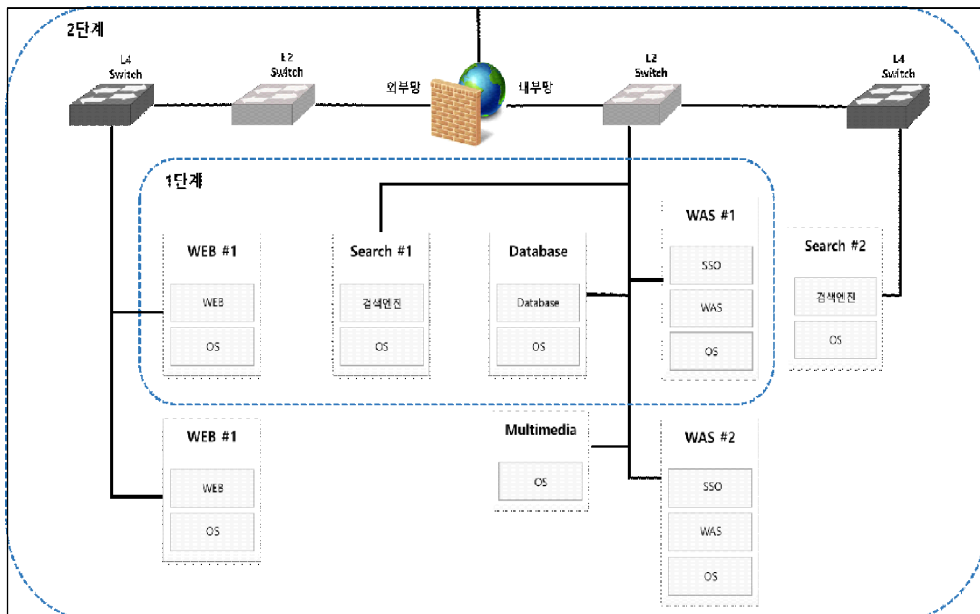


<그림 51> 정부통합전산센터 G-클라우드 구성(KLCMS 3단계 H/W 구성 시 적용)

- 1) 개별 하드웨어의 실제 사양은 구매 추진 단계에서 결정한다.
- 2) '09.12월, 범정부 클라우드 활성화 추진계획 발표를 기점으로, 관계부처 합동 클라우드 중합계획수립('10년~), 중앙부처정보 자원통합추진('08~'11년), 공공부문 클라우드 활용범위 확대('12년~)에 이르기까지 범정부적으로 지속적으로 추진 중(세부 사항은 NCIS(정부통합전산센터)의 "G-클라우드 소개자료" 참조.

4.2.3. 소프트웨어 구성

진행 단계	개발 방향	
1단계	한국어 학습자 말뭉치의 구축 관리 및 테스트용 (내부망 중심)	
2단계	내부망	한국어 학습자 말뭉치 자원 관리, 분석 및 테스트
	외부망	한국어 학습자 말뭉치 수집, 활용, 배포 서비스
3단계	사용량에 따른 부하 가변 확장 기능, 정부통합전산 정책 수용, 내외부망 표준 템플릿 적용	



<그림 52> 1,2단계 한국어 학습자 말뭉치 통합 관리 시스템 S/W 구성

공개 소프트웨어 표준 템플릿			템플릿 구성				
OS	구분	솔루션	구분	OS	WEB	WAS	DB
RHEL	WEB	JDK	OS용	RHEL AIX HP	-	-	-
		EWS(Apache)					
	WAS	JDK	WEB 서버용	RHEL	EWS (Apache)	-	-
		EAP(JBoss)					
	DBMS	JDK	WAS 서버용	RHEL	-	EAP (JBoss)	-
		CUBRID					
		CUBRID Manager					
			DB 서버용	RHEL	-	-	CUBRID

<그림 53> 정부통합전산센터 G-클라우드 구성(KLCMS 3단계 S/W 구성 시 적용)

4.3. 기능별 분석

4.3.1. 수집

4.3.1.1. 개요

	분석 결과 요약	핵심 요건 및 개선 방향
현황	- 작문, 음성 파일 우편 수집 - 구축기관 집중식 수집 - 수집원 의존적 자료 수집	- 원시자료, 학습자 정보, 메타데이터의 작업 분산 처리 지원 - 데이터 통합 및 메타데이터 표준화 - 기관별 수집정보 교환 인터페이스 확립
요구 사항	- 수집기관별 보유자료 등록 기능 - 학습자 정보 공유	
고려 사항	- 학습자 자율적 자료 제공 수집 기능 - 학습자 개인정보 보안 조치	

학습자 말뭉치의 수집은 국내외 한국어 교육을 담당하는 기관 및 한국어 보급처를 중심으로 수행된다. 대부분의 기관들은 자체적으로 학습자 말뭉치를 구축, 관리할 수 있는 정보화 체계를 갖추고 있지 않으므로, 학습자의 작문과 발화를 종이와 녹음 파일로 보유하고 있다. 이로 인해 구축기관은 각 수집기관으로부터 원시자료를 받아 디지털화하고, 학습자 정보와 메타데이터를 직접 구축하게 되므로 업무 부담으로 인해 자료의 검증을 충분히 수행하기 어렵다.

학습자 말뭉치 수집 미래 모형의 핵심 요건은 구축기관의 자료 수집 부하 분산과 수집된 데이터 및 메타데이터를 통합할 수 있는 표준화, 그리고 이를 지원하기 위한 각 기관의 정보화 인터페이스 확립에 있다. 이를 통해 구축기관은 수집 업무의 부담을 경감하고, 수집기관은 자체 학습자 말뭉치 구축, 관리하는 기술적 수단을 제공받을 수 있다.

또한, 학습자의 자율적 자료 제공이 가능한 수단을 제공함으로써, 수집된 자료의 저작권 및 개인정보의 침해 문제를 해소하여 자료의 활용성을 높여야 한다. 이러한 권리적 문제와는 별개로 학습자의 개인정보를 철저히 보호할 수 있도록 논리적 구조상에서 보안 설계가 수반되어야 한다.

4.3.1.2. 단계별 개발 계획

단계	개발 목표	목표 결과물
1단계	- 메타데이터 표준화 - 표본 데이터 등록, 전송, 저장 및 데이터 통합 처리 시스템	웹기반 학습자 말뭉치 표본 등록 시스템
2단계	- 기관 간 정보교환 표준 인터페이스 - 데이터 검증 시스템 - 말뭉치 설계 기관 표본 자동 분류/저장 시스템	- 기관 간 연계 시스템 - 학습자 말뭉치 표본 자동 수집 시스템
3단계	준거 말뭉치 등 참조용 데이터 통합 및 확장	학습자 말뭉치 및 참조 데이터 수집/통합 시스템

4.3.1.3. 주요 요구 사항

- 1) 다양한 데이터 수집처로부터 온라인으로 자료 수집
- 2) 수집된 데이터가 일관된 형식으로 저장되고 처리될 수 있어야 함

- 3) 수집간 작업자 오류로 인한 데이터 관리, 처리 혼선 방지
 - 파일명, 파일 분류 등이 잘못되더라도 입력/전사, 주석 등의 데이터의 연계는 유지
- 4) 학습자의 개인 정보 사용은 최소화하고, 온전한 보완 대책이 강구되어야 함
- 5) 학습자 동의서 및 동의 요건 변경

4.3.2. 관리

4.3.2.1. 개요

	분석 결과 요약	핵심 요건 및 개선 방향
현황	• 데이터별 파일 및 폴더화 관리 • 지침 및 주석 유형별 파일화 관리 • 변경에 대한 범위 및 위험관리 부재	• 메타데이터 중심의 추상 모델 기반 통합 관리 • 주석 형식 및 관리 체계의 일원화 • 변경에 대한 범위 및 위험관리 프로세스 정립
요구 사항	• 주석 유형별 주석 형식 관리 • 수집, 가공된 표본들의 현황 제공	
고려 사항	• 변경에 대한 범위 및 위험관리 • 변경에 따른 재작업 소요 최소화 방안	

학습자 말뭉치의 관리는 데이터들의 현황을 관리자에게 제공하고, 데이터의 추가, 삭제, 변경에 대한 물리적 수준의 적용을 처리하는 데이터 처리로 구성된다.

현행 학습자 말뭉치의 관리는 메타데이터(표본 정보 및 학습자 정보 등)와 말뭉치 설계 정보, 주석 체계, 표본 파일을 관리자가 직접 추가, 삭제, 변경, 이동, 저장함으로써 관리 소요의 부담뿐만 아니라, 데이터 무결성과 보존성에 있어서도 위험요인이 존재한다.

학습자 말뭉치 수집 미래 모형은, 관리자에게 메타데이터를 중심으로 한 데이터 관리 수단을 부여함으로써 데이터를 직접 조작으로 인한 위험요인과 관리 소요 부담을 제거한다. 또한 주석 형식 및 관리 체계를 일원화하고, 변경에 대한 범위 및 위험관리 프로세스를 정립하여 제공함으로써 학습자 말뭉치의 가공, 분

석, 배포, 활용 서비스에서 발생할 수 있는 데이터의 비밀관성과 신뢰성 저하를 방지하도록 한다.

4.3.2.3. 단계별 개발 계획

단계	개발 목표	목표 결과물
1단계	- 비정형 데이터 통합 관리 (텍스트, 이미지) - 기계 가독한 통합 주석 모델	- 비정형 데이터 통합 관리 시스템 - 통합 주석체계 관리 시스템
2단계	- 비정형 데이터 통합 관리 (영상, 음성 포함) - 통합 주석관리 호환/확장 - 주석체계, 지침 변경 관리	- 메타데이터 기반 통합 말뭉치 관리 시스템 - 말뭉치 데이터 변경관리 통제 시스템
3단계	대규모 데이터 트랜잭션 관리	

4.3.2.3. 주요 요구 사항

- 1) 파일 및 데이터를 직접 조작하지 않고 관리할 수 있어야 한다.
- 2) 파일정보, 학습자 정보 수정 시 말뭉치 표본의 헤더 정보가 함께 변경될 수 있어야 한다.
- 3) 형태 주석, 오류 주석, 원시 표본 수정 시 연관된 정보들이 함께 변경될 수 있어야 한다.
- 4) 수집된 표본, 설계된 말뭉치의 현황(자료량, 가공 수준, 작업 현황 등)을 직관적으로 볼 수 있어야 한다.

4.3.3. 가공

4.3.3.1. 개요

	분석 결과 요약	핵심 요건 및 개선 방향
현황	- 표현 중심의 수작업 가공(주석 입력) - 입출력 비표준화에 따른 후처리/변환 작업 - 다양한 상용도구 사용 - 비가시적 정보로 인한 데이터 오염 및 손상 - 형식, 지침변경 등에 의한 전수 작업 부담으로 주석 품질관리 제약	- 통합 주석에 의한 주석 데이터 일관성 확립 - 공정간 데이터 검증 및 호환성 - 데이터 변경에 따른 무결성 - 주석 입력(타이핑)에서 주석 삽입 형식의 작업 UI - 작업 유형에 따른 작업환경 자동 구성 및 편의성 UI - 주석 결과 미리보기 - 반복/규칙성 발굴을 통한 작업 환경 및 자동화 지속 개선
요구 사항	- 오류 주석 특성을 반영 (단어, 구, 문장, 담화 등 전 층위 문법 단위/요소에 대한 주석 및 교정값 처리) - 작업 효율을 위한 직관적 UI - 작업자 입력 오류 배제 방안	
고려 사항	작업 진행에 따른 UI 변경 수용성	

학습자 말뭉치의 가공은 주석 작업의 효율성과 편의성, 주석 결과로 인한 데이터의 오염/손상 방지 문제가 중심이 된다.

말뭉치의 주석 가공은 숙련된 전문가에 의한 주석 할당 작업이 필수적이다. 전문가의 숙련도는 할당하는 주석의 내용적 적합성에 대한 것으로, 주석 표지의 이름 입력, 대상 데이터의 전산적 처리와는 구분된다. 또한 전문가의 전문성은 대체로 주석 유형(입력/전사, 형태주석, 오류주석 등)에 제한적인 것이므로, 연관된 선후행 작업에 대해서는 영향성을 고려하지 않는 경우가 많다.

그러므로 학습자 말뭉치 가공에 있어서 미래 모형은, 데이터 내/외적 처리 일관성을 고려한 통합 주석 체계 및 처리 모듈을 바탕으로 공정간 데이터를 검증하고, 호환성과 무결성이 유지되도록 데이터를 처리하여 주석 가공 작업자에게 제공하는 것이 핵심이다.

이와 함께 주석의 명칭과 형식을 가공 작업자가 직접 입력하던 방식에서 도구가 제공하는 작업환경상의 주석을 선택 입력하는 방식으로 개선함으로써 작업자 입력에 의한 오류를 방지한다. 또한, 주석 가공에 필요한 문맥, 메타데이터, 선행

작업 결과 및 주석 결과 미리보기 등과 반복/규칙성 발굴을 통한 자동화 기능 등 편의성 UI를 지속적으로 제공함으로써 작업 효율과 품질을 향상시킨다.

4.3.3.2. 단계별 개발 계획

단계	개발 목표	목표 결과물
1단계	통합 주석 체계를 이용한 일관된 주석 작업 지원	- 주석 편집 작업도구 (작업 편의성 UI) - 공정간 데이터 관리 시스템 - 공정별 데이터 제어 모듈
2단계	변경 관리 통제 프로세스	- 입출력 표준화를 통한 외부 가공 자료 통합 시스템 - 주석, 데이터 변경/수정 추적관리 체계 및 시스템
3단계	- 데이터 무결성 검증 관리 - 주석, 지침, 데이터 변경에 따른 신뢰성 예측 모델	말뭉치 통합 검증 시스템

4.3.3.3. 주요 요구 사항

- 1) 단어, 구, 문장, 담화 등 전 층위 문법 단위/요소에 대한 주석 및 교정값 처리 등 다층위, 다중 주석이 필요한 학습자 오류 주석 특성 반영
- 2) 작업자 입력 오류를 배제한 주석 편집 UI
- 3) 문맥, 선행 분석(형태 주석, 본문 주석 등)의 작업 시 필요한 관련 정보 제공
- 4) 주석 변경 이력 저장 및 조회
- 5) 주석 입력 단축키 등 작업 편의성 기능 및 UI
- 6) 주석 부분 제거/일관 제거 기능
- 7) 주석 형식 오류 검증
- 8) 형태 주석 자동/반자동 생성 및 편집 기능
- 9) 가공 공정별 작업자 접근 권한 관리
- 10) 오류 항목 신고 기능

4.3.4. 분석

4.3.4.1. 개요

	분석 결과 요약	핵심 요건 및 개선 방향
현황	- 주석별 빈도 중심의 통계 1회적 분석 정보 생성 - 데이터 처리 중심의 연구과제별 분석	- 모범적 사례 제시 기반 분석 서비스 제공 - 분석결과를 재사용할 수 있도록 가공 가능한 결과 내려받기 제공
요구 사항	다양한 연구/서비스를 위해 사용자 분석 부담 최소화	
고려 사항	분석도구의 의존성 최소화	

4.3.4.2. 단계별 개발 계획

단계	개발 목표	목표 결과물
1단계	용례검색, 기초통계(빈도 및 분포), 고급 통계 등 제공	- 주석 유형별 용례 추출 및 검색 시스템 - 기초/고급 통계 계산 모듈 - 통계 결과 검색 시스템 - 통계 시각화 모듈
2단계	기계학습 등 고급 데이터 분석 제공	고급 데이터 분석, 검색, 시각화 시스템
3단계	- 사용자 요구에 따른 대행 분석 - 능동적 분석/검증	사용자 정의형 데이터 분석/검증 서비스 시스템

4.3.4.3. 주요 요구 사항

- 1) 원시 어절, 어절의 일부, 형태, 품사, 오류 유형, 오류 교정값 등 말뭉치 표본의 유형별 값에 의한 용례 검색이 가능해야 한다.
- 2) 검색 및 통계는 부분적으로 실시간 업데이트가 가능해야 한다.
- 3) 서비스화된 기능은 사용자의 사용 패턴, 사용량 등의 이용 현황 통계가 수집되어야 한다.
- 4) 분석결과(용례, 검색, 통계 분석)는 사용자가 재활용 가능한 형태로 내려받을 수 있어야 한다.

4.3.5. 배포 및 활용 서비스

4.3.5.1. 개요

	분석 결과 요약	핵심 요건 및 개선 방향
현황	- 수집, 구축된 결과 파일 중심 서비스 - 가공 형식에 의존적인 데이터 배포 - 정적인 활용 도구 및 서비스	사용자 필요에 따른 다양한 주석 정보 통합 제공
요구 사항	다양한 외부 도구를 활용할 수 있는 다양한 형식의 배포	템플릿 기반의 사용자 배포 형식 지원
고려 사항	- 개인 정보, 저작권 등에 따른 배포, 서비스 권한 통제 - 서비스 데이터에 대한 변경 제한 등 데이터 안전성 확보	서비스 확장, 변경이 용이한 동적 서비스 구성 기반 확립

4.3.5.2. 단계별 개발 계획

단계	개발 목표	목표 결과물
1단계	- 사용자 요구에 따라 주석별, 파일 형식별로 형식을 재구성하는 배포 - 사용자 필요에 따라 표본을 재구성한 말뭉치 배포	- 형식 재구성 배포 시스템 - 사용자별 말뭉치 재구성 서비스³⁾
2단계	- 배포 자료의 내용, 형식에 대한 사용자 정의 구성 - 주요 상용/공개 S/W와의 호환	- 템플릿 기반 사용자 정의 형 배포 시스템 - Export Wizard
3단계	- 대규모 데이터 배포 관리 - 배포 없는 대규모 데이터 활용 지원 시스템	클라우드 기반 개인형 학습자 말뭉치 통합 관리 시스템

4.3.5.3. 주요 요구 사항

- 1) 사용자는 필요에 따라 내려받을 표본 또는 말뭉치의 내용과 형식을 결정할 수 있어야 한다.
- 2) 사용자가 배포 자료의 내부 구성을 조정할 수 있는 수단(템플릿 등)을 제공해야 한다.

3) 사용자는 자신만의 말뭉치를 설계하고 표본을 구성하여, 그 분석결과를 내려받아 활용할 수 있다.

- 3) 내려받은 자료는 기계가독성이 보장되어야 한다.
- 4) Excel, 텍스트 편집기, DBMS SQL 등 상용/공개 S/W와 호환 가능한 출력 결과를 생성하는 도구를 지속적으로 추가할 수 있어야 한다.
- 5) 사용자가 자체적인 표본 가공, 구축, 재구성이 가능한 수단 및 도구의 제공을 고려한다.
- 6) 사용자 요청에 의한 서비스 제공에 의해 서비스 데이터가 변경되지 않아야 한다.

4.4. 학습자 말뭉치 통합 관리 시스템 개발 로드맵

구분		1단계		2단계		3단계	
		2015	2016	2017	2018	2019	2020
목표		기초 연구 - 공정, 데이터의 관리, 처리 방안 개발 및 검증		도입 및 안정화 - 표준화 및 호환성 수립 - 외부 기관 및 자원 연계		서비스 고도화 - 데이터 및 서비스의 확장성, 다양성, 신뢰성 확립	
시스템 개발		학습자 말뭉치 구축 지원 도구	학습자 말뭉치 통합 관리 시스템 (기관 연계)		학습자 말뭉치 연구/활용 서비스 시스템		학습자 말뭉치 연구/활용 서비스 시스템 (클라우드 기반)
DB 구축 (학습자 말뭉치)		학습자 말뭉치 구축 및 기초 연구 참조					
인프라 도입	하드웨어	해당 없음	웹서버 (2식) 웹어플리케이션 서버 (2식) 검색엔진 (3식) 비정형 데이터 관리 서버 (1식) 데이터베이스 서버 (1식)			G-클라우드로 전환	
	소프트웨어	해당 없음	웹서버(1식) 웹어플리케이션 서버(1식) 검색엔진(1식) 데이터베이스 서버 (1식)	웹서버(1식) 웹어플리케이션 서버(1식) 비정형 데이터 관리 서버 (1식)	검색엔진 (2식)		

<표 12> 학습자 말뭉치 통합 관리 시스템 개발 로드맵

Ⅲ. 한국어 학습자 말뭉치 구축 지원 도구 개발

한국어 학습자 말뭉치 통합 관리 시스템은 웹서비스를 통해 모든 기능을 제공하는 것을 원칙으로 한다. 다만, 본 연구 과정에서는 사용자의 GUI 컴포넌트 조작이 많은 주석 편집의 효율적인 개발을 위해 PC-Application 개발 후 web으로 통합해 나가는 Proto-type 기반의 점진적 개발 방식을 적용한다. 그러므로 ‘웹페이지 개발’ 등의 표현은 전통적으로 웹페이지에서 주로 구현되는 기능을 의미하는 정도일 뿐, 최종 배포 환경의 제약을 의미하지 않는다.

여기에서는 화면 설계를 중심으로 세부 기능, 동작 절차(Business Logic), 데이터 스키마, 클래스 다이어그램 등을 제시한다.

1. 한국어 학습자 말뭉치 웹페이지 개발

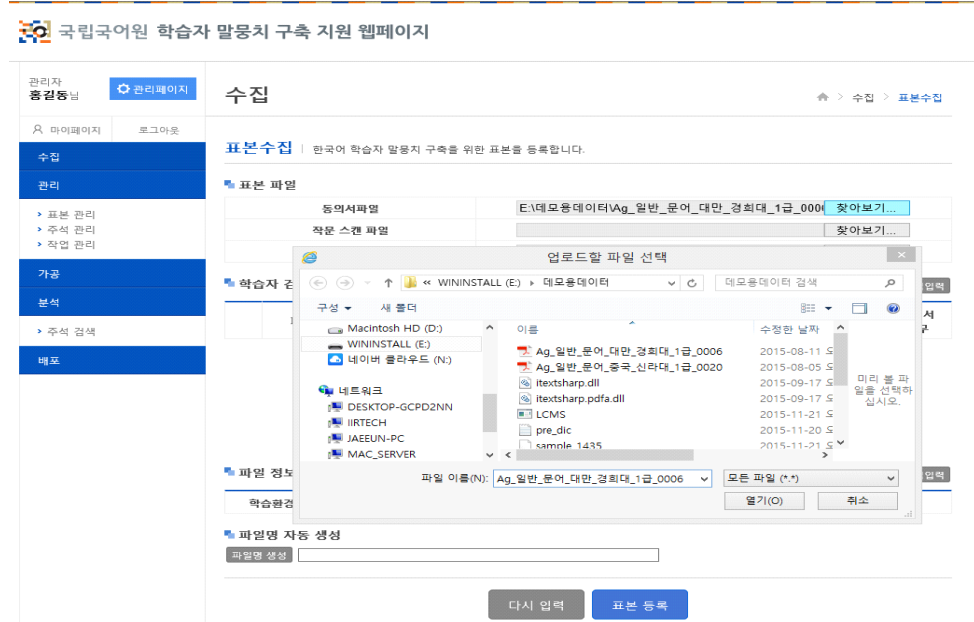
ID	화면명		화면 설명
1		메인 화면	웹 메인 페이지 (통계 그림 포함)
2		로그인	관리자, 작업자, 교/강사, 학습자별 로그인
3	수집	표본 수집 화면	표본 파일/정보 등록 페이지
4		표본 수집 - 학습자 정보 직접 입력	학습자 정보 입력 페이지
5	관리	표본 관리 화면	표본 검색 조회 페이지
6		표본 관리 - 상세 보기	표본 파일 상세 보기 페이지
7		표본 관리 - 표본 수정	표본 파일/정보 수정 페이지
8		주석 관리 화면	주석 추가/수정/삭제 페이지
9	분석	주석 검색 화면	표본의 주석 검색 페이지
10	배포	배포 화면	표본 다운로드 페이지
11		배포 - 표본 다운로드 정보 선택 창	표본 다운로드 정보 선택 페이지

<표 13> 한국어 학습자 말뭉치 웹페이지 화면 목록

1.1. 표본 수집

1.1.1. 파일 업로드

가. 화면



나. 요구 명세

ID	1.1.1	요구명	표본파일 업로드 기능
내용		사용자는 말뭉치 표본파일을 업로드한다.	
사전조건		사용자는 웹 페이지의 회원이다.	
		사용자는 관리자/교사, 강사/학습자/일반사용자를 말한다.	
사후조건		페이지는 이동하지 않는다.	
작업 흐름			
1. 사용자는 해당하는 파일별로 찾아보기 버튼을 누른다. 2. 파일 선택창이 뜬다. 3. 사용자는 업로드 할 해당 표본 파일을 선택한다. 4. 선택된 파일의 경로가 표시된다.			

1.1.2. 학습자 정보 검색

가. 화면

국립국어원 학습자 말뭉치 구축 지원 웹페이지

관리자
홍길동님
관리페이지
수집
수집 > 수집 > 표본수집

마이페이지
로그아웃
수집
관리
표본 관리
주석 관리
작업 관리
가공
분석
주석 검색
배표

표본수집 | 한국어 학습자 말뭉치 구축을 위한 표본을 등록합니다.

표본 파일

동역서파일
E:\대용데이터\Ag_일반_문어_대만_경희대_1급_000(
찾아보기...

작문 스캔 파일
E:\대용데이터\일반_문어_대만_경희대_1급_0006.pdf
찾아보기...

업로드/전사 파일
E:\대용데이터\일반_문어_대만_경희대_1급_0006.txt
찾아보기...

학습자 검색결과
1918
학습자검색
학습자작업입력

	ID	나이	성별	국적	모국어	한국어 등급	학습 기관	학습 기간	가주 기간	동역서 유무
☐	1918	99	남	중국	영어	1급	경희대학교	9	9	유

파일 정보 입력

학습환경
☐ 국내
☐ 국외

표본유형
☐ 구어
☐ 문어

파일명 자동 생성

파일명 생성

나. 요구 명세

ID	1.1.2	요구명	학습자 정보 검색 기능
내용		사용자는 학습자 정보를 검색한다.	
사전조건		사용자는 웹 페이지의 회원이다.	
		사용자는 관리자/작업자/교사, 강사를 말한다.	
사후조건		페이지는 이동하지 않는다.	
작업 흐름			
1. 사용자는 학습자 ID을 검색한다. 2. DB에 tbl_learnerinfo테이블에서 해당 ID값을 조회한다. 3. 학습자 ID와 일치하는 ID값을 찾는다. 4. 일치하는 ID값의 데이터 행이 학습자 검색 결과창에 출력된다. 4-1. 일치하는 데이터가 없으면 아무 반응이 일어나지 않는다. 5. 조회한 학습자 정보를 선택한다.			

1.1.3. 파일 정보 입력

(1) 학습자 정보 직접 입력

가. 화면

나. 요구 명세

ID	1.1.3	요구명	학습자 정보 직접 입력 기능
내용		사용자는 학습자 정보를 직접 등록한다.	
사전조건		사용자는 웹 페이지의 회원이다.	
		사용자는 관리자/작업자/교사, 강사를 말한다.	
		사용자는 학습자 조회를 했다.	
사후조건		페이지는 이동하지 않는다.	
작업 흐름			
1. 사용자는 조회되지 않는 학습자 정보를 직접 입력한다.			
2. 학습자 직접 입력 버튼을 누른다.			
3. 학습자 직접입력창이 뜨면 등록하고자 하는 학습자 정보를 입력한다.			
4. 확인 버튼을 눌러 학습자 정보를 저장한다.			
4-1. 모든 입력창에 정보를 입력하지 않고 확인 버튼을 누른다.			
4-2. 입력 경고 알람창이 뜬다.			
4-3. 확인 버튼을 눌러 알림창을 닫는다.			
5. 등록한 학습자 ID로 학습자 정보를 검색한다.			
제약조건		학습자는 ID에 의해서만 검색한다.	

(2) 파일 정보 입력

가. 화면

관리자
홍길동님

관리페이지

마이페이지

로그아웃

수집

관리

표본 관리

주석 관리

작업 관리

가공

분석

주석 검색

배포

수집

표본수집

표본수집 | 한국어 학습자 맞춤형 구축을 위한 표본을 등록합니다.

표본 파일

동역서파일	E:\데모용데이터\Ag_일반_문어_대만_경희대_1급_0006_ 찾아보기...
작문 스캔 파일	E:\데모용데이터\일반_문어_대만_경희대_1급_0006.pdf_ 찾아보기...
압력/전사 파일	E:\데모용데이터\일반_문어_대만_경희대_1급_0006.txt_ 찾아보기...

학습자 검색결과

1918

학습자 검색

학습자 직접입력

	ID	나이	성별	국적	모국어	한국어 등급	학습 기관	학습 기간	거주 기간	동역서 유무
☒	1918	99	남	중국	영어	1급	경희대학교	9	9	유

파일 정보 입력

파일정보 직접입력

학습환경	☒ 국내 ☐ 국외	표본유형	☐ 구어 ☒ 문어
------	--	------	--

파일명 자동 생성

파일명 생성

일반 문어 중국 경희대학교 1급

나. 요구명세

ID	1.1.5	요구명	파일 정보 등록 기능
내용		사용자는 파일 정보를 등록한다.	
사전조건	사용자는 학습자 정보를 등록한다.		
	사용자는 일반사용자/학습자/교사, 강사/관리자를 말한다.		
사후조건			
작업 흐름			
1. 사용자는 해당하는 파일의 정보를 체크한다.			
2. 파일명 자동 생성 버튼을 누른다.			
3. 파일 정보가 DB에 저장된다.			

1.1.4. 파일명 자동 생성

가. 화면

국립국어원 학습자 말뭉치 구축 지원 웹페이지

관리자 홈결통님 관리페이지

수집

표본수집 | 한국어 학습자 말뭉치 구축을 위한 표본을 등록합니다.

표본 파일

동역서파일 E:\데모용데이터\Ag_일반_문어_대만_경희대_1급_000(...) 찾아보기 ...

작문 스캔 파일 E:\데모용데이터\일반_문어_대만_경희대_1급_0006.pdf 찾아보기 ...

입력/전사 파일 E:\데모용데이터\일반_문어_대만_경희대_1급_0006.txt 찾아보기 ...

학습자 검색결과 1918 학습자검색 학습자 직접입력

ID	나이	성별	국적	모국어	한국어 등급	학습 기관	학습 기간	거주 기간	동역서 유무
1918	99	남	중국	영어	1급	경희대학교	9	9	유

파일 정보 입력

학습환경 ☒ 국내 ☐ 국외

표본유형 ☐ 구어 ☒ 문어

파일명 자동 생성

파일명 생성 일반_문어_중국_경희대학교_1급

다시 입력

표본 등록

나. 요구 명세

ID	1.1.6	요구명	파일명 자동 생성 기능
내용		사용자는 파일명을 자동 생성한다.	
사전조건		사용자는 표본파일을 등록한다.	
		사용자는 학습자 정보를 등록한다.	
		사용자는 파일 정보를 등록한다.	
사후조건			
작업 흐름			
1. 사용자는 표본의 모든 정보를 입력한 후, 파일명 자동 생성 버튼을 누른다. 2. 입력된 학습자 정보와 파일 정보가 조합되어 파일명이 생성된다. 3. 파일명 자동 생성란에 결합된 값이 출력된다.			

1.1.5. 표본 파일 저장

가. 화면

관리자 **표본수집** 관리페이지

수집 > 수집 > **표본수집**

표본수집 | 한국어 학습자 알동지 구축을 위한 표본을 등록합니다.

표본 파일
 동의서파일: E:\데모용데이터\Ag_일반_문어_대만_경희대_1급_0001_찾아보기...
 작문 스캔 파일: E:\데모용데이터\일반_문어_대만_경희대_1급_0006.pdf_찾아보기...
 입학/전사 파일: C:\Users\mia\Desktop\업무용\데\데모용데이터\데모용데_찾아보기...

학습자 검색결과 1918 학습자검색 학습자 직접입력

ID	나이	학습 기간	학습 기간	거주 기간	동역서 유무
1918	99	경희대학교	9	9	유

파일 정보 입력 학습환경: ☒ 국내 ☐ 국외 ☐ 구어 ☒ 문어

파일명 자동 생성
 파일명 생성 일반 문어 중국 경희대학교 1급

다시 입력 **표본 등록**

© 2015 국립국어원

나. 요구 명세

ID	1.1.7	요구명	표본 저장 기능
내용		사용자는 생성된 파일명으로 표본을 저장한다.	
사전조건		사용자는 표본파일을 등록한다.	
		사용자는 학습자 정보를 등록한다.	
		사용자는 파일 정보를 등록한다.	
		사용자는 파일명 자동 생성을 한다.	
사후조건			
작업 흐름			
1. 사용자는 표본파일 등록 버튼을 누른다. 1-1. 표본파일 업로드, 학습자 정보, 파일 정보, 파일명 생성 중 누락된 것이 있다. 1-2. 모든 정보를 입력하라는 경고창을 띄운다. 1-3. 사용자는 확인 버튼을 눌러 경고창을 닫는다. 1-4. 표본 정보를 다시 입력하려면 다시 입력 버튼을 누른다. 2. 표본파일이 자동 생성된 파일명으로 DB에 저장된다. 3. Tbl_real_file테이블에 파일정보가 자동 생성 파일명으로 저장된다.			

```

graph TD
    Start([시작]) --> Login[로그인 세션  
- UserID  
- UserPW]
    Login --> Auth{권한체크}
    Auth -- N --> Warn1[접근권한X  
경고창]
    Auth -- Y --> Select{파일선택}
    Select -- N --> Warn2[파일선택 경고창]
    Select -- Y --> Search[학습자이름검색]
    Search --> Reg{학습자  
등록여부}
    Reg -- N --> Warn3[조회 불가  
경고창]
    Reg -- Y --> List[학습자 목록 출력]
    List --> Check{동의서  
파일 유무  
체크}
    Check -- N --> Warn4[동의서 파일 등록  
경고창]
    Check -- Y --> RegFile{동의서  
파일 등록}
    RegFile -- N --> Warn5[동의서 파일 등록  
경고창]
    RegFile -- Y --> Info{파일  
정보입력}
    Info -- N --> Warn6[파일정보입력  
경고창]
    Info -- Y --> Gen{파일명  
자동생성}
    Gen -- N --> Warn7[파일명자를 생성  
경고창]
    Gen -- Y --> RegFile2[표본파일등록]
    RegFile2 --> Commit[Commit]
    Commit --> End([종료])
    
```

로그인 세션

- UserID
- UserPW

사용자정보

- UserID(key)
- Usertype
- .
- .
- .

학습자 정보

- UserID(key)
- Name
- Age
- Gender
- National
- M_tongue
- Level
- Learn_period
- Residence_period
- 동의서 유무
- Learning_Organ

파일정보

- FileD(key)
- Mcode_Env
- Mcode_gatherOrgan
- Mcode_LearnerType
- Dcode_MotherTongue
- Dcode_LearnerLevel
- Dcode_SourceType

파일데이터

- Id(key)
- File_infold
- File_size
- File_type
- Blobcata

파일정보

- Useric(key)
- Mcode_Env
- Mcode_gatherOrgan
- Mcode_LearnerType
- Dcode_MotherTongue
- Dcode_LearnerLevel
- Dcode_SourceType

임시학습자정보

- Temp_ic
- name
- age
- gender
- national
- m_tongue
- level
- learn_period
- residence_2
- eriod

1.2. 관리

1.2.1. 표본 관리

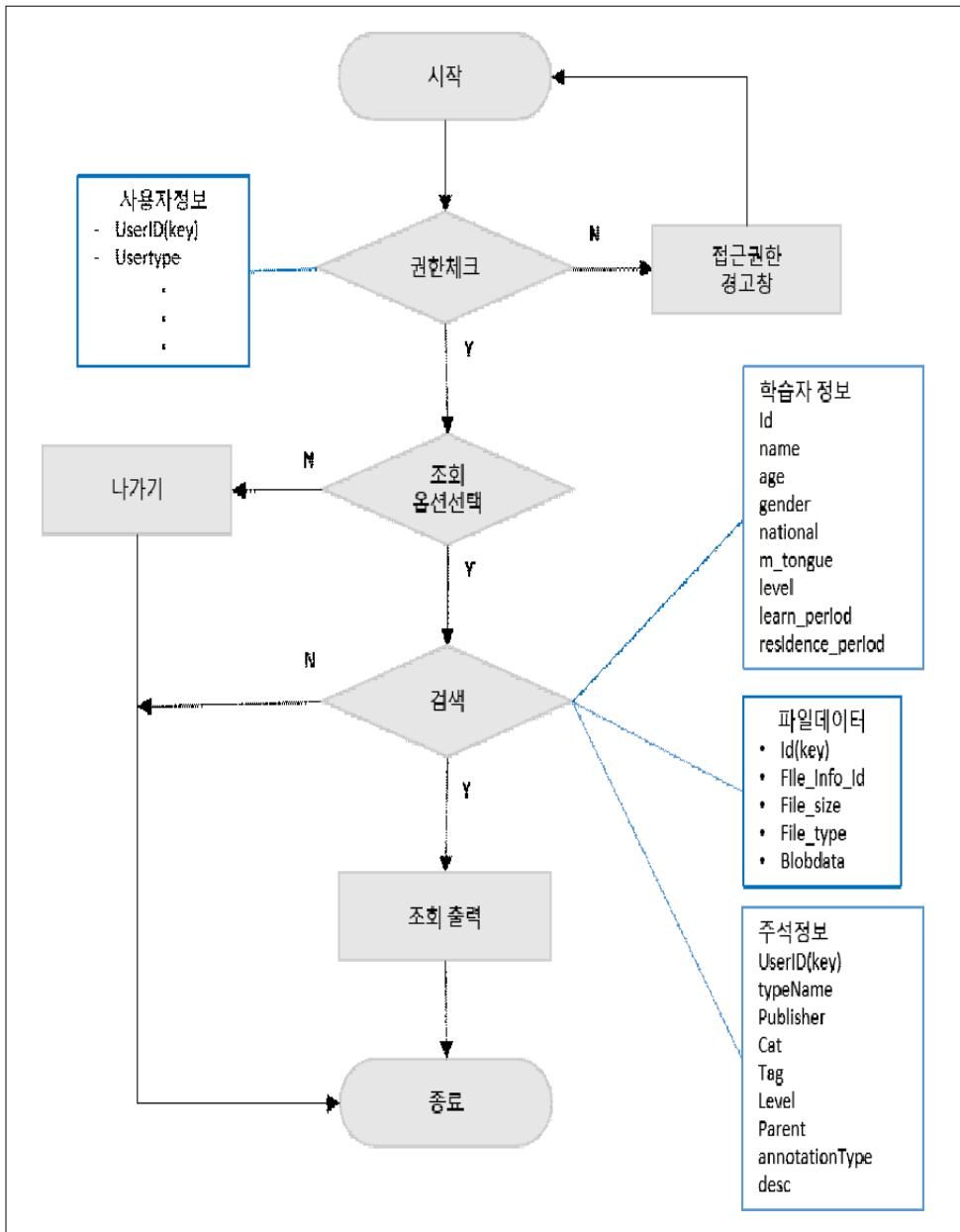
(1) 주식 유형별 조회 및 파일 정보 조회

가. 화면

나. 요구 명세

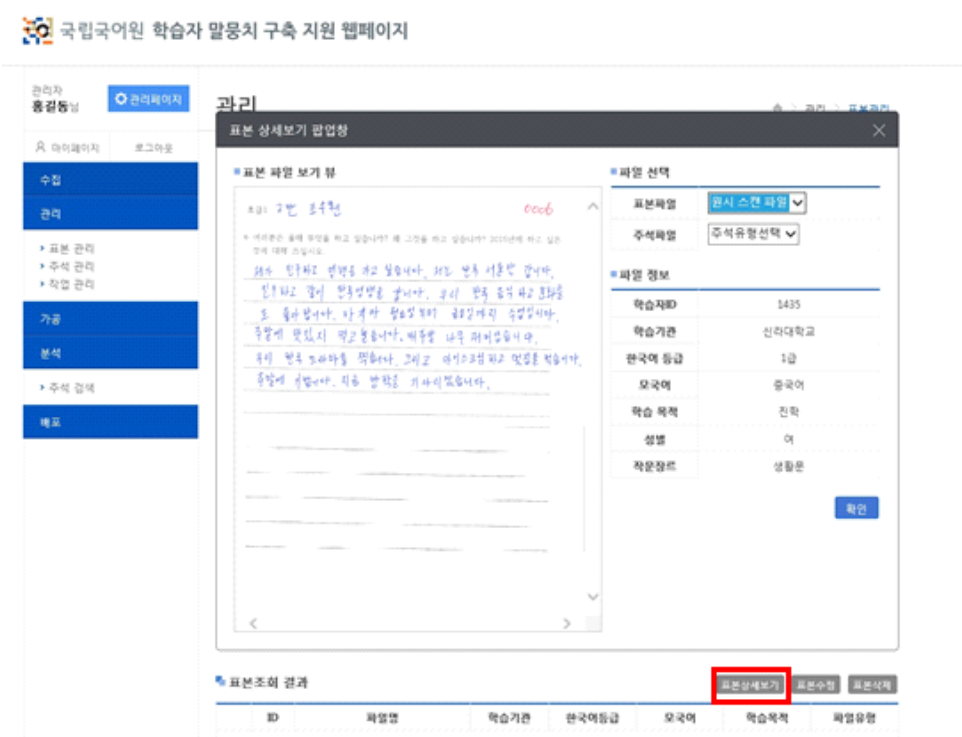
ID	2.2.1.2	요구명	파일 정보별 조회
내용		사용자는 표본을 파일 정보별로 조회한다.	
사전조건		해당 없음	
사후조건		해당 없음	
작업 흐름			
1. 사용자는 표본관리 페이지를 누른다. 2. 사용자는 조회하고자 하는 표본의 파일 정보를 입력한다. 4. 사용자는 검색 조회 버튼을 누른다. 5. 입력한 파일 정보에 해당하는 표본이 표본 조회 결과에 출력된다.			
제약조건		검색 조건은 모두 입력하지 않아도 된다. (일부 검색 기능)	

다. 액티비티 다이어그램



(2) 표본 열람 (상세보기)

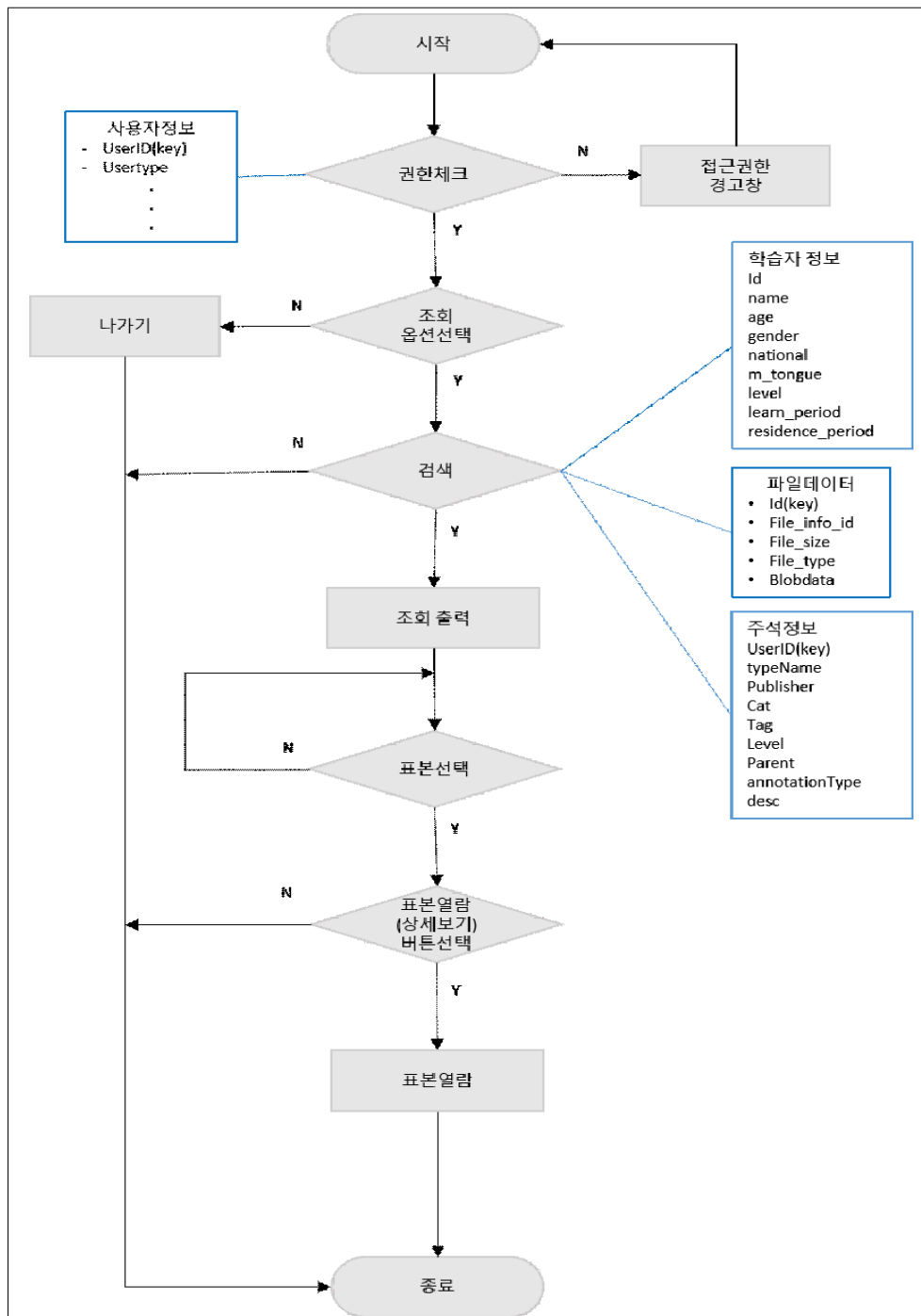
가. 화면



나. 요구 명세

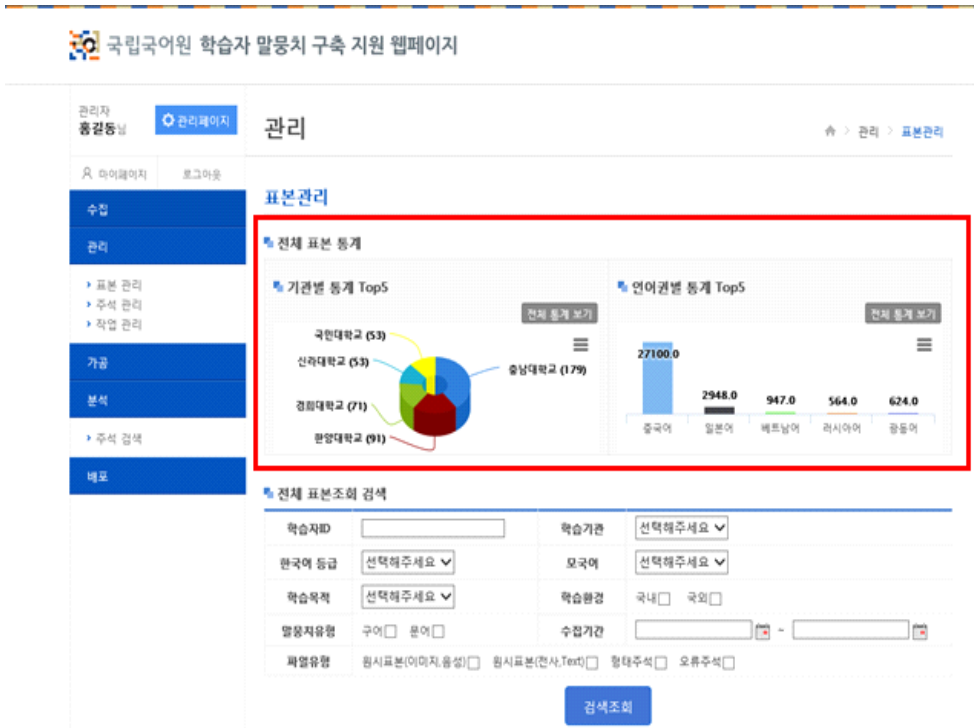
ID	2.2.1.3	요구명	표본 상세보기
내용		사용자는 표본에 대한 정보를 열람(상세보기)한다.	
사전조건		정보를 열람할 표본을 조회한다.	
사후조건			
작업 흐름			
1. 사용자는 표본 관리 페이지를 누른다.			
2. 사용자는 열람할 표본에 대한 조회 옵션을 선택한다.			
3. 표본 검색 조회 버튼을 누른다.			
4. 조회 된 표본 중 열람하고자 하는 표본을 체크한다.			
5. 표본 상세보기 버튼을 누른다.			
6. 표본에 해당하는 이미지, 파일 정보가 상세보기 팝업창에 출력된다.			

다. 액티비티 다이어그램



(3) 표본 가공률 통계 조회

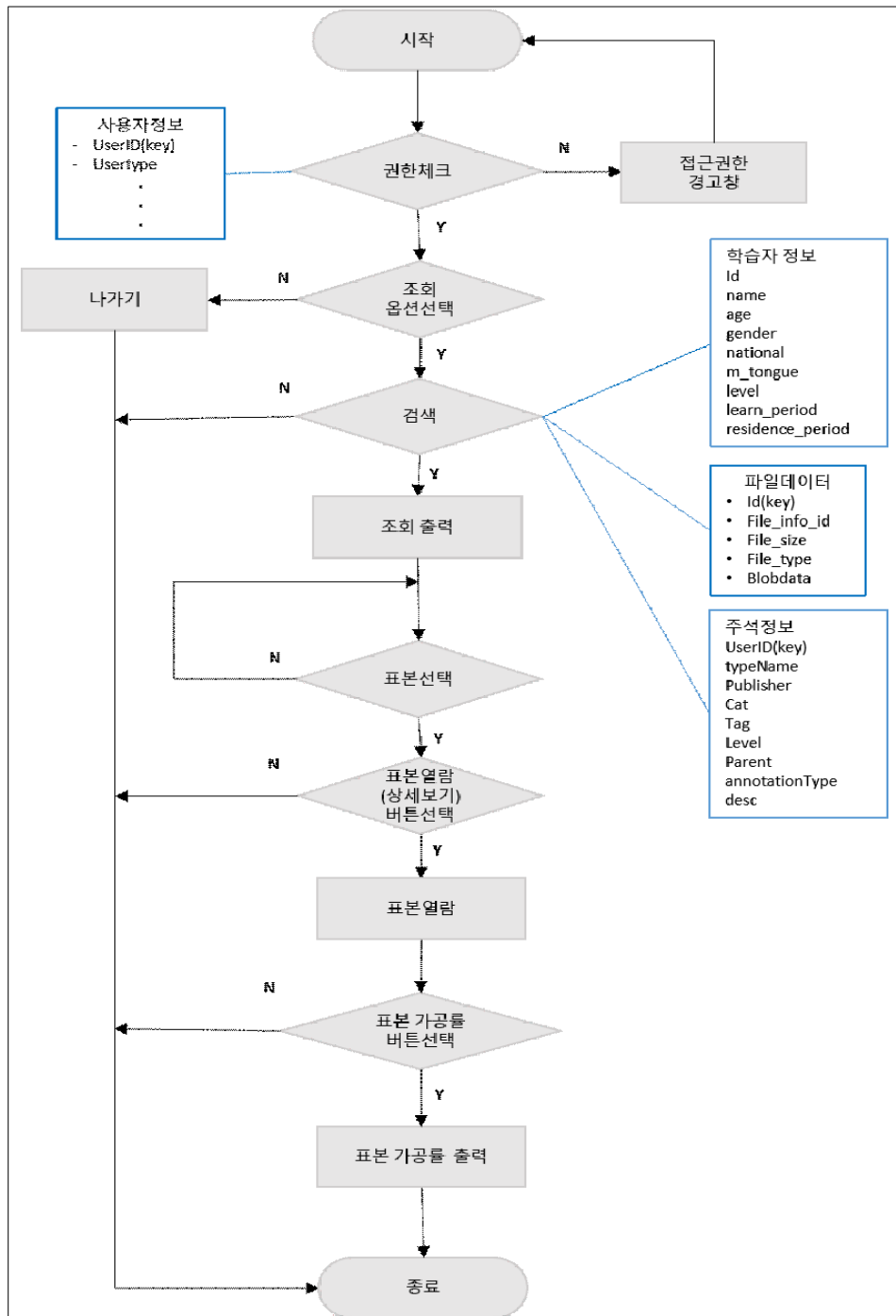
가. 화면



나. 요구 명세

ID	2.2.1.4	요구명	표본 가공률 통계 조회
내용		사용자는 표본에 대한 가공률 통계를 조회한다.	
사전조건		조회 가능한 표본이 있다.	
사후조건			
작업 흐름			
1. 사용자는 표본관리 페이지를 누른다. 2. 표본 관리 메인 페이지 상단에 기관별, 언어별 통계 그래프가 조회된다.			

다. 액티비티 다이어그램



(4)표본 수정

가. 화면

국립국어원 학습자 말뭉치 구축 지원 웹페이지

관리자

홍길동님

관리페이지

마이페이지

로그아웃

수집

관리

표본 관리

주석 관리

작업 관리

가공

분석

주석 검색

배포

관리

표본 관리

표본 수정 팝업창

표본수정

수집 된 표본파일에 대한 정보를 수정할 수 있습니다.

표본파일

찾아보기...

찾아보기...

학습자 정보

학습자 ID	1435	국적	중국
한국어 등급		직업	학생
모국어		학습기간	4개월
생년		기관	신라대학교
한국 거주기간		성별	☐ 남자 ☒ 여자
등역서 유무			

웹 페이지 메시지

?

선택한 표본을 수정하시겠습니까?

확인

취소

학습자 정보

학습환경	☐ 국내 ☐ 국외	표본유형	☐ 구어 ☐ 문어
------	---	------	---

파일명 자동 생성

파일명 생성

일반 문어 중국 신라대 1급 0020

표본 수정

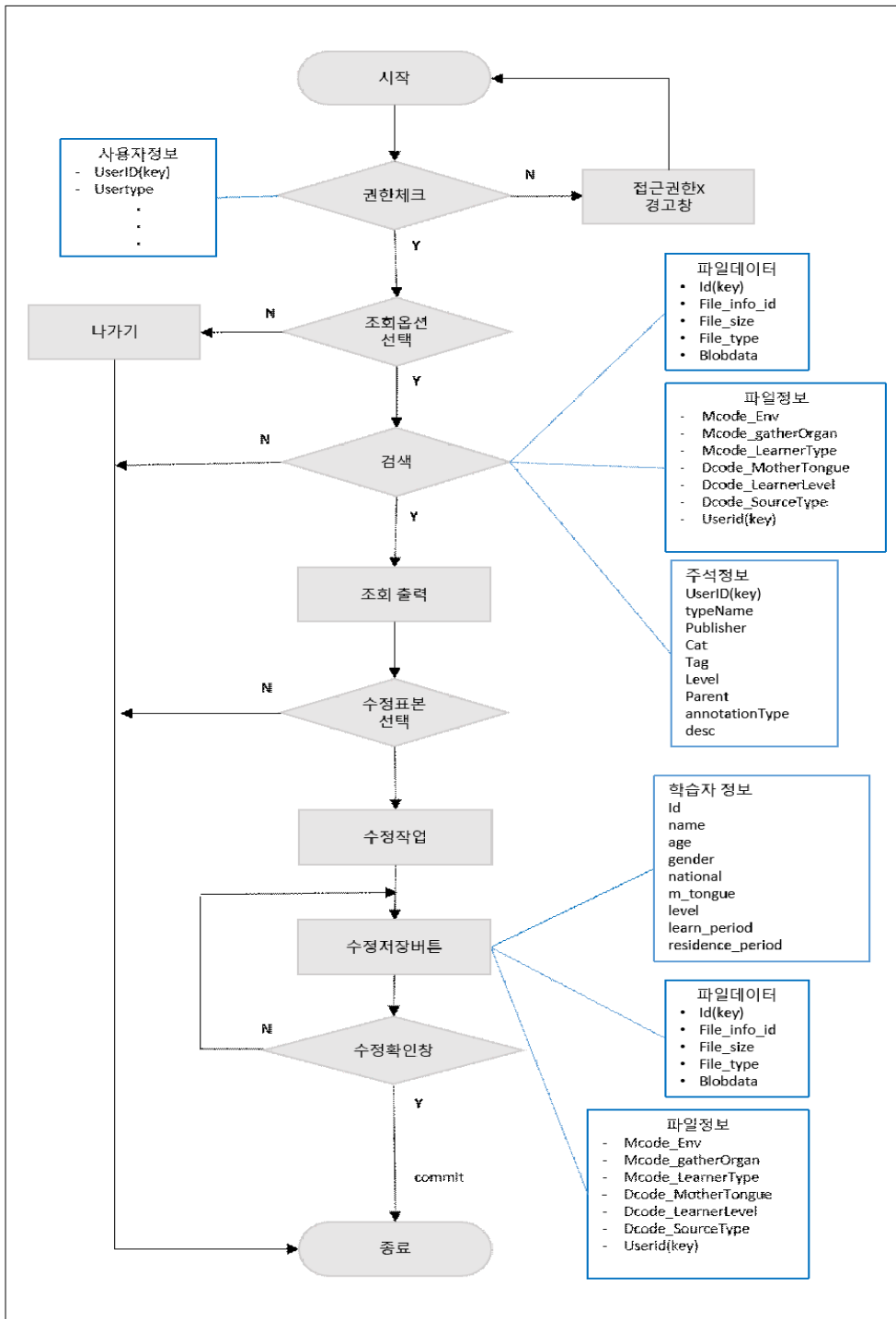
ID	파일명	학습기간	한국어등급	모국어	학습목적	파일유형

나. 요구 명세

ID	2.2.2.1	요구명	표본 수정 기능
내용		사용자는 표본의 원본 파일을 수정(변경)한다.	
사전조건			
사후조건			
작업 흐름			
1. 사용자는 표본관리 페이지를 누른다. 2. 사용자는 표본은 조회한다. 3. 사용자는 수정 할 표본을 체크한다. 4. 표본 수정 버튼을 누른다. 5. 수정 팝업창이 뜬다. 6. 사용자는 찾아보기 버튼을 눌러 원본 파일을 수정한다. 7. 사용자는 저장을 누른다. 8. 수정 확인 알림창이 뜬다. 8-1. 취소 버튼을 누르면 원래 페이지로 돌아간다. 9. 수정된 표본 파일이 저장된다.			

ID	2.2.2.2	요구명	파일정보 수정 기능
내용		사용자는 표본 파일정보를 수정(변경)한다.	
사전조건			
사후조건			
작업 흐름			
1. 사용자는 표본 관리 페이지를 누른다. 2. 사용자는 표본을 조회한다. 3. 사용자는 수정할 표본을 체크한다. 4. 표본 수정 버튼을 누른다. 5. 수정 팝업창이 뜬다. 6. 사용자는 변경할 파일 정보를 수정한다. 7. 사용자는 모두 입력 후 표본 수정 버튼을 누른다. 7-1. 입력창에 입력하지 않은 경우, 기존 정보 그대로 저장된다. 7-2. 정보를 수정하지 않은 경우, 기존 정보 그대로 저장된다. 8. 수정 확인창이 뜬다. 8-1. 취소 버튼을 누르면, 원래 페이지로 돌아간다. 9. 수정된 정보로 표본 파일이 저장된다.			

다. 액티비티 다이어그램



(5)표본 삭제

가. 화면

전체 표본조회 검색

학습자ID 선택해주세요

한국어 등급 선택

학습목적 선택

발음자음형 구어 ☐ ~

파열유형 원시표 ☐ ☐ 국외 ☐ ☐ 오류주석 ☐

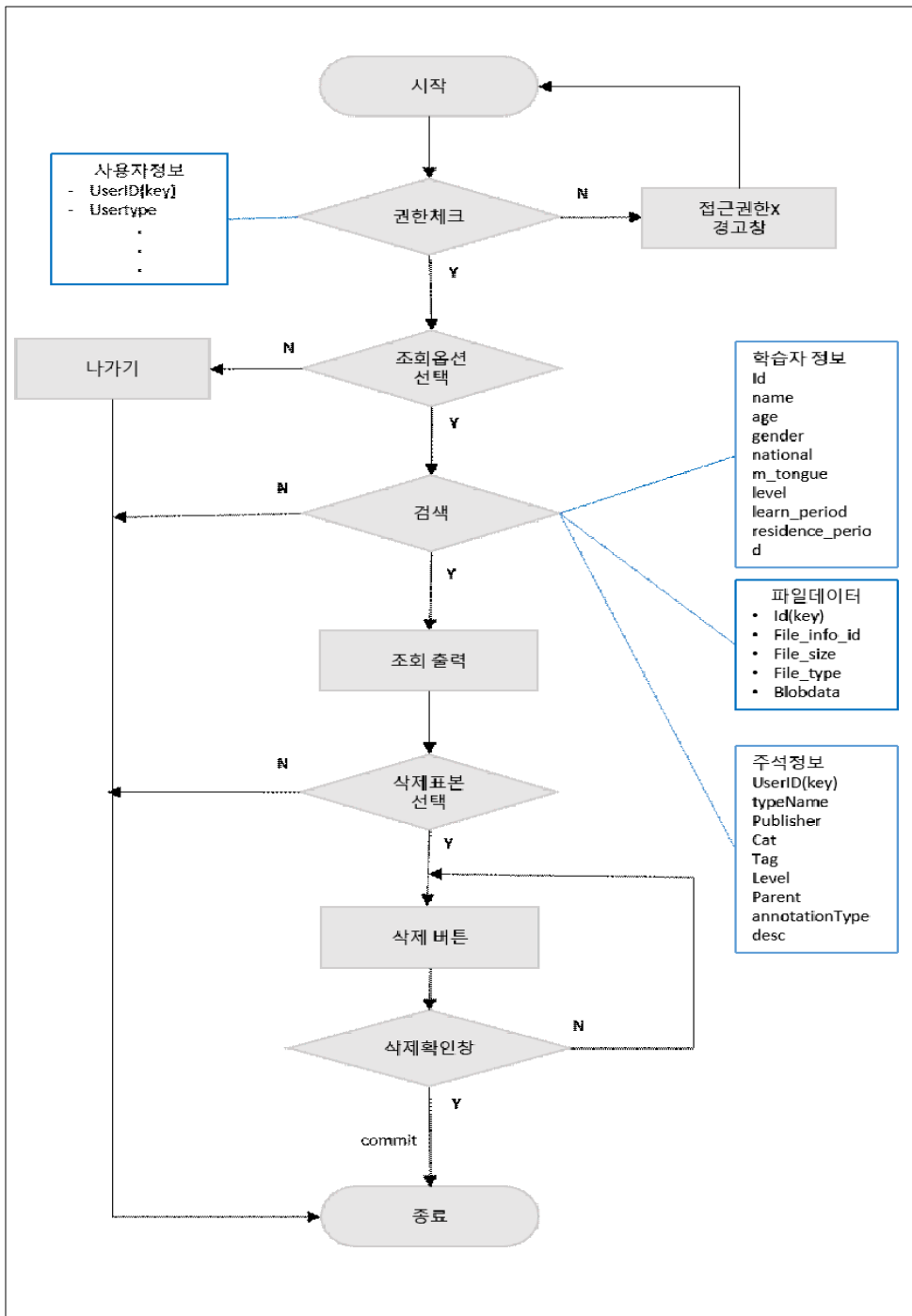
표본조회 결과

	ID	파열형	학습기관	한국어등급	오국어	학습목적	파열유형
☒	1435	일반_문어_중국_신라대_1급_0020	신라대학교	1급	중국어	전학	
☐	1436	일반_문어_중국_신라대_1급_0021	신라대학교	1급	중국어	전학	
☐	1437	일반_문어_중국_신라대_1급_0022	신라대학교	1급	중국어	전학	
☐	1438	일반_문어_중국_신라대_1급_0023	신라대학교	1급	중국어	전학	
☐	1439	일반_문어_중국_신라대_1급_0024	신라대학교	1급	중국어	전학	

나. 요구 명세

ID	2.2.3	요구명	표본파일 삭제기능
내용		사용자는 표본 파일을 삭제한다.	
사전조건			
사후조건			
작업 흐름			
1. 사용자는 표본 관리 페이지를 누른다. 2. 사용자는 표본을 조회한다. 3. 사용자는 삭제 할 표본을 체크한다. 4. 삭제 확인창이 뜬다. 4-1. 취소 버튼을 누르면, 표본이 삭제되지 않는다. 5. 확인 버튼을 누르면 표본이 삭제된다. 6. 검색 조회 목록에 해당 표본이 삭제되어 출력된다.			

다. 액티비티 다이어그램



1.2.2. 주석 관리

(1) 주석 추가

가. 화면

국립국어원 학습자 말뭉치 구축 지원 웹페이지

관리자 **홍길동님** 관리페이지

관리

주석관리 | 관리자가 주석 체계를 변경합니다

오류주석표지 오류주석표지

주석추가 주석삭제

학습자 오류 주석

- 분석 여부
 - 가능[POS]
 - 불가능[IMP]
- 오류 종류
 - 표기[OR]
 - 발음
 - 음절[PS]
 - 음운규칙[PF]
 - 어휘
 - 단일어[VSW]
 - 합성어[VWC]
 - 파생어[VDP]
 - 문법
 - 조사[GPT]
 - 어미[GE]
 - 높임[GH]
 - 시제[GT]
 - 사동[GCT]
 - 피동[GPS]
 - 부정[GN]
 - 표현문형[GF]
 - 문장[GS]
 - 담화
 - 자시[DR]
 - 접속[DC]- 오류 현상...

주석명 이름변경

목록보기 ☐ 목록닫기 ☐ 목록열기

층위 정렬 위 아래 맨위 맨아래

확인

© 2015 국립국어원

■ 오류주석표지 오류주석표지 ▼

+ 주석추가 주석삭제

- 학습자 오류 주석
 - 분석 여부
 - 가능[POS]
 - 불가능[IMP]
 - 오류 총위
 - 표기[OR]
 - 말음
 - 음절[PS]
 - 음운규칙[PF]
 - 어휘
 - 단일어[VSW]
 - 합성어[VCW]
 - 파생어[VD]
 - 문법
 - 어미[GE]
 - 활임[GH]
 - 시제[GT]
 - 사동[GC]
 - 피동[GPS]
 - 부정[GN]
 - 표현문형[GF]
 - 문장[GS]
 - 담화
 - 자시[DR]
 - 전속[DC]
 - 새주석

주석명 이름변경

목록보기 ☐ 목록닫기 ☐ 목록열기

총위 정렬 위 아래 맨위 맨아래

확인

나. 요구 명세

ID	2.3.1	요구명	주석 등록 기능
내용	사용자는 사용할 주석의 체계를 등록한다.		
사전조건			
사후조건	사용자는 '확인' 버튼을 눌러 변경 사항을 저장한다.		
작업 흐름	1. 사용자는 주석 관리 페이지를 누른다. 2. 주석을 추가할 주석 트리를 선택한다. 3. 주석 추가 버튼을 누른다. 4. 선택한 상위 주석의 하위주석으로 주석이 추가된다.		
기타	주석 추가 시, 이름은 따로 입력하여 변경한다.		

(2) 주석 수정

가. 화면

마이페이지

로그아웃

수집

관리

가공

분석

배포

> 표본 관리

> 주석 관리

> 작업 관리

> 주석 검색

주석관리 | 관리자가 주석 체계를 변경합니다

오류주석표지

오류주석표지

+주석추가

주석삭제

학습자 오류 주석

분석 여부

가능[POS]

불가능[IMP]

오류 종류

표기[OR]

말음

음절[PS]

음운규칙[PF]

어휘

단어어[VSW]

합성어[VCW]

파생어[VD]

문법

어미[GE]

높임[GH]

시제[GT]

사동[GC]

피동[GPS]

부정[GN]

표현문형[MF]

문장[GS]

담화

자시[DR]

전속[DC]

주석

주석명

이름변경

목록보기

목록닫기

목록열기

종위 정렬

위

아래

맨위

맨아래

확인

© 2015 국립국어원

(3) 주석 삭제

가. 화면

마이페이지

로그아웃

수집

관리

표본 관리

주석 관리

작업 관리

가공

분석

주석 검색

배포

주석관리

관리자가 주석 체계를 변경합니다

오류주석표지 오류주석표지

주석추가

주석삭제

학습자 오류 주석

분석 여부

가능[POS]

불가능[IMP]

오류 종류

표기[OR]

말음

음절[PS]

음운규칙[PF]

어휘

단말어[VSW]

합성어[VCW]

파생어[VDO]

문법

조사[GPT]

어미[GE]

불침[GH]

시제[GT]

사동[GC]

파동[GPS]

부정[GN]

표현문형[GF]

문장[GS]

담화

자시[DR]

접속[DC]

오류 현상

주석명

조사[GPT]

이름변경

목록보기

목록닫기

목록열기

증위 정렬

위

아래

맨위

맨아래

확인

© 2015 국립국어원

1.3. 검색

(1) 주석 유형별 검색

가. 화면

 국립국어원 학습자 말뭉치 구축 지원 웹페이지

관리자
홍길동님

관리페이지

마이페이지 | 로그아웃

수집

관리

> 표본 관리
 > 주석 관리
 > 작업 관리

가공

분석

> 주석 검색

배포

분석

[주석검색](#) | 등록 된 표본의 형태/오류주석을 검색합니다.

전체 표본 어절 검색

주석유형	형태주석	일반부사	검색어 유형	☐ 어절 ☐ 형태
검색어	벌써			

검색조회

sampleid	Left-Context	Word	Right-Context
1451	한국어 공부한 지	벌써	네 달이 되었습니다.
1471	하지만 이제	벌써	시대를 변해서 사람을 꿈꾸는 것 다르다.
1484	한국에 온지	벌써	4개월 되었다.
1484	지금 한국에 온 지가 안 되었지만	벌써	익숙한다.
1489	한국에 온 지	벌써	여섯 달었다.

나. 요구 명세

ID	5.2.1	요구명	주석 검색 기능
내용		사용자는 주석별 검색을 한다.	
사전조건			
사후조건			
작업 흐름			
1. 사용자는 분석 주석 검색 페이지를 누른다.			
2. 사용자는 검색할 주석의 유형을 선택한다.			
3. 사용자는 검색어를 입력한다.			
3-1. 검색어를 입력하지 않을 경우, 입력 알림창이 뜬다.			
4. 검색 조회 버튼을 누른다.			
4-1. 검색 조건에 해당하는 결과가 없을 경우, 출력을 하지 않는다.			
5. 검색한 조건에 따라 조회 결과가 화면에 출력된다.			

1.4. 배포

(1) 표본 검색 및 다운로드

가. 화면

① 검색

국립국어원 학습자 맞춤형 구축 지원 웹페이지

관리자 **홍길동님** [관리페이지](#)

마이페이지 로그인

배포

배포 : 등록 된 표본을 다운로드 할 수 있습니다.

전체 표본조회 검색

학습자ID: 학습기관: **선택해주세요**
 한국어 등급: **선택해주세요** 모국어: **선택해주세요**
 학습목적: **선택해주세요** 학습환경: ☐ 국내 ☐ 국외
 발표지 유형: ☐ 구어 ☐ 문어 수집기간: ~

검색조회

표본조회 결과 **다운로드**

ID	파일명	학습기관	한국어등급	모국어	학습목적	학습환경	파일유형
☐ 1435	일본_문어_중국_신라대_1급_0020	신라대학교	1급	중국어	진학		
☐ 1436	일본_문어_중국_신라대_1급_0021	신라대학교	1급	중국어	진학		
☐ 1437	일본_문어_중국_신라대_1급_0022	신라대학교	1급	중국어	진학		
☐ 1438	일본_문어_중국_신라대_1급_0023	신라대학교	1급	중국어	진학		
☐ 1439	일본_문어_중국_신라대_1급_0024	신라대학교	1급	중국어	진학		

② 내려받기

관리자 **홍길동님** [관리페이지](#)

마이페이지 로그인

배포

표본 다운로드 정보 선택

다운로드 파일 유형

☒ 원시 표본 (이미지/음성) - PDF파일 ☒ 원시 표본 (입력/전사) - TEXT파일

파일 포함 정보

정보 포함 ☒ 메타포함 ☐ 형태주석 ☐ 오류주석

다운로드

표본조회 결과 **다운로드**

ID	파일명	학습기관	한국어등급	모국어	학습목적	학습환경	파일유형
☐ 1435	일본_문어_중국_신라대_1급_0020	신라대학교	1급	중국어	진학		
☒ 1436	일본_문어_중국_신라대_1급_0021	신라대학교	1급	중국어	진학		
☐ 1437	일본_문어_중국_신라대_1급_0022	신라대학교	1급	중국어	진학		
☐ 1438	일본_문어_중국_신라대_1급_0023	신라대학교	1급	중국어	진학		
☐ 1439	일본_문어_중국_신라대_1급_0024	신라대학교	1급	중국어	진학		
☐ 1440	일본_문어_중국_신라대_1급_0025	신라대학교	1급	중국어	진학		

표본 다운로드 확인

선택한 표본을 다운로드하시겠습니까?

확인 **취소**

③ 내려받은 주석별 XML 파일 내용 (헤더, 형태주석, 오류주석을 모두 선택한 경우)

```

<?xml version="1.0" encoding="UTF-8"?>
- <Korean_Learners_Corpus>
  - <Header>
    <id>1435</id>
    <filename>일반_문어_중국_신라대_1급_0020</filename>
    <agreementFilename>Ag_일반_문어_중국_신라대_1급_0020</agreementFilename>
    <collectTime>2015 여름 학기</collectTime>
    <assignmentType>시험 작문</assignmentType>
    <assignmentGenre>생활문</assignmentGenre>
    <assignmentTheme>살고 싶은 집</assignmentTheme>
    <sentenceCount> </sentenceCount>
    <wordCount> </wordCount>
    <inputDate>2015.9.11</inputDate>
    <person_input>박은지</person_input>
    <checkDate> </checkDate>
    <person_check>김나연</person_check>
    <source_type> </source_type>
    <learn_env> </learn_env>
    <file_type> </file_type>
  </Header>
  - <BODY>
    - <T_UNIT>
      <s>저는 방 3개, 큰 거실 하나, 화장실 2개가 있는 집에 살고 싶다</s>
      - <학습자오류>
        <ERROR error_type="OM">방</ERROR>
      </학습자오류>
      - <형태분석>
        - <word>
          <w>저는</w>
          <morph pos="NP">저</morph>
          <morph pos="JX">는</morph>
          <w>방</w>
          <morph pos="NNG">방</morph>
          <w>3개,</w>
          <morph pos="SN">3</morph>
          <morph pos="NNB">개</morph>
          <morph pos="SP">,</morph>
          <w>큰</w>
          <morph pos="VA">크</morph>
          <morph pos="ETM">ㄴ</morph>
          <w>거실</w>
          <morph pos="NNG">거실</morph>
          <w>하나,</w>
          <morph pos="NR">하나</morph>
          <morph pos="SP">,</morph>
          <w>화장실</w>
          <morph pos="NNG">화장실</morph>
          <w>2개가</w>
          <morph pos="SN">2</morph>
          <morph pos="NNB">개</morph>
          <morph pos="JKS">가</morph>
          <w>있</w>
          <morph pos="VV">있</morph>
          <morph pos="ETM">는</morph>
          <w>집에</w>
          <morph pos="NNG">집</morph>
          <morph pos="JKB">에</morph>
          <w>살고</w>
          <morph pos="VV">살</morph>
          <morph pos="EC">고</morph>
          <w>싶다.</w>
          <morph pos="VX">싶</morph>
          <morph pos="EF">다</morph>
          <morph pos="SF">.</morph>
        </word>
      </형태분석>
    </T_UNIT>

```

①헤더

②오류 주석

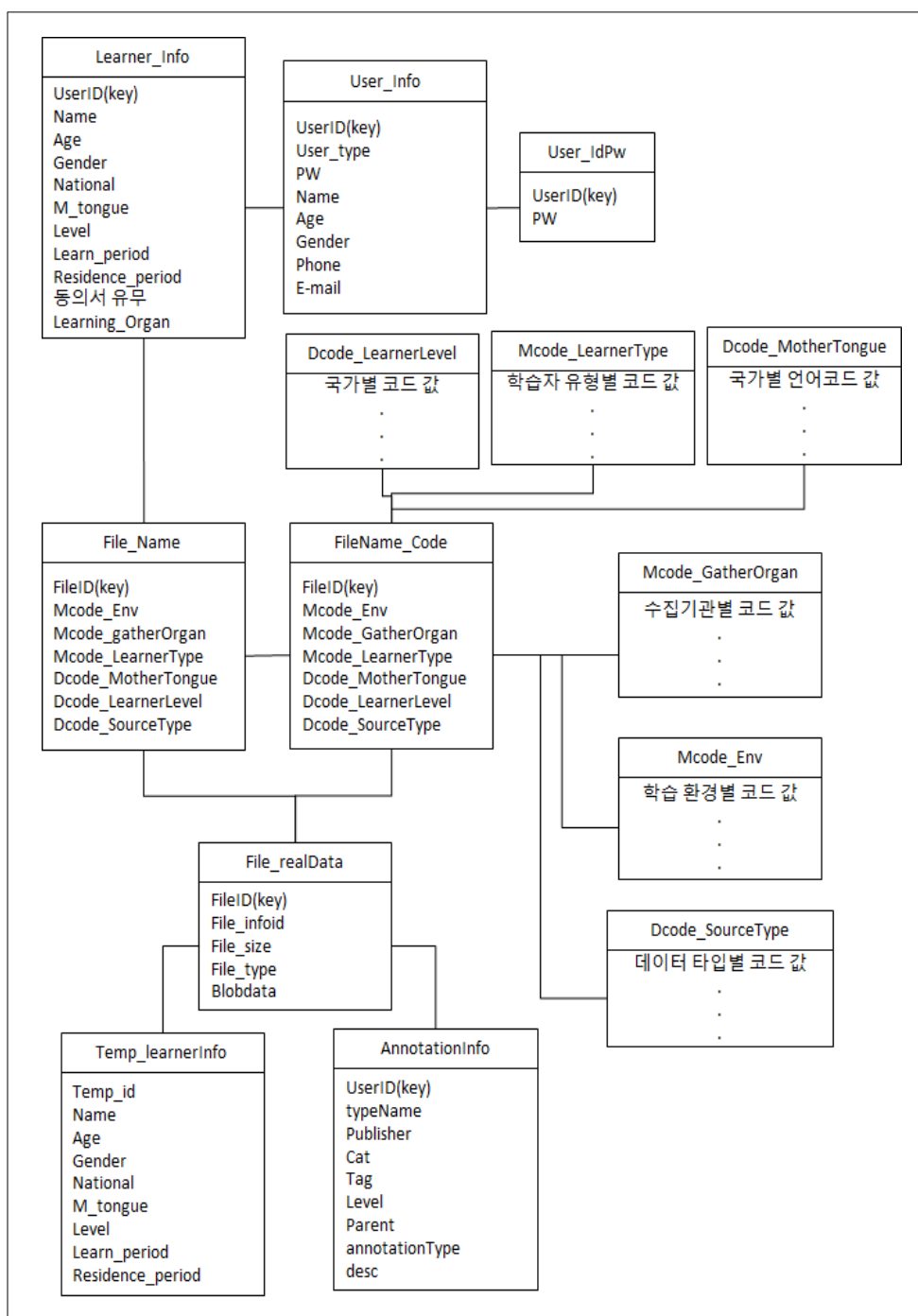
③형태 주석

나. 요구 명세

ID	4.1.1	요구명	파일별 배포 기능
내용	사용자는 등록된 표본의 파일 별로 다운로드 한다.		
사전조건			
사후조건			
작업 흐름			
1. 사용자는 배포 페이지를 누른다. 2. 사용자는 해당하는 검색 조건을 입력한다. 2-1. 검색 조건을 입력하지 않을 경우, 전체 표본이 검색 출력된다. 3. 사용자는 검색 조회 버튼을 누른다. 4. 검색 조회 결과가 화면에 출력된다. 5. 다운로드 받을 표본을 체크한다. 6. 다운로드 버튼을 클릭한다. 7. 파일이 지정된 경로로 저장된다.			

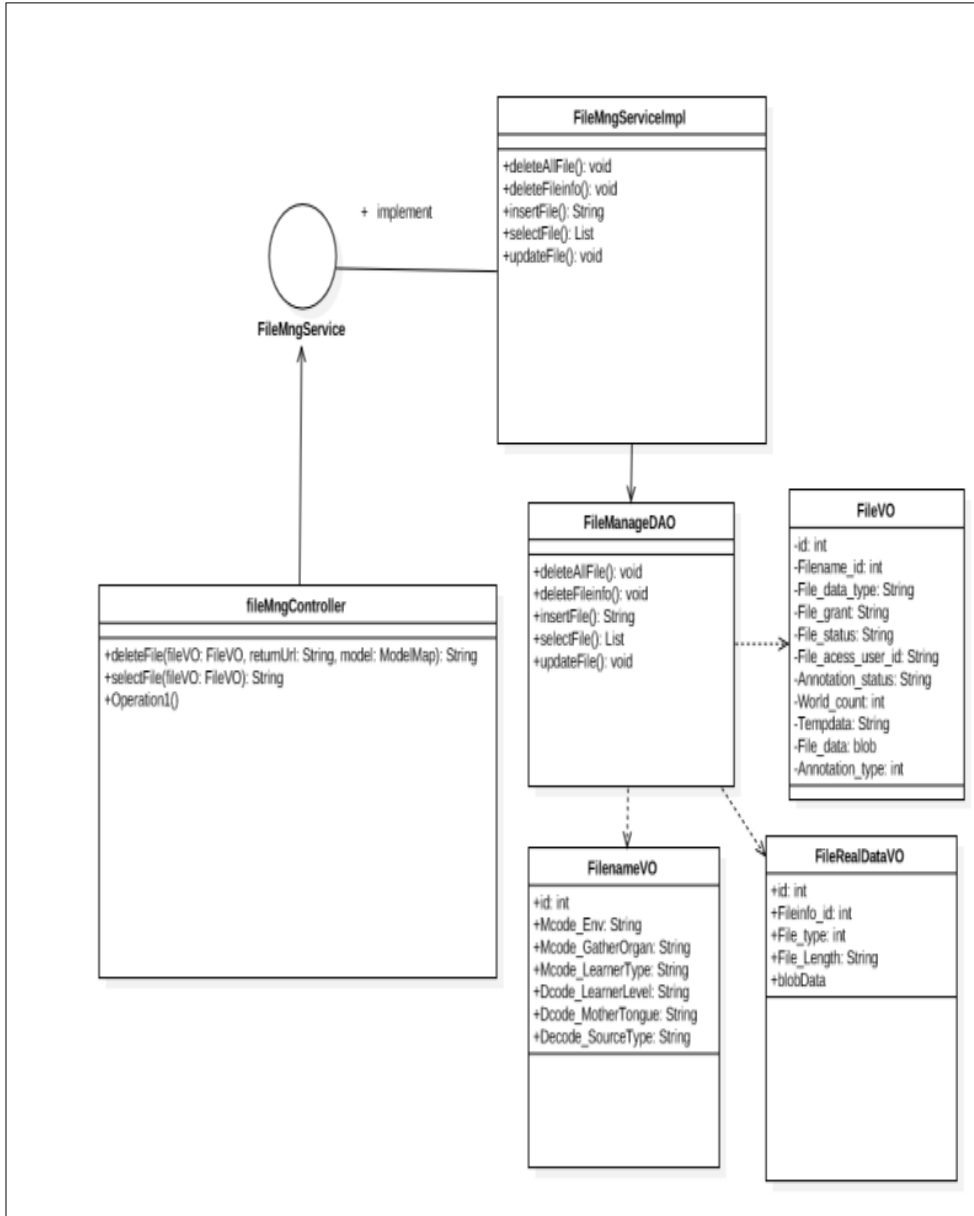
ID	4.1.1	요구명	분석 / 통계 / 검색 배포 기능
내용	사용자는 표본들의 분석/통계/검색 자료를 다운로드 한다.		
사전조건			
사후조건			
작업 흐름			
1. 사용자는 배포 페이지를 누른다. 2. 메인 페이지 상단에 통계 자료가 출력된다. 3. 다운받고자 하는 통계 자료의 우측 상단을 클릭하여 원하는 형식에 맞게 다운로드 한다. 3-1. 통계 자료 인쇄 기능도 포함한다. 4. 지정한 경로로 해당 파일이 저장된다.			
기타	통계, 검색 등은 차후에 구현 (통계 및 검색 기능, 출력용 템플릿 확정 후)		

1.5. 데이터 스키마

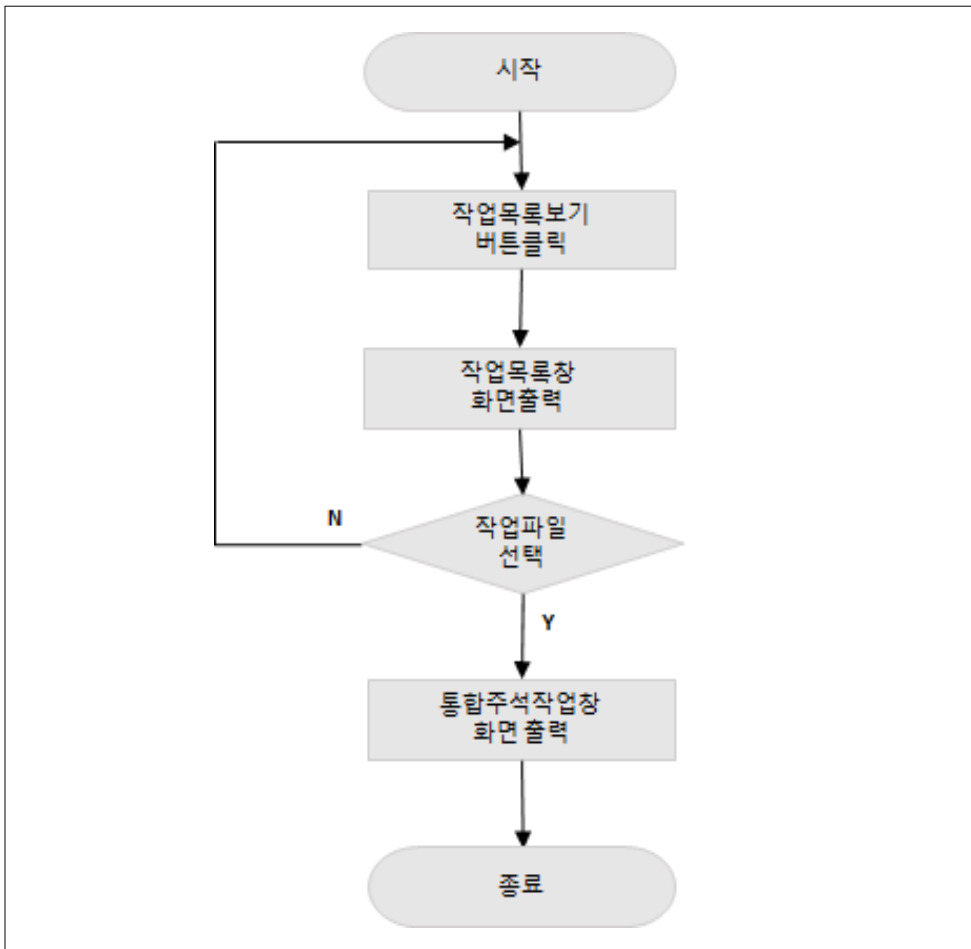


1.6. 클래스 다이어그램

<파일 관리>

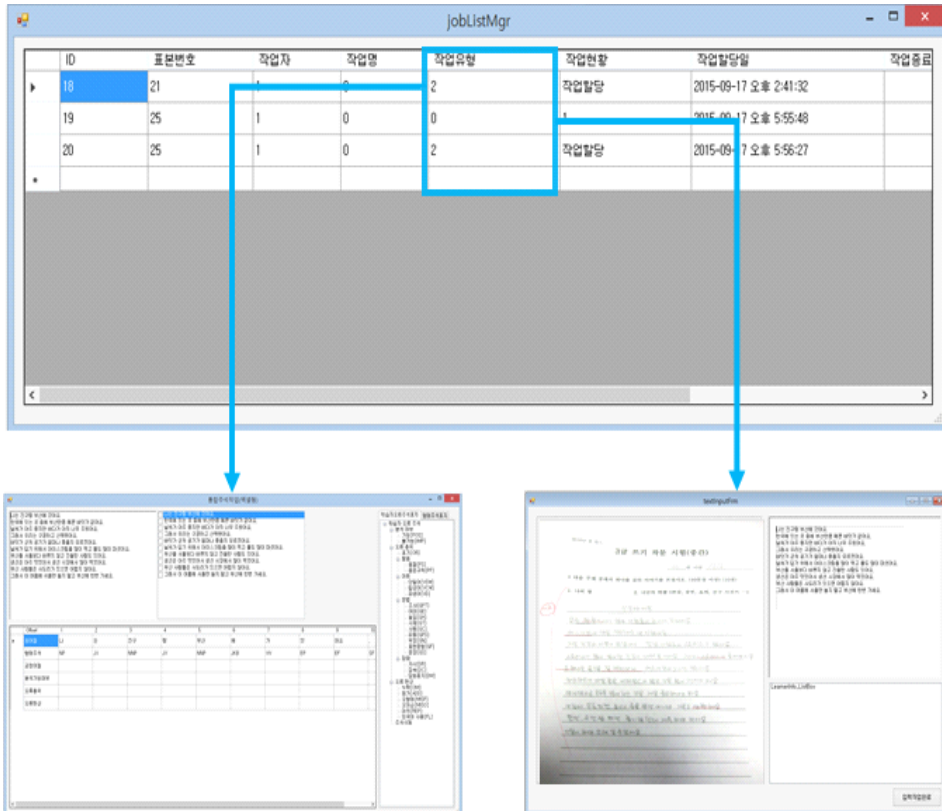


다. 액티비티 다이어그램



2.1.2. 작업 유형별 작업창 할당

가. 화면



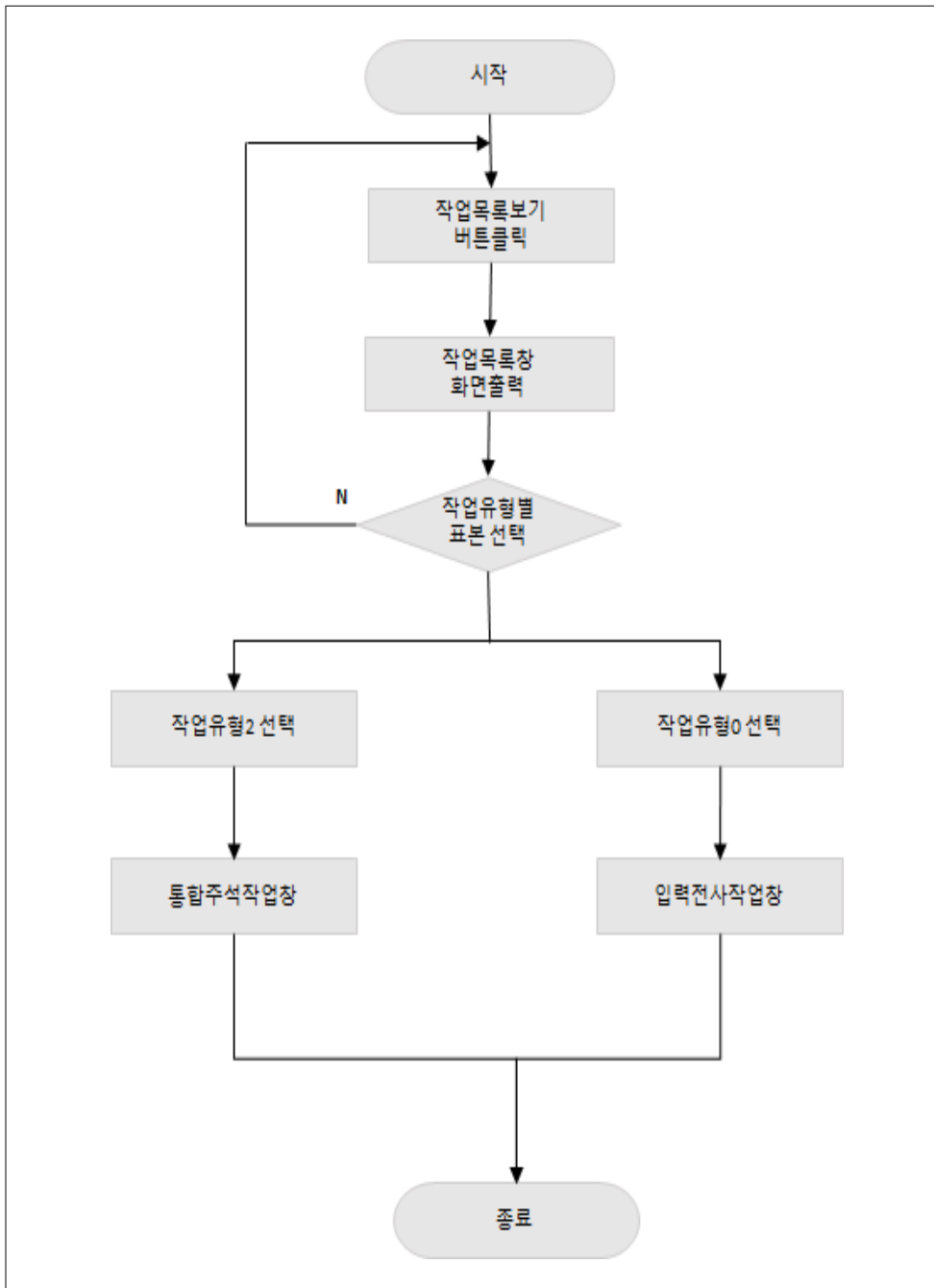
예)작업유형2:통합주석작업

작업유형0:입력 전사작업

나. 요구 명세

ID	3.1.1.2	요구명	작업 유형별 작업
내용		사용자는 표본 목록의 작업 유형에 따라 작업한다.	
사전조건		표본마다 작업 유형이 등록되어 있다.	
사후조건			
작업 흐름			
1. ‘한국어 학습자 말뭉치 관리 도구’ 메인 화면 우측에 있는 작업 목록 보기 버튼을 누른다.			
2. 파일 목록창이 뜬다.			
3. 파일 목록창에 작업이 할당된 표본 목록이 출력된다.			
4. 사용자는 작업유형을 확인한다.			
5. 사용자가 선택한 파일의 작업 유형에 따라 작업창이 뜬다.			

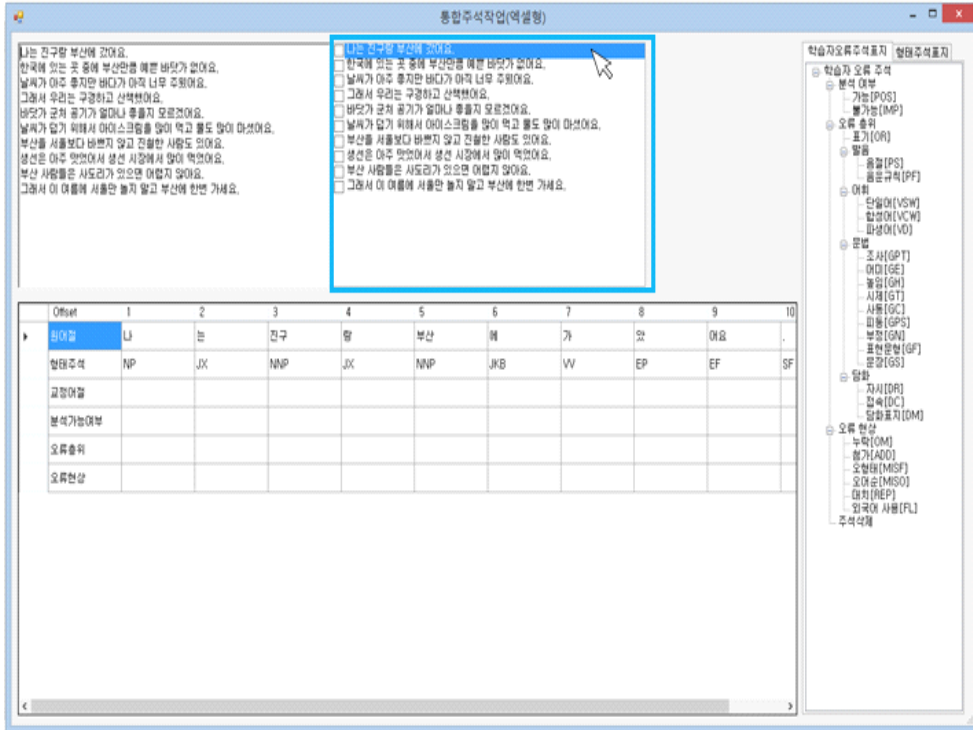
다. 액티비티 다이어그램



2.2. 주석 유형별 가공 작업

2.2.1. 오류 주석 작업 도구

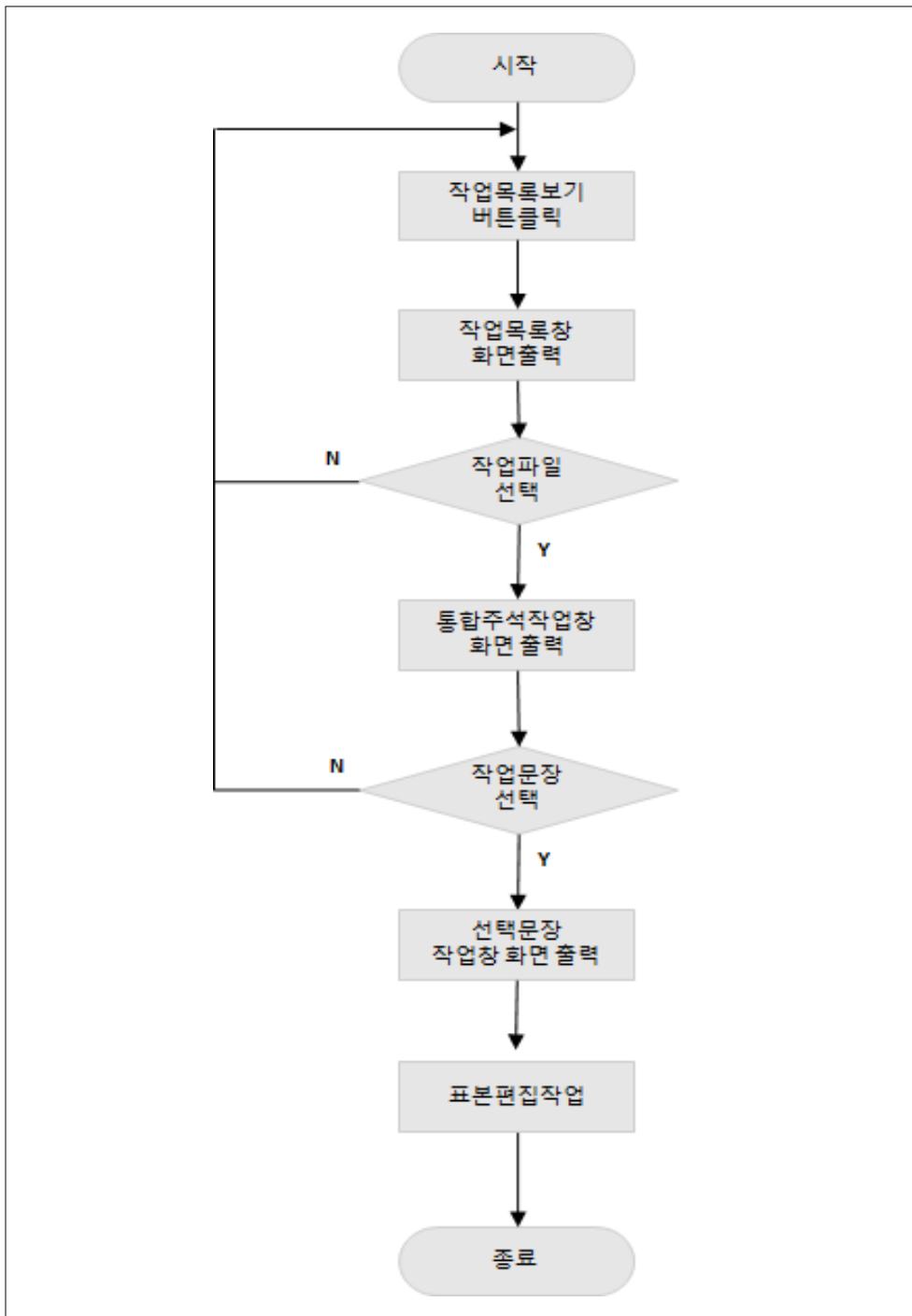
가. 화면



나. 요구 명세

ID	3.1.2.1	요구명	표본 편집
내용		사용자는 표본을 문장 단위로 편집한다.	
사전조건		표본파일에 대한 통합 주석창이 열려 있다.	
사후조건			
작업 흐름			
1. 원문 TEXT를 문장별로 보여주는 박스에서 원하는 문장을 선택한다.			
2. 주석 작업 그리드창에 선택한 문장이 출력된다.			
3. 표본 편집 작업을 한다.			

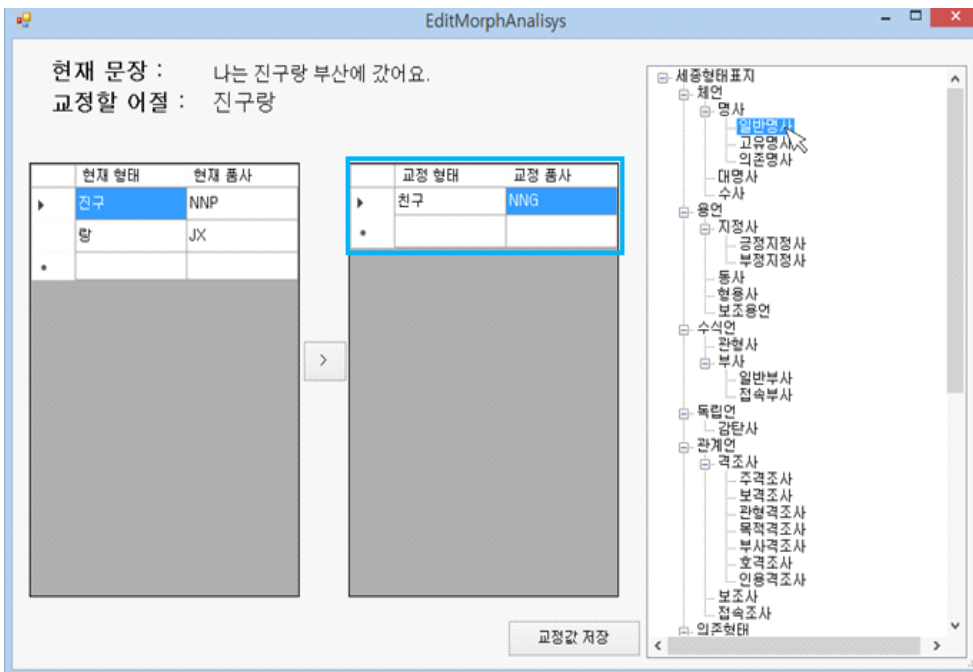
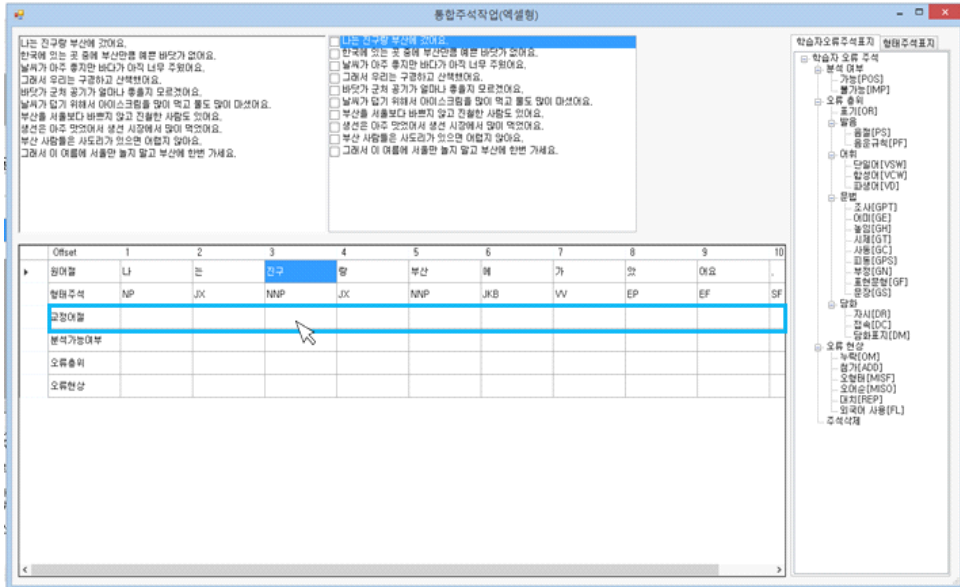
다. 액티비티 다이어그램

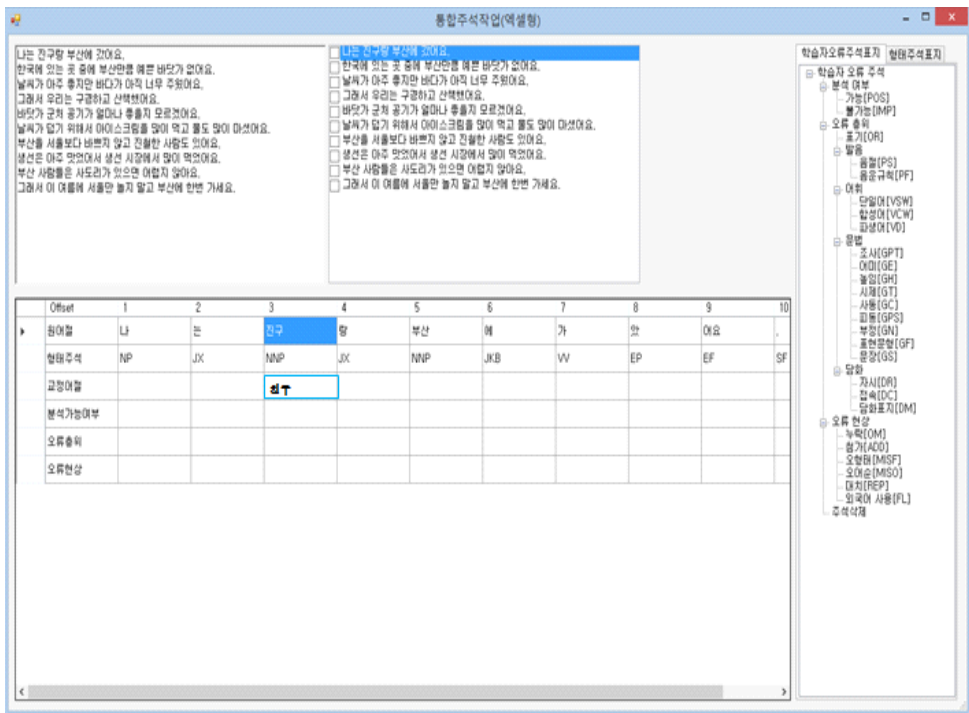


2.2.2. 오류 주석 편집 작업

(1) 표본 어절 교정

가. 화면

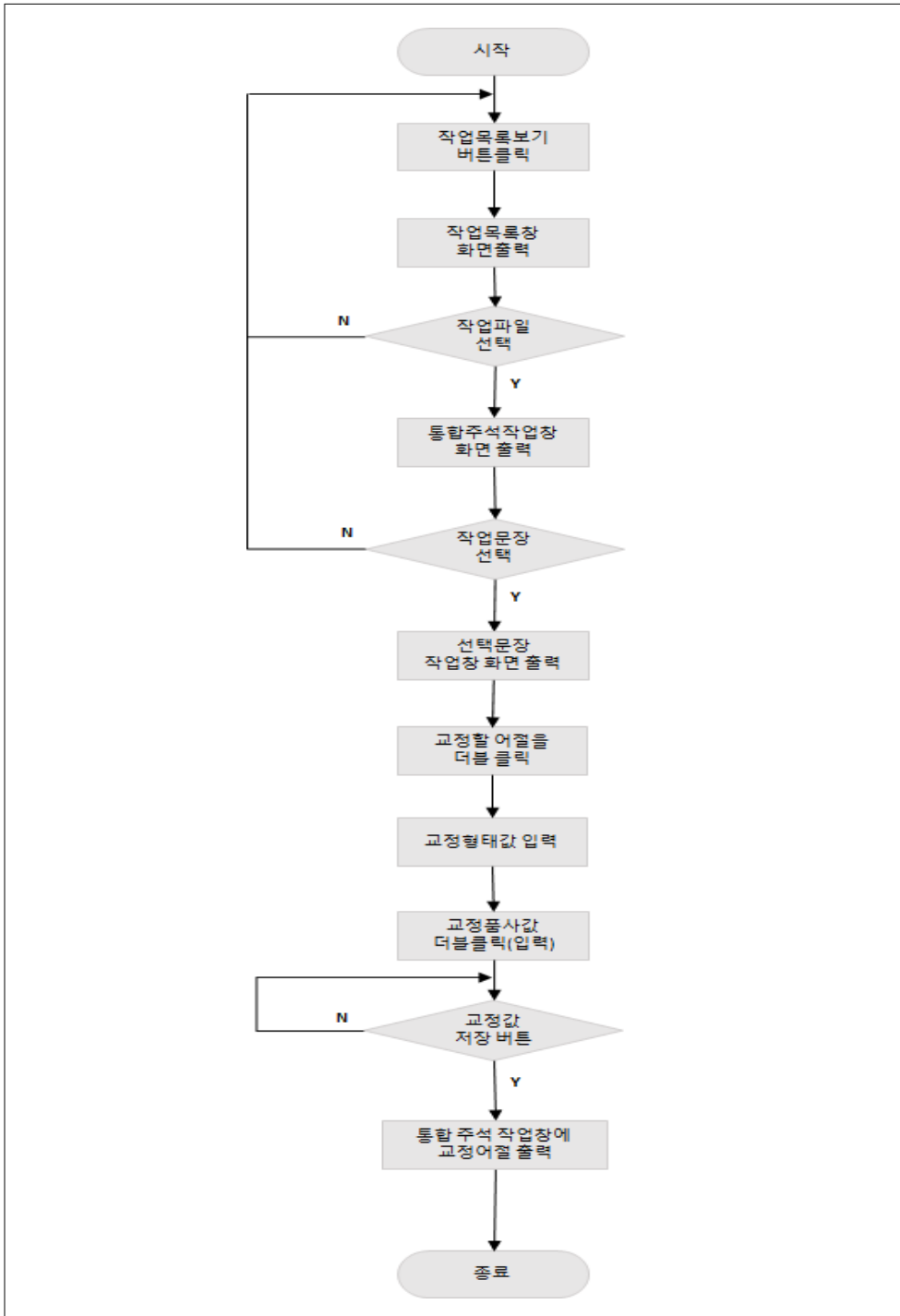




나. 요구 명세

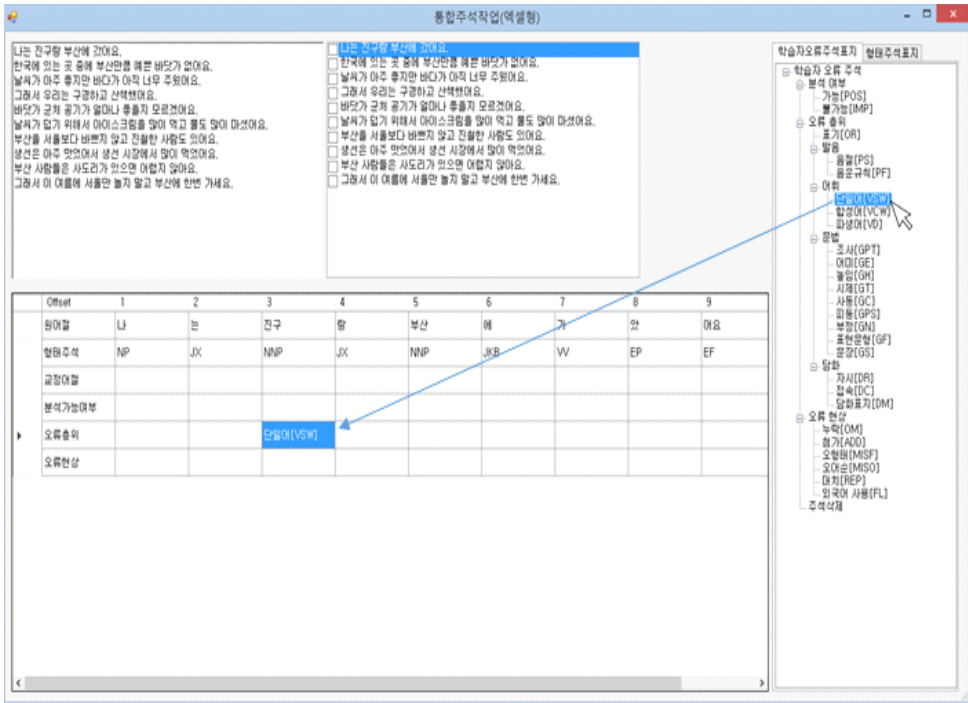
ID	3.1.2.2	요구명	어절 교정
내용		사용자는 표본의 어절 단위로 교정한다.	
사전조건		표본파일에 대한 통합 주석창이 열려 있다.	
		작업은 어절 단위를 선택하였다.	
사후조건			
작업 흐름			
1.	통합 주석 작업창에서 교정할 어절에 해당하는 셀을 더블클릭한다.		
2.	교정 작업창에서 교정형태 셀을 클릭한다.		
3.	선택된 셀에 키보드 입력으로 형태를 교정한다.		
4.	교정 품사 셀을 클릭한다.		
5.	우측에 있는 '세종형태표지'에 있는 품사를 더블클릭하여 해당 품사를 교정한다.		
6.	어절에 대한 교정이 완료되면 교정값 저장 버튼을 누른다.		
7.	통합 주석 작업창에 교정어절 셀에 교정값이 삽입된다.		

다. 액티비티 다이어그램



(2) 주석 신규 등록

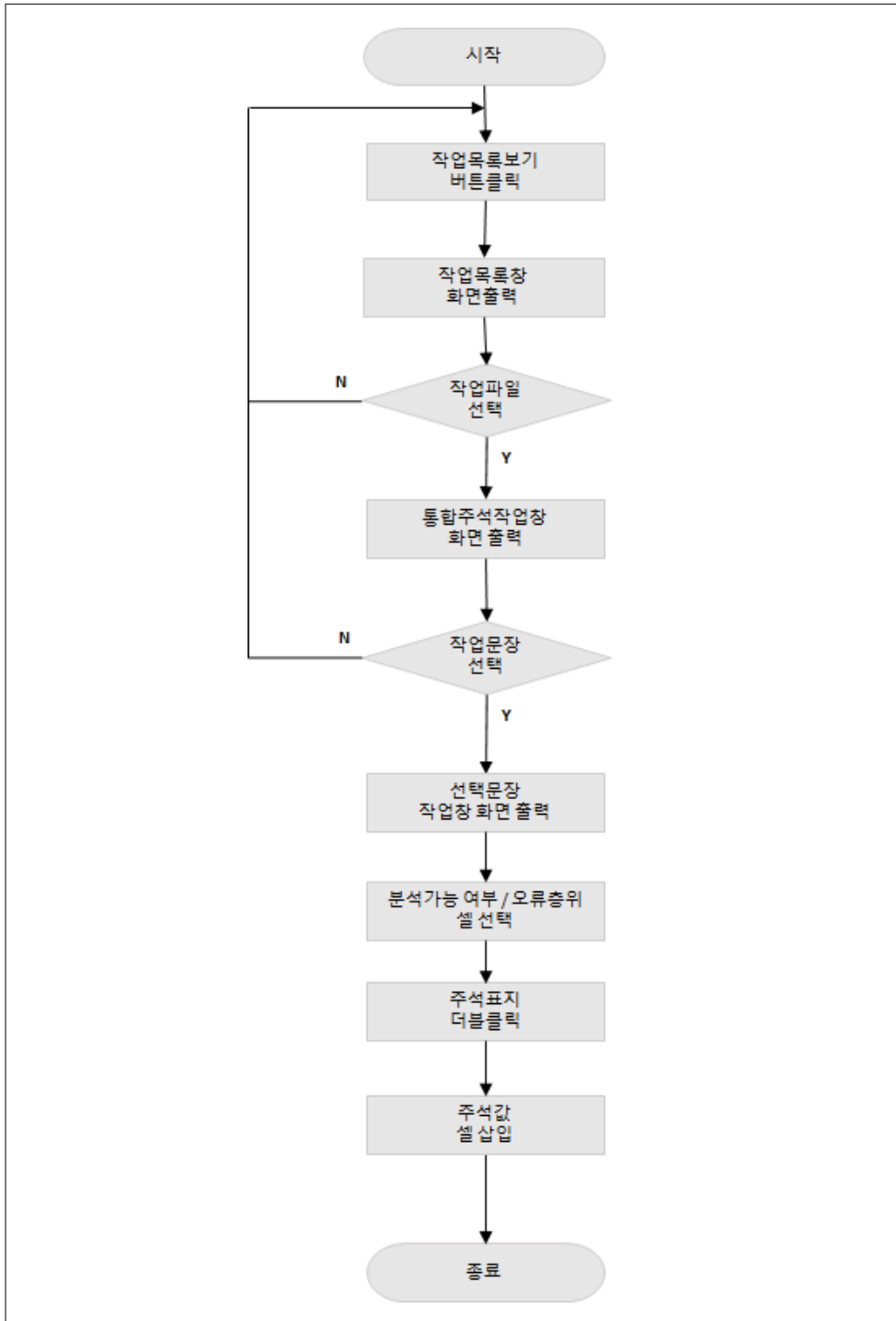
가. 화면



나. 요구 명세

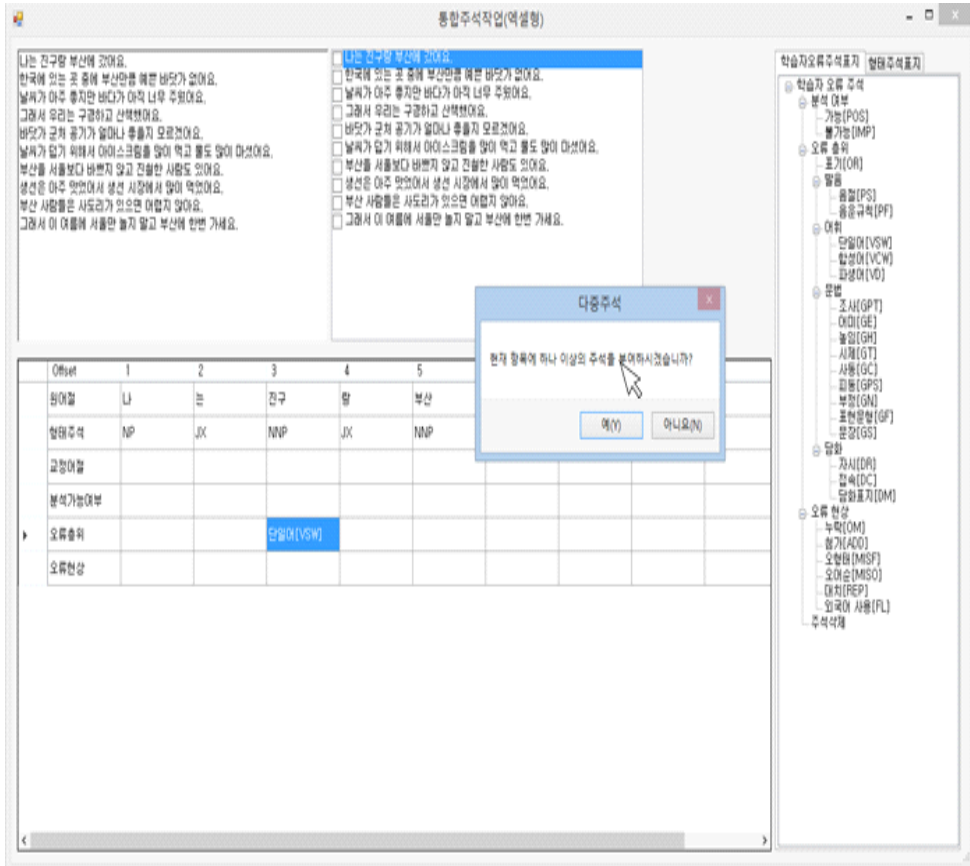
ID	3.1.2.3	요구명	주석 신규 등록
내용		사용자는 주석을 신규로 부가한다.	
사전조건		주석 작업을 할 표본 문장을 선택하였다.	
사후조건			
작업 흐름			
1. 통합 주석 작업창에서 작업할 분석 가능 여부, 오류 층위의 셀을 클릭한다. 2. 우측에 있는 학습자 오류 주석 표지에서 해당 주석을 더블클릭한다. 3. 선택한 주석값이 셀에 삽입된다.			

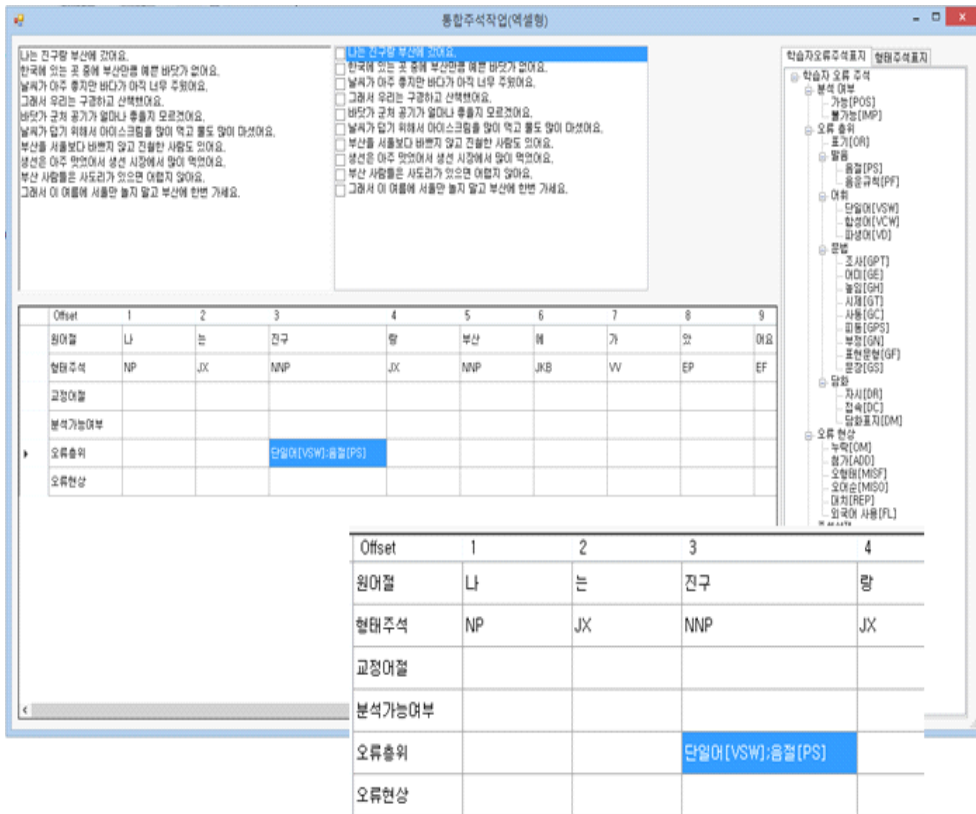
다. 액티비티 다이어그램



(3) 다중 주석 부가

가. 화면



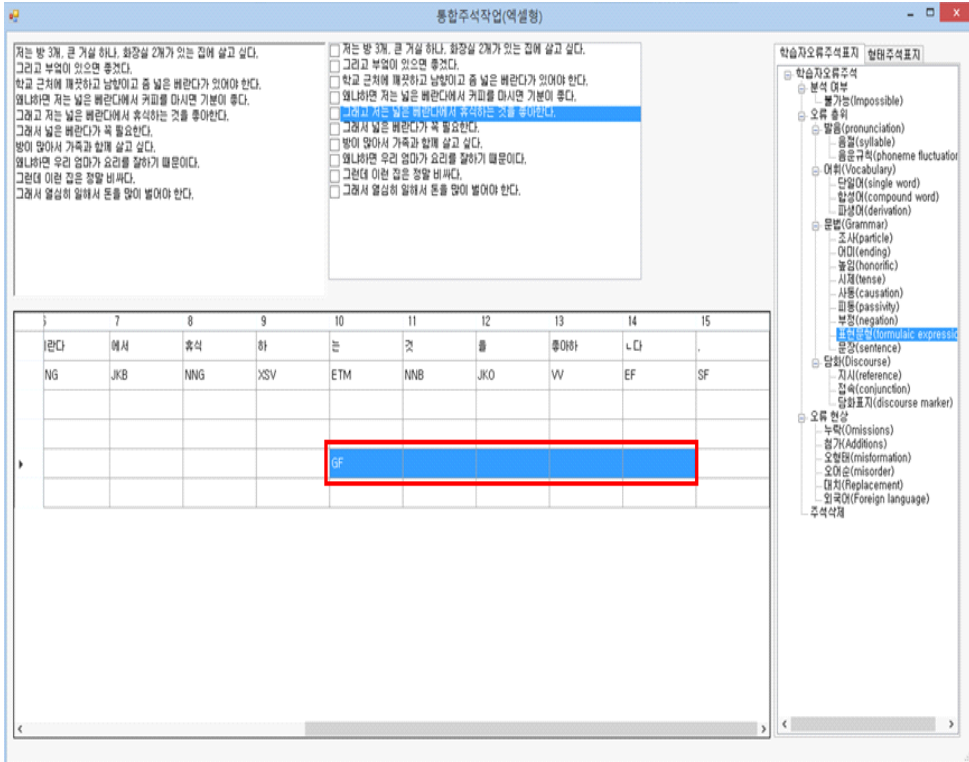


나. 요구 명세

ID	3.1.2.4	요구명	다중 주석 부가
내용		사용자는 다중 주석을 부가한다.	
사전조건		해당 셀에 주석 표지가 신규로 부가되었다.	
사후조건			
작업 흐름			
1. 주석값이 들어있는 셀을 클릭한다. 2. 우측의 학습자 오류 주석 표지에서 다른 주석값을 더블클릭한다. 3. 다중 주석 알림창이 뜬다. 4. 다중 주석 확인 버튼(예)을 누른다. 5. 해당 셀에 추가한 주석값이 삽입된다.			

(4) 다층위 주석 부가

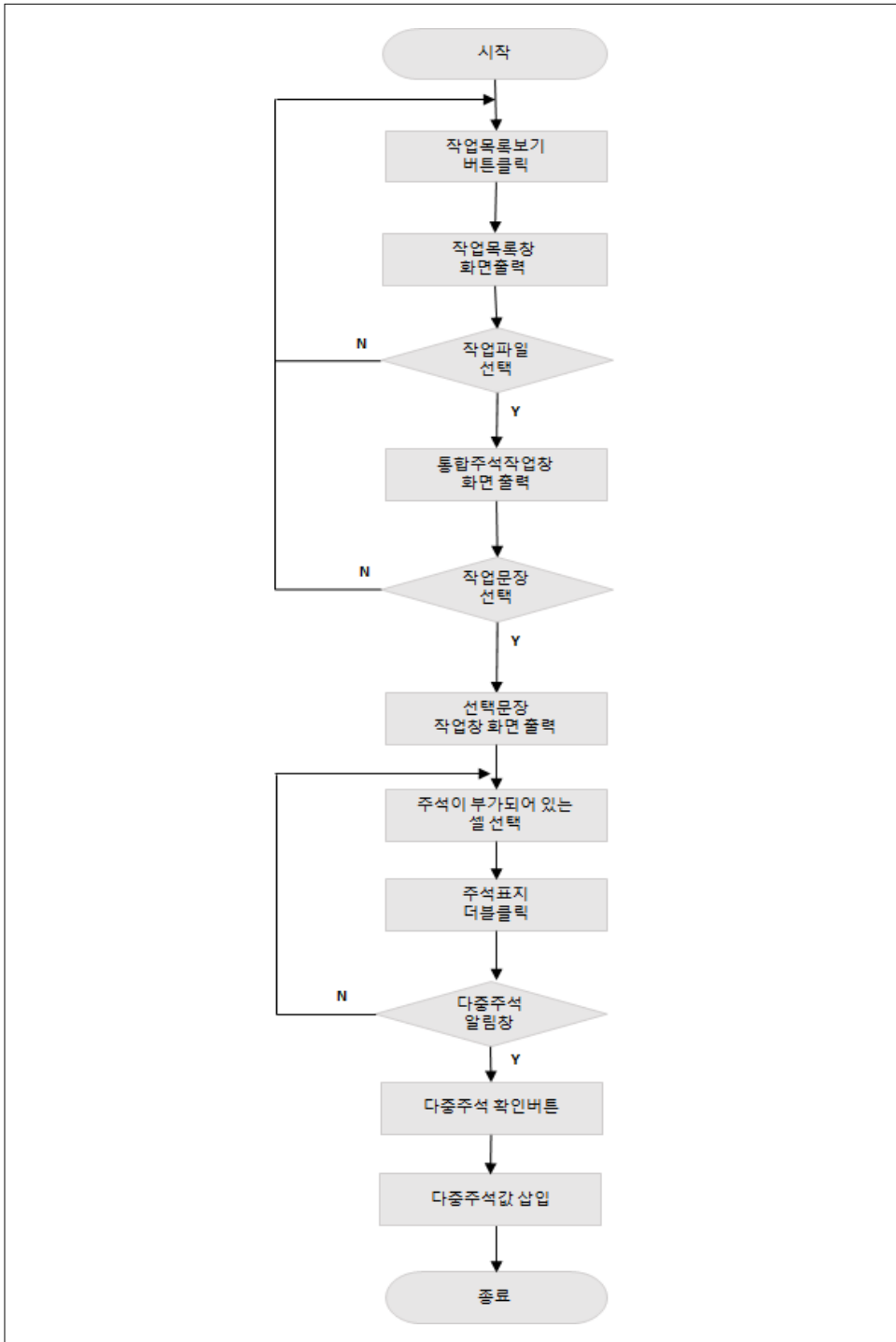
가. 화면



나. 요구 명세

ID	3.1.2.4	요구명	다층위 단위 주석 부가
내용		사용자는 다층위 단위로 주석을 부가한다.	
사전조건			
사후조건			
작업 흐름			
1. 주석을 부가할 셀을 모두 드래그하여 선택한다. 2. 우측의 학습자 오류 주석 표지에서 주석값을 더블클릭한다. 3. 해당 셀에 추가한 주석값이 삽입된다.			

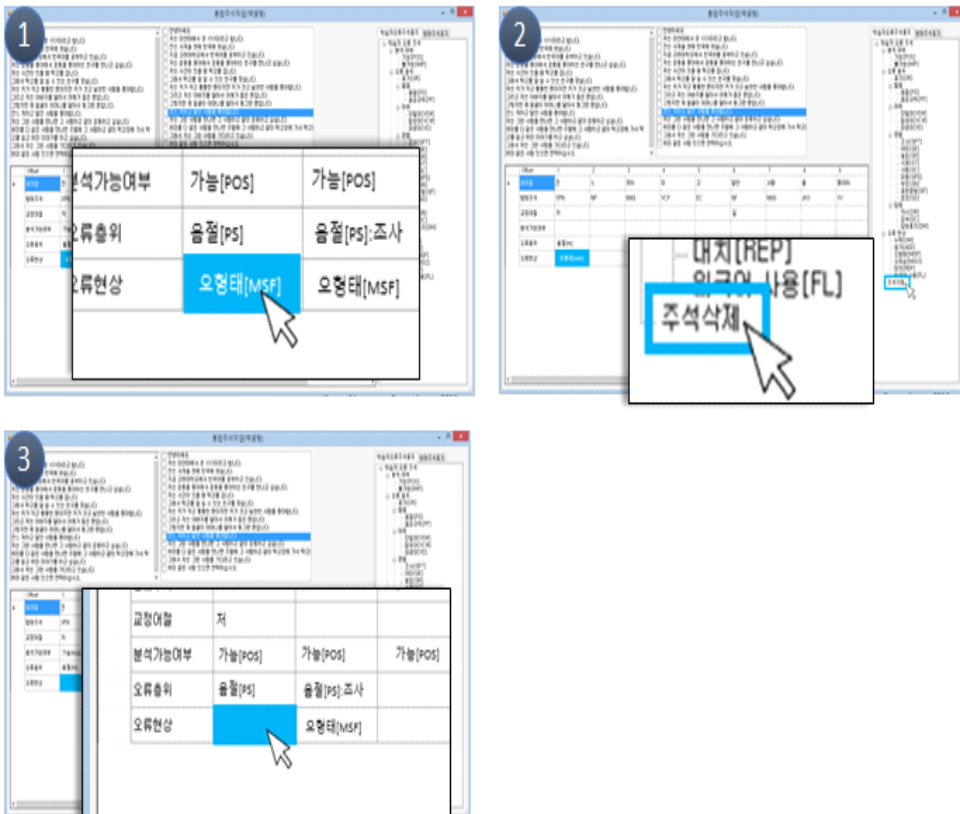
다. 액티비티 다이어그램



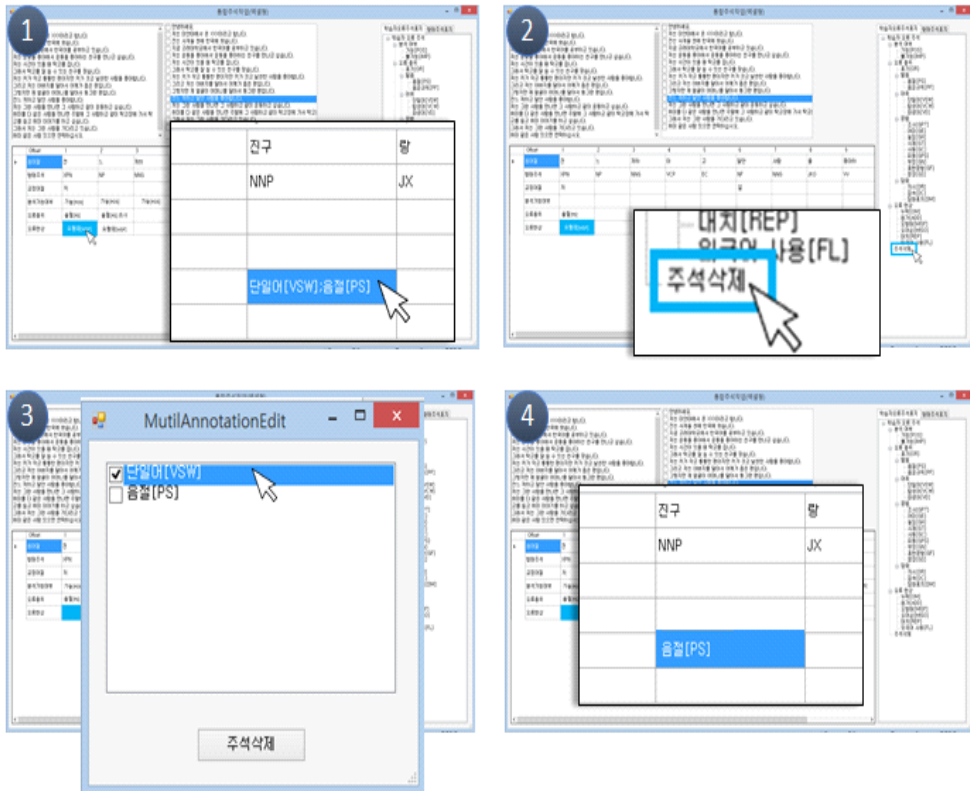
(4) 주석 삭제

가. 화면

<기본 삭제 화면>



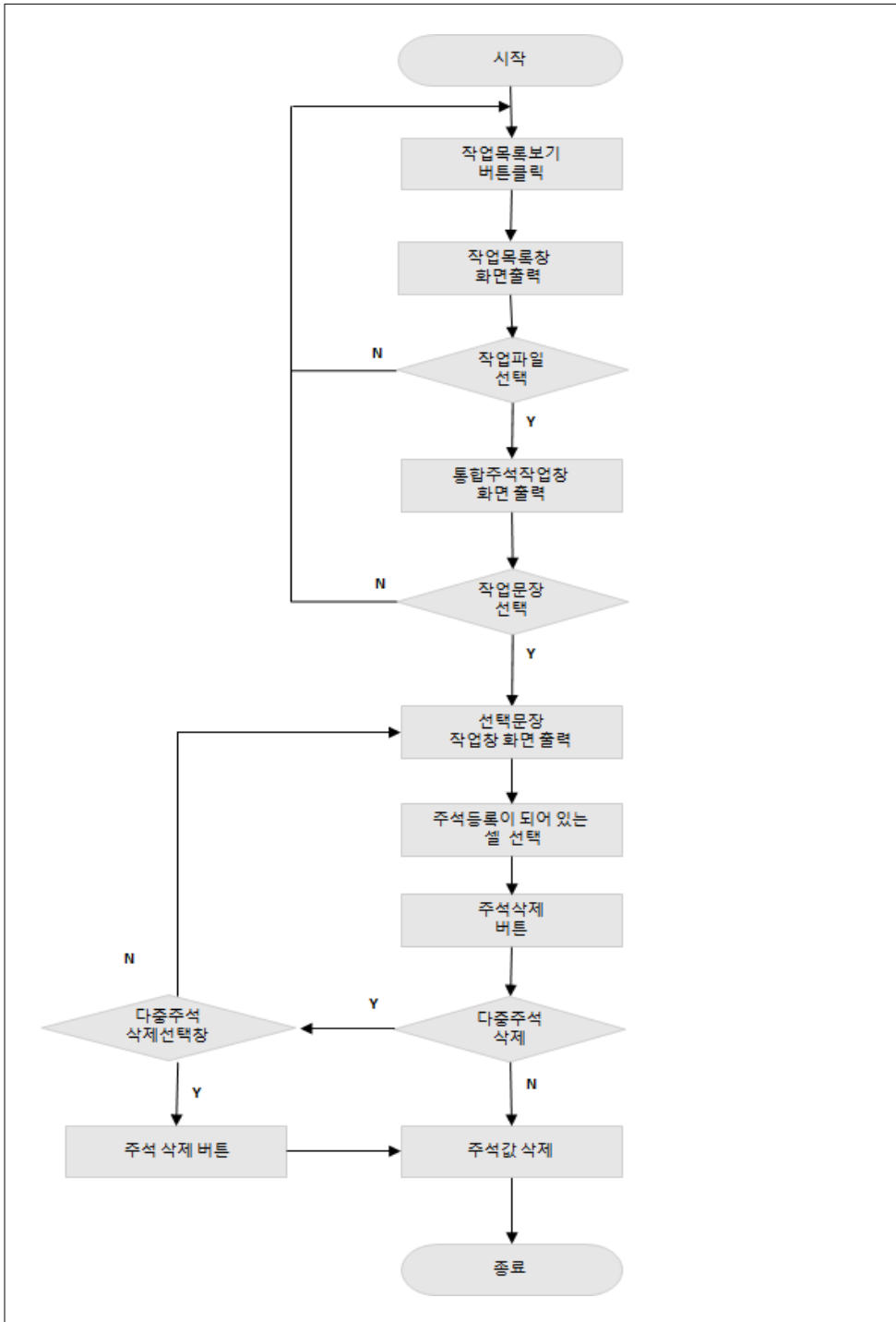
<다중 주석 삭제 화면>



나. 요구 명세

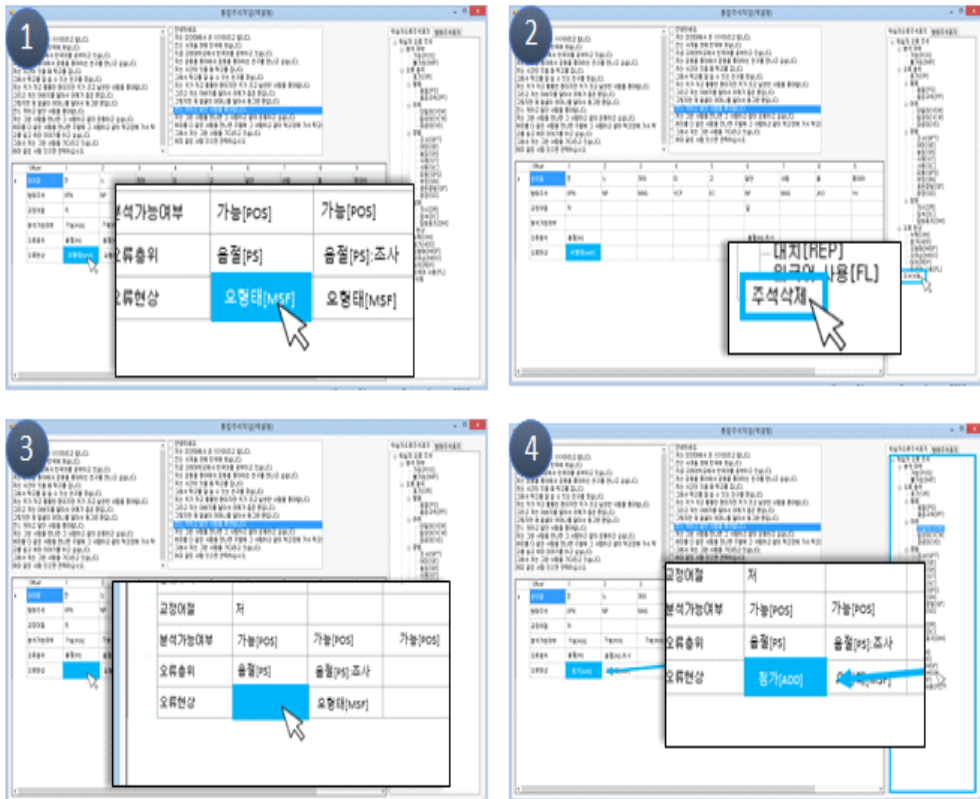
ID	3.1.2.5	요구명	주석 삭제
내용	사용자는 부가된 주석을 삭제한다.		
사전조건	해당 셀에 주석 표시가 부가되어 있다.		
사후조건			
작업 흐름			
1. 주석값이 부가되어 있는 셀을 클릭한다. 2. 우측의 학습자 오류 주석 표지에서 주석 삭제를 더블클릭한다. 2.1 다중 주석값이 부가되어 있는 셀의 경우, 삭제 주석값 선택창이 뜬다. 2.2 다중 주석값 중 삭제할 주석값을 선택한다. 2.3 주석 삭제 버튼을 누른다. 3. 셀에 있는 주석값이 삭제된다.			

다. 액티비티 다이어그램



(5) 주석 수정

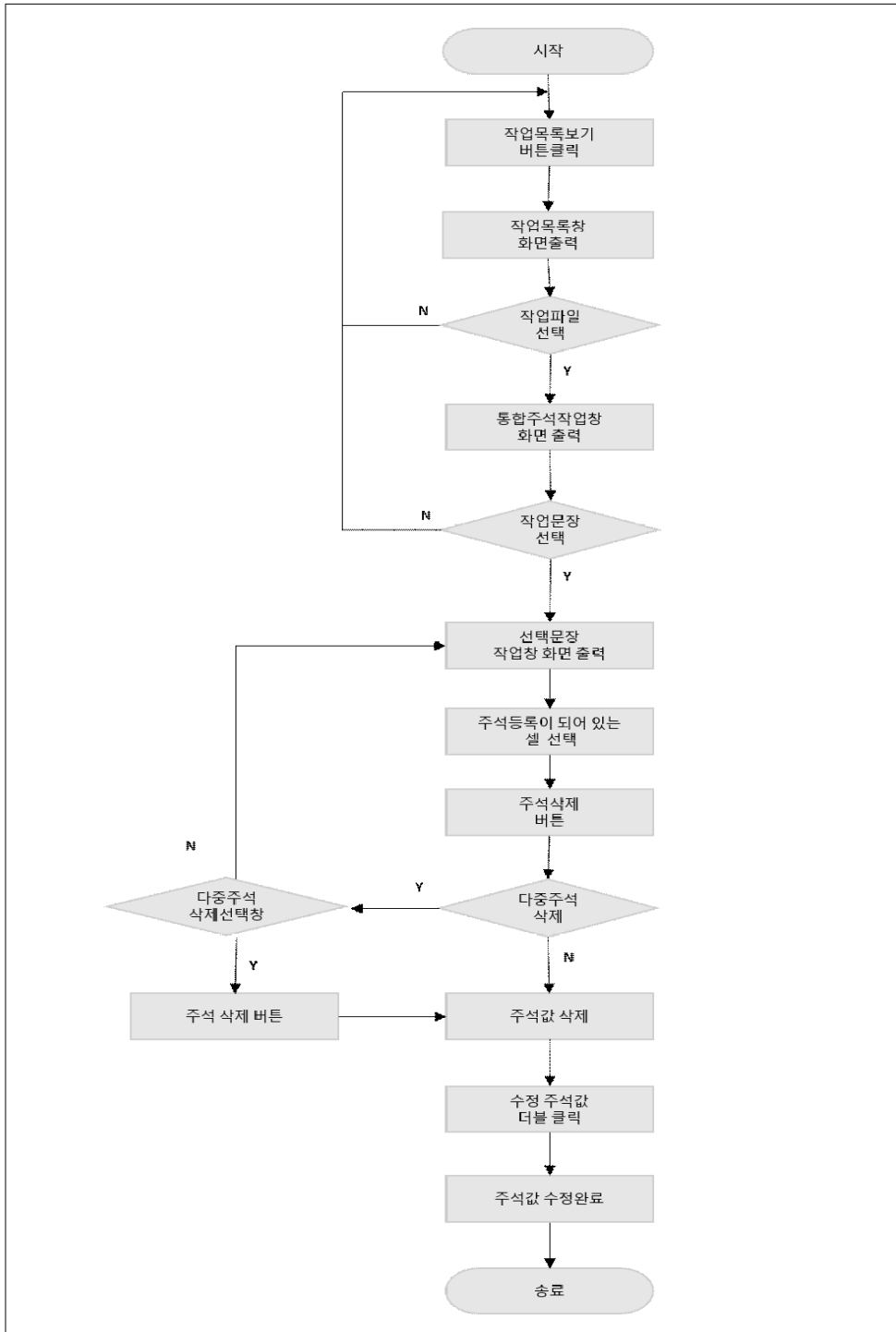
가. 화면



나. 요구 명세

ID	3.1.2.6	요구명	주석 수정
내용	사용자는 부가된 주석을 삭제 후 수정한다.		
사전조건	해당 셀에 주석 표시가 부가되어 있다.		
사후조건			
작업 흐름			
1. 주석값이 부가되어 있는 셀을 클릭한다. 2. 우측의 학습자 오류 주석 표시에서 주석 삭제를 더블클릭한다. 3. 셀에 있는 주석값이 삭제된다. 4. 셀을 다시 클릭한다. 5. 수정할 주석값을 더블클릭한다. 6. 수정한 주석값이 셀에 삽입된다.			

다. 액티비티 다이어그램



(6) 주석 작업 상태 확인

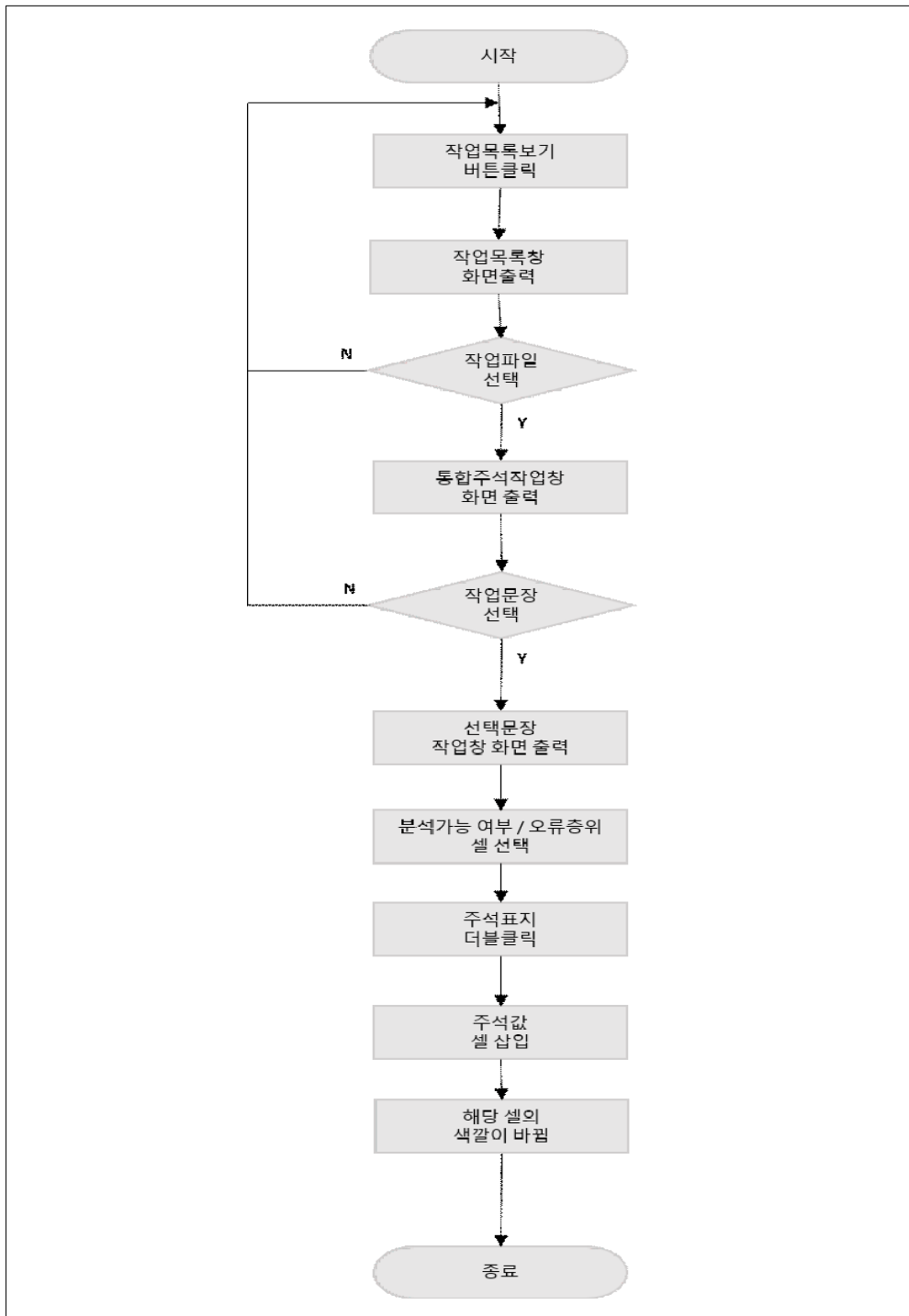
가. 화면

	Offset	1	2	3	4	5	6
	원어절	나	는	진구	랑	부산	에
	형태주석	NP	JX	NNP	JX	NNP	JKB
	교정어절						
▶	분석가능여부			가능[POS]			
	오류종위			단일어[VSW];...			
	오류현상			오형태[MISF]			

나. 요구 명세

ID	3.1.2.7	요구명	주석 작업 상태 확인
내용	사용자는 변경된 주석 상태를 확인한다.		
사전조건	해당 표본 파일에 주석 작업을 하였다.		
사후조건			
작업 흐름	- 셀마다 해당하는 주석 작업이 완료되면 셀의 색이 변한다. - 작업을 하지 않은 경우, 셀의 기본 색은 흰색이다. - 셀의 색이 변하는 것을 통해 자신의 주석 작업 상태를 확인한다.		

다. 액티비티 다이어그램



IV. 한국어 학습자 말뭉치 활용 방안

1. 공개/상용 소프트웨어를 이용한 활용 방안

1.1. 학습자 말뭉치의 구조

현재 학습자 말뭉치는 다음과 같은 모습으로 되어 있다.

전체 번호	파일명	원어절	형태	주석	형태	주석	교정	어절	교정	어절	분석	오류	유형	현상	원어절	포함	문장	파일명
1	1	요즘	요즘/NNG												[요즘]인터넷폭력로인한사회문제가날마다늘어나고있다.			일반. 문어. 대만. 경희대. 6급. 0003
2	2	인터넷	인터넷/NNG												요즘[인터넷]폭력로인한사회문제가날마다늘어나고있다.			일반. 문어. 대만. 경희대. 6급. 0003
3	3	폭력	폭력/NNG												요즘인터넷[폭력]로인한사회문제가날마다늘어나고있다.			일반. 문어. 대만. 경희대. 6급. 0003
4	4	로	로/JKB												요즘인터넷[폭력]로인한사회문제가날마다늘어나고있다.			일반. 문어. 대만. 경희대. 6급. 0003
5	5	인한	인한/VV												요즘인터넷폭력로인한사회문제가날마다늘어나고있다.			일반. 문어. 대만. 경희대. 6급. 0003
6	6	인한	인한/VV												요즘인터넷폭력로인한사회문제가날마다늘어나고있다.			일반. 문어. 대만. 경희대. 6급. 0003
7	7	사회	사회/NNG												요즘인터넷폭력로인한[사회]문제가날마다늘어나고있다.			일반. 문어. 대만. 경희대. 6급. 0003
8	8	문제	문제/NNG												요즘인터넷폭력로인한[사회]문제가날마다늘어나고있다.			일반. 문어. 대만. 경희대. 6급. 0003
9	9	가	가/JKS												요즘인터넷폭력로인한[사회]문제가날마다늘어나고있다.			일반. 문어. 대만. 경희대. 6급. 0003
10	10	날마다	날/NNG												요즘인터넷폭력로인한사회문제가날마다늘어나고있다.			일반. 문어. 대만. 경희대. 6급. 0003
11	11	날마다	마다/JX												요즘인터넷폭력로인한사회문제가날마다늘어나고있다.			일반. 문어. 대만. 경희대. 6급. 0003
12	12	늘어나고	늘어나/VV												요즘인터넷폭력로인한사회문제가날마다늘어나고있다.			일반. 문어. 대만. 경희대. 6급. 0003
13	13	늘어나고	고/EC												요즘인터넷폭력로인한사회문제가날마다늘어나고있다.			일반. 문어. 대만. 경희대. 6급. 0003
14	14	있다	있/VX												요즘인터넷폭력로인한사회문제가날마다늘어나고있다.			일반. 문어. 대만. 경희대. 6급. 0003
15	15	있다	다/EF												요즘인터넷폭력로인한사회문제가날마다늘어나고있다.			일반. 문어. 대만. 경희대. 6급. 0003
16	16	있다	/SF												요즘인터넷폭력로인한사회문제가날마다늘어나고있다.			일반. 문어. 대만. 경희대. 6급. 0003
17	17	그	그/MM												[그]중에서가장심각한문제[가]넷폭력이라고생각한다.			일반. 문어. 대만. 경희대. 6급. 0003
18	18	중에서	중/NNB												[그]중에서가장심각한문제[가]넷폭력이라고생각한다.			일반. 문어. 대만. 경희대. 6급. 0003
19	19	중에서	에서/JKB												[그]중에서가장심각한문제[가]넷폭력이라고생각한다.			일반. 문어. 대만. 경희대. 6급. 0003
20	20	가장	가장/MAG												그중에서가장심각한문제[가]넷폭력이라고생각한다.			일반. 문어. 대만. 경희대. 6급. 0003
21	21	심각한	심각하/VA												그중에서가장[심각한]문제[가]넷폭력이라고생각한다.			일반. 문어. 대만. 경희대. 6급. 0003
22	22	심각한	하/ETM												그중에서가장[심각한]문제[가]넷폭력이라고생각한다.			일반. 문어. 대만. 경희대. 6급. 0003
23	23	문제	문제/NNG												그중에서가장심각한[문제]가넷폭력이라고생각한다.			일반. 문어. 대만. 경희대. 6급. 0003
24	24	문제	가/JKS												그중에서가장심각한[문제]가넷폭력이라고생각한다.			일반. 문어. 대만. 경희대. 6급. 0003
25	25	넷	넷/NNG												그중에서가장심각한[문제]가넷[넷]폭력이라고생각한다.			일반. 문어. 대만. 경희대. 6급. 0003
26	26	폭력이라고	폭력/NNG												그중에서가장심각한[문제]가넷[넷]폭력이라고생각한다.			일반. 문어. 대만. 경희대. 6급. 0003
27	27	폭력이라고	이/VCP												그중에서가장심각한[문제]가넷[넷]폭력이라고생각한다.			일반. 문어. 대만. 경희대. 6급. 0003
28	28	폭력이라고	라고/EC												그중에서가장심각한[문제]가넷[넷]폭력이라고생각한다.			일반. 문어. 대만. 경희대. 6급. 0003
29	29	생각한다	생각/NNG												그중에서가장심각한[문제]가넷[넷]폭력이라고생각한다.			일반. 문어. 대만. 경희대. 6급. 0003
30	30	생각한다	하/XSV												그중에서가장심각한[문제]가넷[넷]폭력이라고생각한다.			일반. 문어. 대만. 경희대. 6급. 0003
31	31	생각한다	다/EF												그중에서가장심각한[문제]가넷[넷]폭력이라고생각한다.			일반. 문어. 대만. 경희대. 6급. 0003

<그림 54> 학습자 말뭉치 구조

- 필드1, 필드2 일련번호
- 필드3 원어절
- 필드4 형태 주석
- 필드5 형태 주석 수정
- 필드6 교정 어절
- 필드7 교정 어절의 형태 주석
- 필드8 분석 여부
- 필드9 오류 층위
- 필드10 오류 현상
- 필드11 문장
- 필드12 파일명
-

이렇게 필드(field; column이라고도 함)와 레코드(record; row라고도 함)로 이루어져 있는 2차원 구조를 테이블(table) 구조라고 한다. 테이블 구조의 문서를 처리하는 데에는 엑셀(Excel) 같은 spreadsheet 소프트웨어가 적격이다. 여기서는 엑셀 활용법을 우선 살펴본 뒤, 21세기 세종계획에서 개발한 말뭉치 용례 검색 프로그램인 한마루와 자동으로 형태소를 분석해 주는 지능형 형태소 분석기를 학습자 말뭉치에 적용해 보도록 한다. 이어 국외에서 개발된 말뭉치 분석 프로그램인 Antconc와 엑셀과 비슷한 기능을 제공하는 공개/무료 소프트웨어인 Cygwin을 활용한 학습자 말뭉치 활용법을 알아보고자 한다. 이 장에서 학습자 말뭉치를 기존의 공개/상용 소프트웨어들을 통해 활용하는 방안을 기술할 때 학습자 말뭉치 내용 구축팀에서 제공한 초기 샘플 데이터를 사용하였음을 밝혀 둔다.

1.2. 엑셀을 이용한 말뭉치 활용

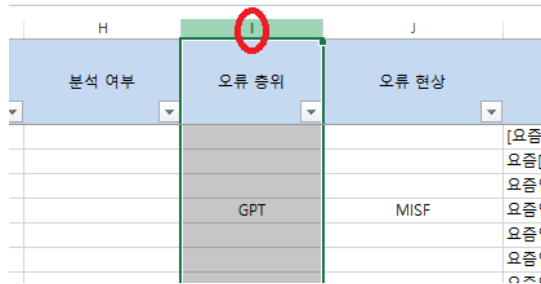
1.2.1. 정렬

학습자 말뭉치를 엑셀에서 열면 다음과 같이 나타난다.

	A	B	C	D	E	F	G	H	I	J	K	L
	전체 번호	과일 번호	원어	형태 주석	형태 수식	교정 어절	교정 어절의 형태 주석	분석 여부	오류 종류	오류 현상	원어/교정 문장	과일명
2	1	1	오물	오물/NNG							[오물]인태생복합표인한사회문제가날아다물어나고있다[물, 분어, 대안, 경피대, 6급, 0003]	
3	2	2	인태생	인태생/NNG							[오물]인태생복합표인한사회문제가날아다물어나고있다[물, 분어, 대안, 경피대, 6급, 0003]	
4	3	3	복합표	복합/NNG		복합으로					[오물]인태생복합표인한사회문제가날아다물어나고있다[물, 분어, 대안, 경피대, 6급, 0003]	
5	4	4	복합표	복합/NNG		복합으로	오물/NB		GPT	MSF	[오물]인태생복합표인한사회문제가날아다물어나고있다[물, 분어, 대안, 경피대, 6급, 0003]	
6	5	5	인태생	인태생/VV							[오물]인태생복합표인한사회문제가날아다물어나고있다[물, 분어, 대안, 경피대, 6급, 0003]	
7	6	6	인태생	인태생/VV							[오물]인태생복합표인한사회문제가날아다물어나고있다[물, 분어, 대안, 경피대, 6급, 0003]	
8	7	7	사회	사회/NNG							[오물]인태생복합표인한사회문제가날아다물어나고있다[물, 분어, 대안, 경피대, 6급, 0003]	
9	8	8	문제가	문제가/NNG							[오물]인태생복합표인한사회문제가날아다물어나고있다[물, 분어, 대안, 경피대, 6급, 0003]	
10	9	9	문제가	문제가/NNG							[오물]인태생복합표인한사회문제가날아다물어나고있다[물, 분어, 대안, 경피대, 6급, 0003]	
11	10	10	날아다	날아다/NNG							[오물]인태생복합표인한사회문제가날아다물어나고있다[물, 분어, 대안, 경피대, 6급, 0003]	
12	11	11	날아다	날아다/NNG							[오물]인태생복합표인한사회문제가날아다물어나고있다[물, 분어, 대안, 경피대, 6급, 0003]	
13	12	12	날아다	날아다/VV							[오물]인태생복합표인한사회문제가날아다물어나고있다[물, 분어, 대안, 경피대, 6급, 0003]	
14	13	13	날아다	날아다/VV							[오물]인태생복합표인한사회문제가날아다물어나고있다[물, 분어, 대안, 경피대, 6급, 0003]	
15	14	14	있다	있다/EC							[오물]인태생복합표인한사회문제가날아다물어나고있다[물, 분어, 대안, 경피대, 6급, 0003]	
16	15	15	있다	있다/EC							[오물]인태생복합표인한사회문제가날아다물어나고있다[물, 분어, 대안, 경피대, 6급, 0003]	
17	16	16	있다	있다/EC							[오물]인태생복합표인한사회문제가날아다물어나고있다[물, 분어, 대안, 경피대, 6급, 0003]	
18	17	17	그	그/MM							[그]중에서가장심각한문제가닷물복합이라고생각한다	
19	18	18	중에서	중에서/NNG							[그]중에서가장심각한문제가닷물복합이라고생각한다	
20	19	19	중에서	중에서/NNG							[그]중에서가장심각한문제가닷물복합이라고생각한다	
21	20	20	가장	가장/MAG							[그]중에서가장심각한문제가닷물복합이라고생각한다	
22	21	21	심각한	심각한/VV							[그]중에서가장심각한문제가닷물복합이라고생각한다	
23	22	22	심각한	심각한/VV							[그]중에서가장심각한문제가닷물복합이라고생각한다	
24	23	23	문제가	문제가/NNG							[그]중에서가장심각한문제가닷물복합이라고생각한다	
25	24	24	문제가	문제가/NNG							[그]중에서가장심각한문제가닷물복합이라고생각한다	
26	25	25	닷물	닷물/NNG							[그]중에서가장심각한문제가닷물복합이라고생각한다	
27	26	26	복합이라고	복합이라고/EC							[그]중에서가장심각한문제가닷물복합이라고생각한다	
28	27	27	복합이라고	복합이라고/EC							[그]중에서가장심각한문제가닷물복합이라고생각한다	
29	28	28	복합이라고	복합이라고/EC							[그]중에서가장심각한문제가닷물복합이라고생각한다	
30	29	29	생각한다	생각한다/VV							[그]중에서가장심각한문제가닷물복합이라고생각한다	
31	30	30	생각한다	생각한다/VV							[그]중에서가장심각한문제가닷물복합이라고생각한다	
32	31	31	생각한다	생각한다/VV							[그]중에서가장심각한문제가닷물복합이라고생각한다	
33	32	32	생각한다	생각한다/VV							[그]중에서가장심각한문제가닷물복합이라고생각한다	
34	33	33	이한	이한/MM							[이한]사회문제를가인하러면우선물바론인태생사용해도	
35	34	34	사회	사회/NNG							[이한]사회문제를가인하러면우선물바론인태생사용해도	
36	35	35	문제가	문제가/NNG							[이한]사회문제를가인하러면우선물바론인태생사용해도	
37	36	36	문제가	문제가/NNG							[이한]사회문제를가인하러면우선물바론인태생사용해도	
38	37	37	가인하러면	가인하러면/VV							[이한]사회문제를가인하러면우선물바론인태생사용해도	
39	38	38	가인하러면	가인하러면/VV							[이한]사회문제를가인하러면우선물바론인태생사용해도	
40	39	39	가인하러면	가인하러면/VV							[이한]사회문제를가인하러면우선물바론인태생사용해도	
41	40	40	우선	우선/MAG							[이한]사회문제를가인하러면우선물바론인태생사용해도	
42	41	41	물바론	물바론/VV							[이한]사회문제를가인하러면우선물바론인태생사용해도	
43	42	42	물바론	물바론/VV							[이한]사회문제를가인하러면우선물바론인태생사용해도	

<그림 55> 학습자 말뭉치를 엑셀에서 열었을 때의 모습

가장 먼저 손쉽게 할 수 있는 것은 데이터를 정렬(sort)하는 것이다. 우선 정렬 기준으로 삼고자 하는 필드를 클릭하여 선택한다. 9번째 필드인 '오류 총위'를 정렬 기준으로 삼고자 한다면, 오류 총위 필드의 맨 위에 있는 'I'자(<그림 56>의 붉은 색 원 부분)를 클릭하면 된다.



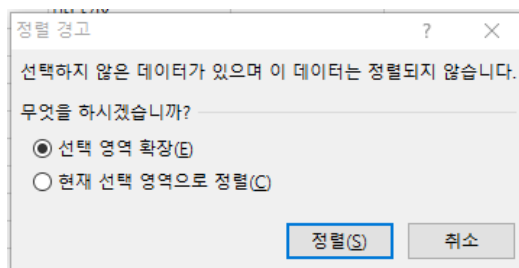
<그림 56> 정렬 기준 필드 선택

그 다음에 메뉴바의 '데이터'를 선택한 뒤 '정렬' 메뉴에서 '오름차순 정렬' 아이콘(<그림 57>의 붉은 색 원 부분)을 클릭한다.



<그림 57> 데이터-정렬

그러면 다음과 같은 팝업창이 뜬다.



<그림 58> 정렬 시 팝업 메뉴

‘선택 영역 확장’을 선택하면, 선택한 필드뿐 아니라 레코드 전체가 정렬 대상이 되고, ‘현재 선택 영역으로 정렬’을 선택하면 레코드의 나머지 부분은 그 자리에 그대로 있고 선택한 필드만 정렬되어 위치가 바뀐다. 대개의 경우는 전자를 선택해야 한다. 각 레코드가 지닌 자료적 통합성(integrity)을 보존하기 위해서이다. <그림 58>에서 ‘정렬’ 버튼을 클릭하면 정렬이 이루어져 다음과 같이 된다.

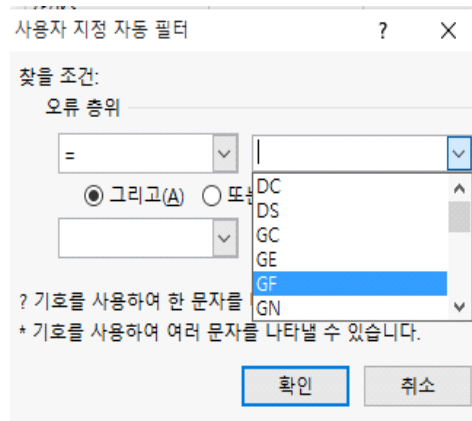
	A	B	C	D	E	F	G	H	I	J	K	L
	전체 번호	파일 번호	원어열	형태 주석	형태 주석 수정	교정 어열	교정 어열의 형태 주석	본식 여부	오류 종류	오류 현상	원어열 포함 문장	파일명
1									DC	REP	[그림 58]주말에 구출되었습니다.	일본, 중국, 러시아, 국민대, 1군, 2004
2	1198	26	그라서	그라서/NAU		그리고	그리고/NAU		DC	REP	[그림 58]주말에 구출되었습니다.	일본, 중국, 러시아, 국민대, 1군, 2004
3	2531	178	그라서	그라서/NAU		그리고	그리고/NAU		DC	REP	[그림 58]주말에 구출되었습니다.	일본, 중국, 러시아, 국민대, 1군, 2004
4	2866	78	그라서	그라서/NAU		그리고	그리고/NAU		DC	REP	[그림 58]주말에 구출되었습니다.	일본, 중국, 러시아, 국민대, 1군, 2004
5	2866	6	그라서	그라서/NAU		그리고	그리고/NAU		DC	REP	[그림 58]주말에 구출되었습니다.	일본, 중국, 러시아, 국민대, 1군, 2004
6	514	60	무엇인가?	무엇인가?		무엇인가?	무엇인가?		DS	REP	[그림 58]주말에 구출되었습니다.	일본, 중국, 러시아, 국민대, 1군, 2004
7	657	203	무엇인가?	무엇인가?		무엇인가?	무엇인가?		DS	REP	[그림 58]주말에 구출되었습니다.	일본, 중국, 러시아, 국민대, 1군, 2004
8	1048	284	가	가/REP		나	나/REP		DS	REP	[그림 58]주말에 구출되었습니다.	일본, 중국, 러시아, 국민대, 1군, 2004
9	1681	105	이	이/REP		나	나/REP		DS	REP	[그림 58]주말에 구출되었습니다.	일본, 중국, 러시아, 국민대, 1군, 2004
10	2355	2	가	가/REP		나	나/REP		DS	REP	[그림 58]주말에 구출되었습니다.	일본, 중국, 러시아, 국민대, 1군, 2004
11	466	12	발한하기	발한하기/XSV		발한하기	발한하기/XSV		GC	REP	[그림 58]주말에 구출되었습니다.	일본, 중국, 러시아, 국민대, 1군, 2004
12	153(2)					발	발/REP		GE	OM	[그림 58]주말에 구출되었습니다.	일본, 중국, 러시아, 국민대, 1군, 2004
13	318	10	종나	종나/REP		종나	종나/REP		GE	REP	[그림 58]주말에 구출되었습니다.	일본, 중국, 러시아, 국민대, 1군, 2004
14	337	29	가	가/REP		가	가/REP		GE	MISF	[그림 58]주말에 구출되었습니다.	일본, 중국, 러시아, 국민대, 1군, 2004
15	352	44	가	가/REP		가	가/REP		GE	MISF	[그림 58]주말에 구출되었습니다.	일본, 중국, 러시아, 국민대, 1군, 2004
16	455	147	실	실/REP		실	실/REP		GE	REP	[그림 58]주말에 구출되었습니다.	일본, 중국, 러시아, 국민대, 1군, 2004
17	454(1)					것	것/REP		GE	REP	[그림 58]주말에 구출되었습니다.	일본, 중국, 러시아, 국민대, 1군, 2004
18	454(2)					것	것/REP		GE	REP	[그림 58]주말에 구출되었습니다.	일본, 중국, 러시아, 국민대, 1군, 2004
19	454(3)					것	것/REP		GE	REP	[그림 58]주말에 구출되었습니다.	일본, 중국, 러시아, 국민대, 1군, 2004
20	476	22	지	지/REP		지	지/REP		GE	ADO	[그림 58]주말에 구출되었습니다.	일본, 중국, 러시아, 국민대, 1군, 2004
21	477	23	지	지/REP		지	지/REP		GE	MISF	[그림 58]주말에 구출되었습니다.	일본, 중국, 러시아, 국민대, 1군, 2004
22	630	176	한	한/REP		한	한/REP		GE	REP	[그림 58]주말에 구출되었습니다.	일본, 중국, 러시아, 국민대, 1군, 2004
23	709	255	한	한/REP		한	한/REP		GE	REP	[그림 58]주말에 구출되었습니다.	일본, 중국, 러시아, 국민대, 1군, 2004
24	713	259	한	한/REP		한	한/REP		GE	MISF	[그림 58]주말에 구출되었습니다.	일본, 중국, 러시아, 국민대, 1군, 2004
25	725	271	한	한/REP		한	한/REP		GE	MISF	[그림 58]주말에 구출되었습니다.	일본, 중국, 러시아, 국민대, 1군, 2004
26	754	300	한	한/REP		한	한/REP		GE	MISF	[그림 58]주말에 구출되었습니다.	일본, 중국, 러시아, 국민대, 1군, 2004
27	829	65	한	한/REP		한	한/REP		GE	MISF	[그림 58]주말에 구출되었습니다.	일본, 중국, 러시아, 국민대, 1군, 2004
28	856	94	한	한/REP		한	한/REP		GE	MISF	[그림 58]주말에 구출되었습니다.	일본, 중국, 러시아, 국민대, 1군, 2004
29	1026	262	한	한/REP		한	한/REP		GE	MISF	[그림 58]주말에 구출되었습니다.	일본, 중국, 러시아, 국민대, 1군, 2004
30	1087	323	한	한/REP		한	한/REP		GE	MISF	[그림 58]주말에 구출되었습니다.	일본, 중국, 러시아, 국민대, 1군, 2004
31	1204	32	한	한/REP		한	한/REP		GE	REP	[그림 58]주말에 구출되었습니다.	일본, 중국, 러시아, 국민대, 1군, 2004
32	1343	52	한	한/REP		한	한/REP		GE	REP	[그림 58]주말에 구출되었습니다.	일본, 중국, 러시아, 국민대, 1군, 2004
33	1417	126	한	한/REP		한	한/REP		GE	MISF	[그림 58]주말에 구출되었습니다.	일본, 중국, 러시아, 국민대, 1군, 2004
34	1550	159	한	한/REP		한	한/REP		GE	MISF	[그림 58]주말에 구출되었습니다.	일본, 중국, 러시아, 국민대, 1군, 2004
35	1669	93	한	한/REP		한	한/REP		GE	REP	[그림 58]주말에 구출되었습니다.	일본, 중국, 러시아, 국민대, 1군, 2004
36	1701	125	한	한/REP		한	한/REP		GE	MISF	[그림 58]주말에 구출되었습니다.	일본, 중국, 러시아, 국민대, 1군, 2004
37	1723	147	한	한/REP		한	한/REP		GE	MISF	[그림 58]주말에 구출되었습니다.	일본, 중국, 러시아, 국민대, 1군, 2004
38	1737	161	한	한/REP		한	한/REP		GE	OM	[그림 58]주말에 구출되었습니다.	일본, 중국, 러시아, 국민대, 1군, 2004
39	1810	234	한	한/REP		한	한/REP		GE	REP	[그림 58]주말에 구출되었습니다.	일본, 중국, 러시아, 국민대, 1군, 2004
40	1926	40	한	한/REP		한	한/REP		GE	MISF	[그림 58]주말에 구출되었습니다.	일본, 중국, 러시아, 국민대, 1군, 2004
41	2179	150	한	한/REP		한	한/REP		GE	REP	[그림 58]주말에 구출되었습니다.	일본, 중국, 러시아, 국민대, 1군, 2004
42	2238	209	한	한/REP		한	한/REP		GE	REP	[그림 58]주말에 구출되었습니다.	일본, 중국, 러시아, 국민대, 1군, 2004
43	2239	210	한	한/REP		한	한/REP		GE	REP	[그림 58]주말에 구출되었습니다.	일본, 중국, 러시아, 국민대, 1군, 2004

<그림 59> ‘오류 총위’를 기준으로 정렬한 결과

오류 총위 DC 4개, DS 5개, GC 1개, GE 135개 …… 하는 식으로 오류 총위가 동일한 레코드들이 한데 모여 나타난다. 이렇게 정렬 기능을 이용하여, 같은 성격의 레코드들을 한데 모아서 살펴볼 수 있다.

정렬 기능을 이용하면 같은 성격의 레코드들을 한데 모을 수 있기는 하나, 원본 데이터의 배열순서는 깨어진다. 대개의 경우 원본 데이터의 배열순서도 중요한 정보를 담고 있으므로, 정렬을 한 결과는 다른 이름으로 저장하는 것이 좋다. 즉 원본 파일과 정렬된 파일을 각각 따로 저장해 두는 것이 좋다.

해당 필드에 들어 있는 문자열에 대해 ‘같음’, ‘같지 않음’, ‘시작 문자’, ‘끝 문자’, ‘포함’, ‘포함하지 않음’ 등의 여러 조건을 지정할 수 있다. 가장 기본적인 ‘같음’을 선택하면, 다시 <그림 62>와 같은 팝업창이 뜬다. ‘=’는 ‘같음’이라는 뜻이고, 그 오른쪽의 드롭다운 메뉴를 클릭하여 원하는 문자열을 선택할 수 있다. 드롭다운 메뉴에 원하는 문자열이 없으면 직접 타이핑해서 입력할 수도 있다.



<그림 62> 필터 설정 2

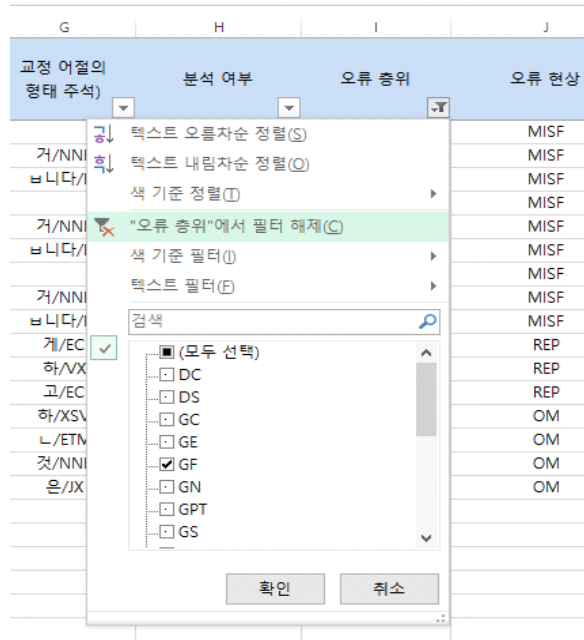
‘GF’를 선택하여 ‘확인’ 버튼을 클릭하면 다음과 같이 된다.

	A	B	C	D	E	F	G	H	I	J	K	L
	현재 번호	과일 번호	원어명	형태 주석	형태 주석 수정	고정 어휘	고정 어휘의 형태 주석	분석 여부	오류 종류	오류 현상	원어명 포함 문장	과실명
1												
1276	1247	75	갈	비/ETM					GF	MISF	저변형N서술타워(잡)입니다.	일본, 문어, 러시아, 국민대, 1급, 0024
1277	1248	76	잡니다.	가/VV		잡니다.	가/NNB		GF	MISF	저변형N서술타워(잡)입니다.	일본, 문어, 러시아, 국민대, 1급, 0024
1278	1249	77	잡니다.	비니다/EF		잡니다.	비니다/EF		GF	MISF	저변형N서술타워(잡)입니다.	일본, 문어, 러시아, 국민대, 1급, 0024
1297	1272	100	죽을	죽/ETM					GF	MISF	그래서N서술타워에서연구하고알아스파가리물죽을(가) 일본, 문어, 러시아, 국민대, 1급, 0024	
1298	1273	101	가습니다.	가습니다/NA	가/NNB	잡니다.	가/NNB		GF	MISF	그래서N서술타워에서연구하고알아스파가리물죽을(가) 일본, 문어, 러시아, 국민대, 1급, 0024	
1299	1273(1)		가습니다.		습니다/EF	잡니다.	비니다/EF		GF	MISF	그래서N서술타워에서연구하고알아스파가리물죽을(가) 일본, 문어, 러시아, 국민대, 1급, 0024	
1306	1280	108	갈	비/ETM					GF	MISF	그래서N서술타워에서연구하고알아스파가리물죽을(가) 일본, 문어, 러시아, 국민대, 1급, 0024	
1307	1281	109	가습니다.	갈/NA	가/NNB	잡니다.	가/NNB		GF	MISF	그래서N서술타워에서연구하고알아스파가리물죽을(가) 일본, 문어, 러시아, 국민대, 1급, 0024	
1308	1282	110	가습니다.	습니다/EF	습니다/EF	잡니다.	비니다/EF		GF	MISF	그래서N서술타워에서연구하고알아스파가리물죽을(가) 일본, 문어, 러시아, 국민대, 1급, 0024	
1690	1654(1)	79	살고	고/EC		살게	게/EC		GF	REP		
1691	1654(2)					하고	하/NX		GF	REP		
1692	1655					하고	고/EC		GF	REP	나는영화가본지(살고)살지않다.	일본, 문어, 러시아, 국민대, 3급, 0028
2084	2828(1)					한	하/XSV		GF	OM		
2085	2828(2)					한	하/ETM		GF	OM		
2086	2828(3)					것은	것/NNB		GF	OM		
2087	2828(4)					것은	은/OX		GF	OM		
43997												
43998												
43999												

<그림 63> 필터 적용 결과

즉, '오류 총위' 필드가 'GF'인 16개의 레코드만이 나타나는 것이다.

우하단의 역삼각형 아이콘을 다시 클릭하여 '필터 해제'를 선택하면(<그림 64>에서 연두색 부분), 원본 데이터 전체가 다시 나타난다.



<그림 64> 필터 해제

1.2.3. 수식

수치 데이터를 포함하고 있는 테이블 구조의 문서에서 수치 계산을 할 때 엑셀을 이용하면 편리하다. 이해를 돕기 위해 매우 단순한 사례를 들어 살펴보자.

	A	B	C	D	E
1	수준	국적	학습자 수	오류 빈도	오류 빈도 합계
2	초급	중국	3	10	
3		일본	2	15	
4		미국	2	20	
5	중급	중국	1	10	
6		미국	3	20	
7		베트남	2	30	
8	고급	중국	3	10	
9		일본	4	15	
10		미국	2	20	
11		베트남	3	30	

<그림 65> 학습자 오류 현황

<그림 65>는 학습자의 오류 현황을 기록한 표이다. ‘오류 빈도 합계’ 필드를 채워 넣으려면 ‘학습자 수’ 필드의 값과 ‘오류 빈도’ 필드의 값을 곱해야 할 것이다. 이를 위해서는 우선 결과값이 들어갈 E2~E11 부분을 블록 지정하고, 수식 입력란에 ‘=C2:C11*D2:D11’이라고 입력한 뒤 Ctrl+Shift+Enter를 친다.

	A	B	C	D	E
1	수준	국적	학습자 수	오류 빈도	오류 빈도 합계
2	초급	중국	3	10	=C11*D2:D11
3		일본	2	15	
4		미국	2	20	
5	중급	중국	1	10	
6		미국	3	20	
7		베트남	2	30	
8	고급	중국	3	10	
9		일본	4	15	
10		미국	2	20	
11		베트남	3	30	

<그림 66> 수식 입력

그러면 다음과 같이 수식 계산 결과가 나온다.

	A	B	C	D	E
1	수준	국적	학습자 수	오류 빈도	오류 빈도 합계
2	초급	중국	3	10	30
3		일본	2	15	30
4		미국	2	20	40
5	중급	중국	1	10	10
6		미국	3	20	60
7		베트남	2	30	60
8	고급	중국	3	10	30
9		일본	4	15	60
10		미국	2	20	40
11		베트남	3	30	90

<그림 67> 수식 계산 결과

모든 학습자의 오류 빈도를 합산하려면, E2~E11을 블록 지정한 상태에서 메뉴바의 ‘수식’ - ‘자동합계’ - ‘합계’를 클릭한다(<그림 68>).

E2		fx {=C2:C11*D2:D11}			
	A	B	C	D	E
1	수준	국적	학습자 수	오류 빈도	오류 빈도 합계
2	초급	중국	3	10	30
3		일본	2	15	30
4		미국	2	20	40
5	중급	중국	1	10	10
6		미국	3	20	60
7		베트남	2	30	60
8	고급	중국	3	10	30
9		일본	4	15	60
10		미국	2	20	40
11		베트남	3	30	90

<그림 68> 합계 함수 적용 방법

그러면 다음 <그림 69>처럼 E12 셀에 합계가 나온다.

E2		fx {=C2:C11*D2:D11}			
	A	B	C	D	E
1	수준	국적	학습자 수	오류 빈도	오류 빈도 합계
2	초급	중국	3	10	30
3		일본	2	15	30
4		미국	2	20	40
5	중급	중국	1	10	10
6		미국	3	20	60
7		베트남	2	30	60
8	고급	중국	3	10	30
9		일본	4	15	60
10		미국	2	20	40
11		베트남	3	30	90
12					450

<그림 69> 합계 함수 적용 결과

요컨대, 엑셀의 대표적인 기능은 테이블 구조의 문서에서 일정한 필드에 기록된 수치들을 가지고 일관된 계산을 행하여 그 결과를 다른 필드의 대응하는 셀에 기록하는 일이다.

1.2.4. 함수

엑셀은 사용자들이 자주 사용하는 수식들을 아예 함수로 만들어 놓았다. 학습자 말뭉치에서 ‘오류 총위’ GPT의 출현 빈도를 알고 싶다고 하자. 그러면 함수를 적용할 범위는 ‘오류 총위’ 필드 처음부터 끝까지이므로 I2~I11826이 된다. 그리고 count를 할 조건은 “GPT”이다. 결과값을 출력한 셀을 선택한 상태에서(<그림 70>에서는 I11827), 수식 및 함수 입력란에 =COUNTIF(I2:I11826, “GPT”)라고 입력한다.

=COUNTIF(I2:I11826,"GPT")										
	A	B	C	D	E	F	G	H	I	J
1	전제 번호1	파일 번호	원어절	형태 주석	형태 주석 수정	교정 어절	교정 어절의 형태 주석)	분석 여부	오류 종류	오류 현상
11797	11655	145	그렇지만	그렇지만/MAJ						
11798	11656	146	학교에	학교/NING						
11799	11657	147	학교에	에//KB						
11800	11658	148	가야	가/AV						
11801	11659	149	가야	아아/EC						
11802	11660	150	해.	하/VX						
11803	11661	151	해.	아/EF						
11804	11662	152	해.	/SF						
11805	11663	153	다음에	다음/NING						
11806	11664	154	다음에	에//KB						
11807	11665	155	기회	기회/NING						
11808	11666	156	있으면	있/AV						
11809	11667	157	있으면	오면/EC						
11810	11668	158	부모님께	부모님/NING	부모님이					
11811	11669	159	부모님께	에//KB	부모님이		이//KS		GPT	REP
11812	11670	160	한국에	한국/NINP						
11813	11671	161	한국에	에//KB						
11814	11672	162	출	오/AV						
11815	11673	163	출	ㅁ/ETM						
11816	11674	164	때	때/NING						
11817	11675	165	꼭	꼭/MAG						
11818	11676	166	가지고	가지/AV		같이	같이/MAG		VV	REP
11819	11677	167	가지고	고/EC		<삭제>				
11820	11678	168	오세요.	오/AV		가세요.	가/AV		VV	REP
11821	11679	169	오세요.	시/EP		가세요.				
11822	11680	170	오세요.	어요/EF		가세요.				
11823	11681	171	오세요.	/SF		가세요.				
11824	11682	172	재미있어	재미있/VA						
11825	11683	173	재미있어	어/EF	재미있어요	어요/EF			GE	REP
11826	11683(1)				재미있어요.	/SF			GS	OM
11827									=COUNTIF(I2:I11826	
11828										

<그림 70> COUNTIF 함수로 GPT의 빈도 세기

커서를 다른 셀로 이동하면 I11827 셀에 242라는 결과값이 표시된다.

	A	B	C	D	E	F	G	H	I	J
1	전제 번호1	파일 번호	원어절	형태 주석	형태 주석 수정	교정 어절	교정 어절의 형태 주석)	분석 여부	오류 종류	오류 현상
11797	11655	145	그렇지만	그렇지만/MAJ						
11798	11656	146	학교에	학교/NING						
11799	11657	147	학교에	에//KB						
11800	11658	148	가야	가/AV						
11801	11659	149	가야	아아/EC						
11802	11660	150	해.	하/VX						
11803	11661	151	해.	아/EF						
11804	11662	152	해.	/SF						
11805	11663	153	다음에	다음/NING						
11806	11664	154	다음에	에//KB						
11807	11665	155	기회	기회/NING						
11808	11666	156	있으면	있/AV						
11809	11667	157	있으면	오면/EC						
11810	11668	158	부모님께	부모님/NING	부모님이					
11811	11669	159	부모님께	에//KB	부모님이		이//KS		GPT	REP
11812	11670	160	한국에	한국/NINP						
11813	11671	161	한국에	에//KB						
11814	11672	162	출	오/AV						
11815	11673	163	출	ㅁ/ETM						
11816	11674	164	때	때/NING						
11817	11675	165	꼭	꼭/MAG						
11818	11676	166	가지고	가지/AV		같이	같이/MAG		VV	REP
11819	11677	167	가지고	고/EC		<삭제>				
11820	11678	168	오세요.	오/AV		가세요.	가/AV		VV	REP
11821	11679	169	오세요.	시/EP		가세요.				
11822	11680	170	오세요.	어요/EF		가세요.				
11823	11681	171	오세요.	/SF		가세요.				
11824	11682	172	재미있어	재미있/VA						
11825	11683	173	재미있어	어/EF	재미있어요	어요/EF			GE	REP
11826	11683(1)				재미있어요.	/SF			GS	OM
11827									242	
11828										

<그림 71> COUNTIF 함수 적용 결과

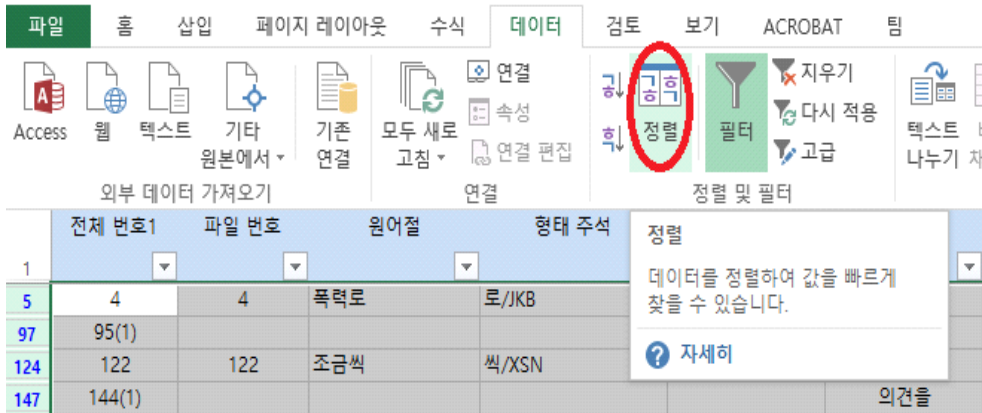
1.2.5. 필터와 정렬의 결합 활용

학습자 말뭉치에서 ‘오류 층위’의 유형들 중에서 GPT(조사 오류)가 242개로 가장 많다. 오류 층위 중 1등을 차지한 GPT(조사 오류)를 더 자세히 분석해 보자. 조사 오류에는 어떤 하위 유형들이 있을까? 즉, 외국인 학습자는 한국어의 조사 중 어떤 조사를 가장 자주 틀릴까? 이 과제를 수행하기 위해서는 우선 ‘오류 층위’ 필드에 필터를 적용해서 ‘GPT’인 것들만 걸러낼 필요가 있다. 이것은 앞서 1.2.2.에서 설명한 대로 쉽게 할 수 있다. <그림 62>의 단계에서 ‘GPT’를 선택하면 된다.

	A	B	C	D	E	F	G	H	I	J	K	L
	원래 번호1	과일 번호	원어명	항목 주석	항목 주석 수정	교정 유형	교정 이유의 요약	분석 여부	오류 종류	오류 현상	원어명 표상 문장	과일명
5	4	4	목련목	목(X)	목련목	오류(X)	오류(X)		GPT	MSF	소름나무의(목련목)한사화재가발발하다니라고도 들린다. 물론, 다만, 한사화재 , 6.6, 2003	
97	95(1)					복합	복합		REP	OM		
109	12	122	조금씩	복(X)YN	조금더하다	오류(X)	오류(X)		REP	OM		
147	144(1)					오직	오직		REP	OM		
148	343	35	불대역	목(X)	불대역	오류(X)	오류(X)		REP	OM		
378	372(1)					상당량과	2/100		GPT	MSF	이상(상당량)의와안이다불대역과2/100을형용사형용사까지도 들린다. 물론, 다만, 한사화재 , 6.6, 2003	
438	450	142	한여름	목(X)YN	한여름	오류(X)	오류(X)		REP	OM		
475	463	9	막	목(X)NN	변역	오류(X)	오류(X)		REP	OM		
477	464(1)					강각목	강각목		REP	OM		
530	517	63	예물	목(X)	목(X)	오류(X)	오류(X)		GPT	MSF	한대도시의문화유산이무엇인가가(예물)이대란의사도 들린다. 물론, 다만, 한사화재 , 6.6, 2008-02	
538	523	69	수조	목(X)	수조	오류(X)	오류(X)		REP	OM		
537	523(1)					수조	수조		REP	OM		
539	544(1)					문화유산	문화유산		REP	OM		
585	568(1)					나뭇잎	나뭇잎		REP	OM		
602	585	131	다기모양	목(X)	다기모양	오류(X)	오류(X)		REP	OM		
611	594	140	노년문화	목(X)	노년문화	오류(X)	오류(X)		REP	OM		
657	640	186	도시사보	목(X)	도시사보	오류(X)	오류(X)		REP	OM		
658	648	195	문화유산	목(X)	문화유산	오류(X)	오류(X)		REP	OM		
702	685	231	집문	목(X)	집문	오류(X)	오류(X)		REP	OM		
723	706	252	자랑	목(X)	자랑	오류(X)	오류(X)		REP	OM		
733	716	262	장문	목(X)	장문	오류(X)	오류(X)		REP	OM		
750	733	279	한여름	목(X)	한여름	오류(X)	오류(X)		REP	OM		
841	823	59	간역	목(X)	간역	오류(X)	오류(X)		ADO	GPT		
845	825(1)					한여름	한여름		REP	OM		
847	826(1)					자전역	자전역		REP	OM		
864	843	79	논산읍	목(X)	논산읍	오류(X)	오류(X)		REP	OM		
879	856(1)					문화유산	문화유산		REP	OM		
891	907(1)					문화유산	문화유산		REP	OM		
892	945(1)					문화유산	문화유산		REP	OM		
906	961	197	물	목(X)	물	오류(X)	오류(X)		REP	OM		
1002	976	212	산간문화	목(X)	산간문화	오류(X)	오류(X)		REP	OM		
1003	1024(1)					산간문화	산간문화		REP	OM		
1081	1055(1)					자전역	자전역		REP	OM		
1098	1070	306	가발도산국역	목(X)	가발도산국역	오류(X)	오류(X)		REP	OM		
1146	1117	353	가발도산국역	목(X)	가발도산국역	오류(X)	오류(X)		REP	OM		
1181	1132	368	가발도산국역	목(X)	가발도산국역	오류(X)	오류(X)		REP	OM		
1193	1164	400	부대역	목(X)	부대역	오류(X)	오류(X)		REP	OM		
1201	1232	60	카트에서	목(X)	카트에서	오류(X)	오류(X)		REP	OM		
1203	1282(1)					나	나		REP	OM		
1481	1334	43	성	목(X)	성	오류(X)	오류(X)		REP	OM		
1481	1384	93	이름	목(X)	이름	오류(X)	오류(X)		REP	OM		

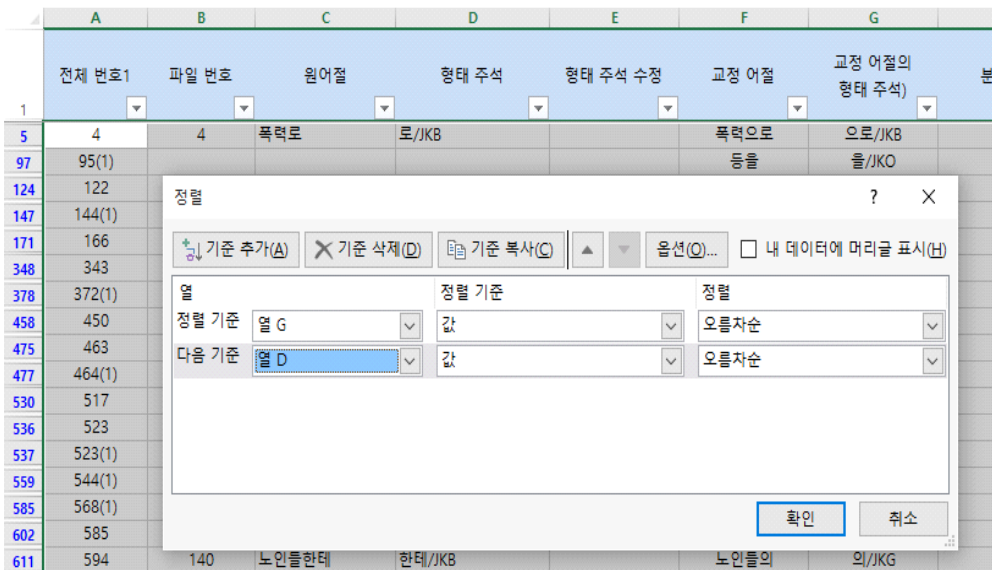
<그림 72> ‘오류 총위’ 필드에 필터를 적용하여 ‘GPT’만 걸러낸 결과

그 다음에는 정렬 대상이 되는 242개의 레코드 전체를 선택한다. 맨 좌측의 레코드 번호를 클릭하여 죽 아래로 스크롤하면 된다. 242개의 레코드 전체가 선택된 상태에서 메뉴바의 '데이터' - '정렬 및 필터' - '정렬'(<그림 73>의 붉은색 원부분)을 선택한다.



<그림 73> 정렬

그러면 정렬 기준을 설정하기 위한 팝업창이 뜬다(<그림 74>). 여기서 첫째 정렬 기준으로 필드 G(교정 어절의 형태 주석: 문맥상 올바른 조사)를 선택하고, '기준 추가' 버튼을 클릭하여 둘째 정렬 기준으로 필드 D(형태 주석: 사용자가 잘못 사용한 조사)를 선택한다.



<그림 74> 정렬 기준 설정

정렬을 한 결과는 다음과 같다.

	A	B	C	D	E	F	G	H	I	J	K
1	전체 번호	파일 번호	원어절	형태 주석	형태 주석 수정	교정 어절	교정 어절의 형태 주석	분석 여부	오류 종류	오류 현상	원어절 포함 문장
5	716	262	정부는	는/JK		정부가	가/JKS		GPT	REP	[정부는]친환경으로공급이내나우물들이지않아서인물이
97	2068	39	가치는	는/JK		가치가	가/JKS		GPT	REP	올바른인턴보사올레도의제가없이있지만세[가치]는특히
124	5200	166	기회도	도/JK		기회가	가/JKS		GPT	REP	면접시험이있으므로그런것이없는사람도일사할수있지만
147	3448	52	바나나를	를/JKO		바나나가	가/JKS		GPT	REP	그가가는수박과포도요[바나나]입니다.
171	9820	21	친구들	들/JKO		친구가	가/JKS		GPT	REP	저를한국[친구들]라고합니다.
248	11241	41	숙제를	를/JKO		숙제가	가/JKS		GPT	REP	토요일에[숙제]를남나친구같이소프으려명통에갑니다
378	640	186	다도시보다	보다/JKB		다도시가	가/JKS		GPT	REP	법정통도시물보다[다도시]보다입니다.
458	2409	56	인덕물을	를/JKO		사이트가	가/JKS		GPT	REP	문화지이때문에한국어에대해이해하기어려울텐데[이]
475	1800	224	교수님이	이/JKS		교수가	가/JKS		GPT	REP	시애틀에실리학[교수님]이되고있다.
477	372(1)					친구가	가/JKS		GPT	OM	
530	826(1)					자전거가	가/JKS		GPT	OM	
536	856(1)					빌다가	가/JKS		GPT	OM	
537	1024(1)				화/XSN	지구온난화가	가/JKS		GPT	OM	
559	2332(1)					태도가	가/JKS		GPT	OM	
585	5962(1)					가치가	가/JKS		GPT	OM	
602	10021(1)					크리스마스가	가/JKS		GPT	OM	
611	11285(1)					먹주가	가/JKS		GPT	OM	
657	166	166	상대방이	와/JKB		상대방과	과/JKB		GPT	MISF	[상대방과]의간이더물하고간을완전하여양정하게자신
666	3718	40	사람을	를/JKO		사람과	과/JKB		GPT	REP	그리고저는사람을보고있습니다.[보기]TOEIC그래서저는
702	11449	114	친구들을	를/JKO		친구들과	과/JKB		GPT	REP	나는우석씨도[친구들]같이늘어나니까우석씨는다
723	5444(1)					옷과	과/JKB		GPT	OM	
733	10057(1)					친구들과	과/JKB		GPT	OM	
750	3093	111	친구들은	은/JK		친구들과	과/JK		GPT	REP	남자친구[친구들은]이여기물하고있습니다.
842	7982	107	2005년에	에/JKB		2005년까지	까지/JK		GPT	REP	사람들은원자력발전의필요성을알고있고2002년부터[2]
845	1990(1)					부도일게	게/JKB		GPT	OM	
847	523(1)					수도인	인/JKM		GPT	REP	
864	1293(1)		내가	가/JKS		나는	는/JK		GPT	OM	
879	1482	191	내가	가/JKS		나는	는/JK		GPT	REP	[내가]좋은회사에서일하고있다.
931	3499	103	친구도	도/JK		친구는	는/JK		GPT	REP	어제[나와]친구도[같이]식당에서먹었습니다.
970	11069	192	우리도	도/JK		우리는	는/JK		GPT	REP	오후내시반에[우리도]비가고됩니다.
986	9817	18	저를	를/JKO		저는	는/JK		GPT	REP	[저를]한국친구들라고합니다.
1002	9842	43	영화를	를/JKO		영화는	는/JK		GPT	REP	한국[영화를]미리있습니다.
1051	8937	261	상각하는	어는/JKA	어는/JK	상각하는	는/JK		GPT	MISF	내[상각하는]남편[사람]은정말물한다거나[자]가가지고있
1063	11549	39	오이도에서	에서/JKB		오이도는	는/JK		GPT	REP	[오이도에서]물없이
1086	11564	54	오이도에서	에서/JKB		오이도는	는/JK		GPT	REP	[오이도에서]아름다운곳이에요.
1146	11601	91	오이도에서	에서/JKB		오이도는	는/JK		GPT	REP	[오이도에서]해물요양객.
1181	10066(1)					우리는	는/JK		GPT	OM	
1193	11365(1)					바다는	는/JK		GPT	OM	
1261	11554(1)					우리는	는/JK		GPT	OM	
1320	8605(1)				소/JK	것도	도/JK		GPT	MISF	
1361	4954	59	버스에	에/JKB		버스로	로/JKB		GPT	REP	길속에서다른[버스에]같이타고갑니다.
1411	4946(1)				리/JKB	버스로	로/JKB		GPT	OM	

<그림 75> 필드 G, 필드 D를 기준으로 정렬한 결과

맨 위의 17개는 주격조사 ‘가’를 써야 옳은 경우이다. 여기서 잘못하여 ‘는’을 쓴 것이 2회, ‘도’를 쓴 것이 1회, ‘를’을 쓴 것이 3회, ‘보다’를 쓴 것이 1회, ‘을’을 쓴 것이 1회, ‘이’를 쓴 것이 1회로 나타난다. 이렇게 정렬 기준의 우선순위를 적절히 설정함으로써, 원하는 기준에 있어서 동일한 레코드들이 한데 모여서 나타난다.

1.2.6. 복합 기준에 의한 빈도 추출

위에서 보았듯이 필터를 사용하여 탐색 범위를 줄이고 그 다음에 정렬 기준을 설정하여 정렬함으로써 비슷한 성격의 레코드들을 한데 모아서 살펴보는 식의 작업 방식도 물론 가능하고 유용하지만, 빈도를 측정하고자 하는 대상의 조건이 처음부터 명확하면 함수를 사용하여 한 번에 빈도를 추출할 수도 있다. 예컨대 주격조사 ‘이’를 써야 하는데 다른 조사를 쓰거나 조사를 아예 안 써서 조사 오류가 발생한 경우를 세고 싶다고 치자. 그렇다면 ‘오류 종류’ 필드가 ‘GPT’이고 ‘교정 어절의 형태 주석’ 필드가 ‘이/JKS’인 레코드를 세면 될 것이다. 이렇게 조건이 둘 이상일 때는 COUNTIF 함수 대신 COUNTIFS 함수를 쓰면 된다. 결과 값이 나올 셀(<그림 76>에서는 I11828)을 선택한 상태에서, 수식 및 함수 입력란

에 =COUNTIFS(I2:I11826,"GPT",G2:G11826,"이/JS")라고 입력한다. 커서를 다른 셀로 옮기면 결과값 30이 나타난다.

=COUNTIFS(I2:I11826,"GPT",G2:G11826,"이/JS")										
	A	B	C	D	E	F	G	H	I	J
1	전체 번호1	파일 번호	원어절	형태 주석	형태 주석 수정	교정 어절	교정 어절의 형태 주석)	분석 여부	오류 종류	오류 현상
11807	11665	155	기회	기회/NING						
11808	11666	156	있으면	있/VV						
11809	11667	157	있으면	있면/EC						
11810	11668	158	부모님에	부모님/NING		부모님이				
11811	11669	159	부모님에	예/KB		부모님이	이/JS		GPT	REP
11812	11670	160	한국에	한국/NNP						
11813	11671	161	한국에	예/KB						
11814	11672	162	출	오/VV						
11815	11673	163	출	은/ETM						
11816	11674	164	때	때/NING						
11817	11675	165	꼭	꼭/MAG						
11818	11676	166	가지고	가지/VV		같이	같이/MAG		VV	REP
11819	11677	167	가지고	고/EC		<삭제>				
11820	11678	168	오세요.	오/VV		가세요.	가/VV		VV	REP
11821	11679	169	오세요.	시/EP		가세요.				
11822	11680	170	오세요.	어요/EF		가세요.				
11823	11681	171	오세요.	/SF		가세요.				
11824	11682	172	재미있어	재미있/VV						
11825	11683	173	재미있어	어/EF		재미있어요	어요/EF		GE	REP
11826	11683(1)					재미있어요	/SF		GS	OM
11827									242	
11828									=COUNTIFS(I2:I118	
11829										

<그림 76> COUNTIFS 함수 사용 방법

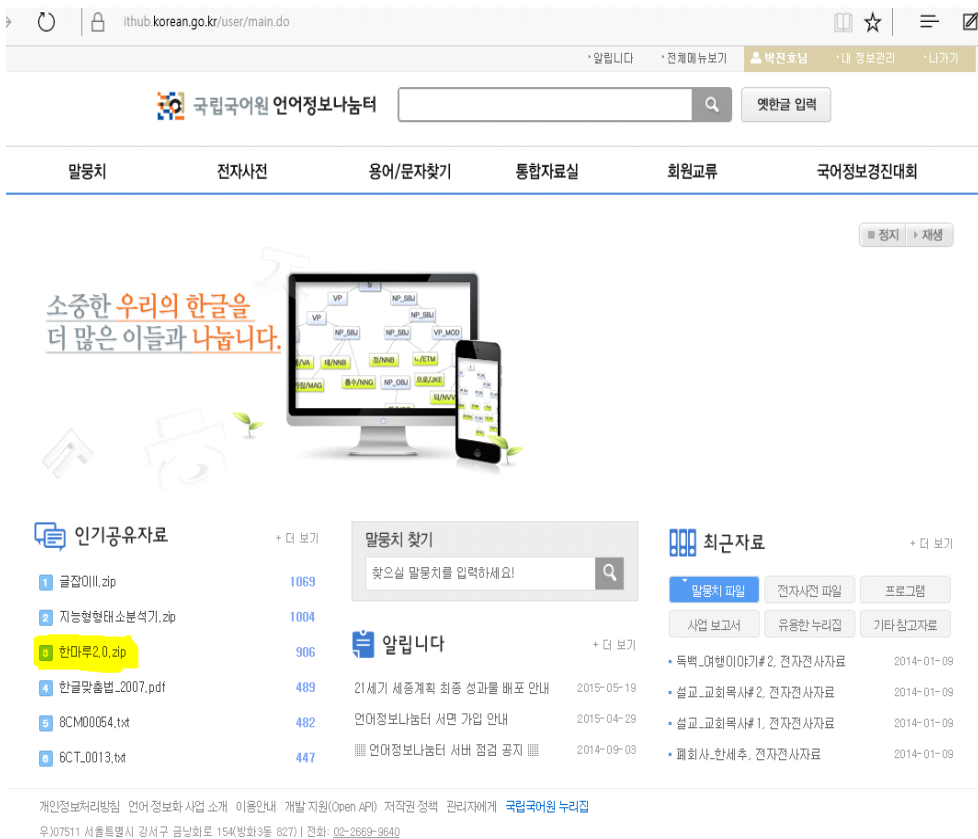
기준을 3개 이상 지정할 수도 있다. 예컨대 주격조사 '이'를 써야 하는데 대격 조사 '을'을 잘못 쓴 레코드를 세려면 =COUNTIFS(I2:I11826,"GPT",G2:G11826,"이/JS", D2:D11826,"을/JO")라고 하면 될 것이다. 결과값은 13이다.

	A	B	C	D	E	F	G	H	I	J
1	전체 번호1	파일 번호	원어절	형태 주석	형태 주석 수정	교정 어절	교정 어절의 형태 주석)	분석 여부	오류 종류	오류 현상
11807	11665	155	기회	기회/NING						
11808	11666	156	있으면	있/VV						
11809	11667	157	있으면	있면/EC						
11810	11668	158	부모님에	부모님/NING		부모님이				
11811	11669	159	부모님에	예/KB		부모님이	이/JS		GPT	REP
11812	11670	160	한국에	한국/NNP						
11813	11671	161	한국에	예/KB						
11814	11672	162	출	오/VV						
11815	11673	163	출	은/ETM						
11816	11674	164	때	때/NING						
11817	11675	165	꼭	꼭/MAG						
11818	11676	166	가지고	가지/VV		같이	같이/MAG		VV	REP
11819	11677	167	가지고	고/EC		<삭제>				
11820	11678	168	오세요.	오/VV		가세요.	가/VV		VV	REP
11821	11679	169	오세요.	시/EP		가세요.				
11822	11680	170	오세요.	어요/EF		가세요.				
11823	11681	171	오세요.	/SF		가세요.				
11824	11682	172	재미있어	재미있/VV						
11825	11683	173	재미있어	어/EF		재미있어요	어요/EF		GE	REP
11826	11683(1)					재미있어요	/SF		GS	OM
11827									242	
11828									30	
11829									13	
11830										

<그림 77> 3가지 기준을 사용한 COUNTIFS 함수 적용 결과

1.3. 한마루 2.0을 이용한 말뭉치 활용

말뭉치에서 원하는 문자열을 검색하여 그 앞뒤 문맥을 보여주는 소프트웨어를 콘코던스(concordance) 프로그램이라고 한다. 여기서는 국내에서 만들어진 한마루와 외국에서 만들어진 AntConc를 중심으로 설명하겠다. 한마루 2.0(이하 '한마루'라 부른다)은 국립국어원 언어정보나눔터 홈페이지에서 회원가입 후 무료로 내려받을 수 있다.



The screenshot shows the website of the National Korean Language Resource Center (국립국어원 언어정보나눔터). The page features a navigation bar with links to '말뭉치' (Corpus), '전자사전' (Electronic Dictionary), '용어/문자찾기' (Term/Character Search), '통합자료실' (Integrated Data Room), '회원교류' (Member Exchange), and '국어정보경진대회' (Korean Language Information Competition). A search bar is located at the top right. Below the navigation bar, there is a section titled '인기공유자료' (Popular Shared Data) with a list of files. The file '한마루2.0.zip' is highlighted in yellow. To the right of this list is a section titled '말뭉치 찾기' (Find Corpus) with a search bar. Below this is a section titled '알립니다' (Notice) with a list of notices. At the bottom of the page, there is a footer with contact information and a link to the '국립국어원 누리집' (National Korean Language Resource Center Website).

<그림 78> 언어정보나눔터 초기 화면

1.3.1. 한마루 사용을 위한 말뭉치 파일 코드 변환: UniConv

한마루는 input 파일의 인코딩이 UTF-16LE(LE는 little endian)로 되어 있을 것

을 전제로 한다. 학습자 원시 말뭉치는 UTF-8으로 되어 있고, 학습자 형태분석 말뭉치는 CP949로 되어 있는데, 둘 다 UTF-16LE로 변환해야 한마루를 이용할 수 있다(향후 배포되는 학습자 말뭉치는 모두 기본적으로 UTF-8으로 인코딩되어 있다. 그러나 사용자가 내려받기 하는 과정에서 필요한 인코딩을 선택할 수 있다. 여기에서는 사용자가 필요에 따라 인코딩을 변경할 경우를 포함하여 설명하였다). 텍스트 파일의 문자 인코딩의 변환을 위해서는 UniConv라는 프로그램이 유용하다. 아래의 주소에서 다운로드할 수 있다.

<http://ringtail.its.monash.edu.au/pub/nihongo/unicnv.zip>

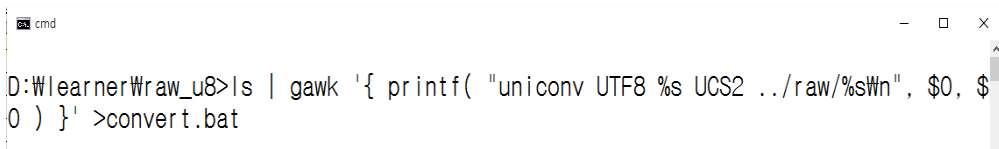
<http://hosting03.snu.ac.kr/~korean/old/data/han/unicnv.zip>

UniConv를 명령프롬프트 상의 모든 디렉토리에서 사용할 수 있게 하려면, uniconv.exe와 btuc21d3.dll(또는 btuc220.dll) 이 두 파일을 1.6.2.에서 path 설정한 디렉토리에 넣어 두어야 한다.

UniConv는 기본적으로 명령프롬프트에서 다음과 같은 문법으로 사용한다.

uniconv input-encoding input-file output-encoding output-file

그런데 변환해야 할 파일의 수가 많을 때에는 배치(batch) 파일을 만들어서 한꺼번에 처리하는 것이 편리하다. 배치 파일은 Cygwin 명령어들을 이용하여 쉽게 만들 수 있다.



<그림 79> ls와 gawk를 이용하여 배치 파일 만들기: 원시 말뭉치

<그림 79>에서 명령어 ls와 gawk가 파이프를 연결되어 있다. ls는 해당 디렉토리에 들어 있는 파일들의 목록을 출력하는 프로그램이다. ls가 출력한 파일들의 이름이 하나하나 gawk에 입력으로 들어간다. gawk의 MAIN 부분은 printf라는 하나의 함수로 이루어져 있는데, “uniconv UTF8 %s UCS2 ../raw/%s\n”라는 문자열을 출력하되 %s로 표시된 두 곳에는 입력 레코드 전체(즉 각 파일 이름)를 집어넣으라는 뜻이다. 출력 결과는 다음과 같다.


```
cmd
D:\WearnerWraw_u8>sed -i 's/$/r/' convert.bat
```

<그림 81> 줄바꿈 문자를 Unix식(LF)에서 Dos식(CR-LF)으로 변환하기

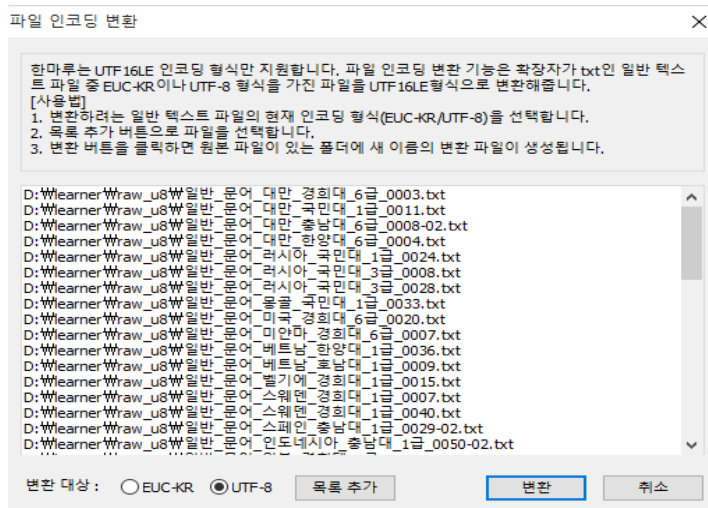
그 다음에 명령프롬프트에서 convert라고 타이핑하고 엔터를 쳐서 배치 파일을 실행하면, 학습자 원시 말뭉치들이 UTF-16LE로 변환되어 d:/learner/raw 디렉토리에 저장된다.

CP949로 되어 있는 학습자 형태분석 말뭉치를 UTF-16LE로 변환하는 방법도 마찬가지로이다(<그림 82>).

```
cmd
D:\WearnerWtagged_cp949>ls | gawk '{ printf( "uniconv CP949 %s UCS2 ../tagged/%s\n", $0, $0 ) }' >convert.bat
```

<그림 82> ls와 gawk를 이용하여 배치 파일 만들기: 형태분석 말뭉치

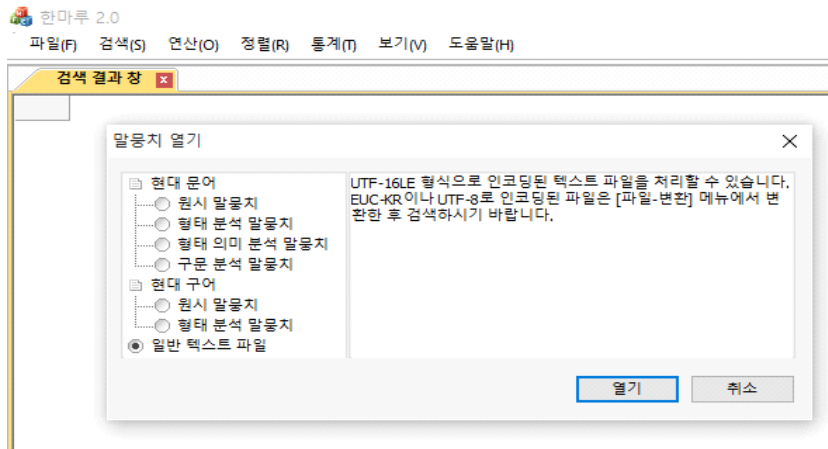
또는 한마루 메뉴바의 ‘파일’ - ‘변환’ 메뉴를 이용할 수도 있다. 원본 파일의 인코딩을 EUC-KR(CP949)와 UTF-8 중 선택하고 ‘목록 추가’ 버튼을 눌러 변환할 원본 파일을 선택한다.(복수의 파일도 선택 가능) 그 다음에 변환 버튼을 누르면 원본 파일이 있는 폴더에 변환된 output 파일이 새로운 이름으로 저장된다.



<그림 83> 한마루 2.0 파일 인코딩 변환

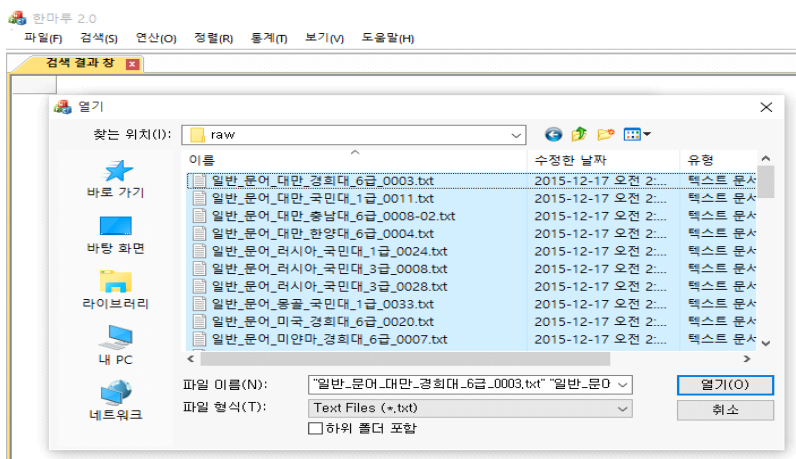
1.3.2. 말뭉치 열기

다운로드한 압축 파일을 풀어 Hanmaru.exe를 실행한 뒤, 메뉴바의 ‘파일’ - ‘열기’ - ‘말뭉치 열기’를 선택하면 다음과 같은 대화상자가 뜨는데 ‘일반 텍스트 파일’을 선택하고 ‘열기’ 버튼을 누른다.



<그림 84> 한마루 2.0의 말뭉치 열기 대화상자

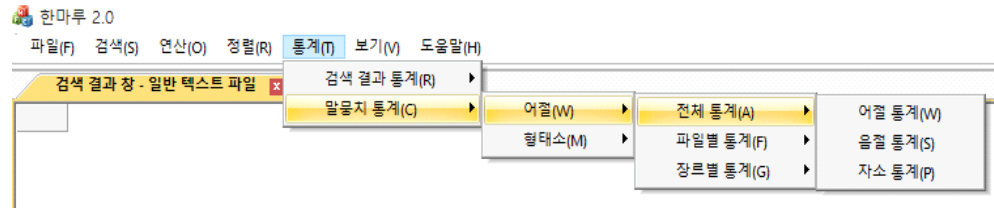
우선 학습자 원시 말뭉치를 UTF-16LE로 변환하여 저장한 파일들을 선택한다. 검색 대상으로 하려는 파일들만 선택하면 된다. 모두 선택하려면 Ctrl-A를 누르고 ‘열기’ 버튼을 클릭한다.



<그림 85> 한마루 2.0 파일 선택 대화상자

1.3.3. 말뭉치 통계

메뉴바의 ‘통계’ - ‘말뭉치 통계’ - ‘어절’ - ‘전체 통계’로 들어가(<그림 86> ‘어절 통계’, ‘음절 통계’, ‘자소 통계’를 각각 내보면 <그림 87>과 같이 된다.



<그림 86> 한마루 2.0의 통계 메뉴

The three screenshots show the '어절 통계' (Word Statistics), '음절 통계' (Syllable Statistics), and '자소 통계' (Character Statistics) windows. Each window displays a table with columns for '어절' (Word), '빈도수(합계: ...)' (Frequency (Total: ...)), and '비율' (Ratio). The '어절 통계' window shows 42 items, the '음절 통계' window shows 42 items, and the '자소 통계' window shows 42 items.

어절	빈도수(합계: 5617)	비율
1 PC방에	2	0.00035606
2 아이들과	1	0.00017803
3 태권도	1	0.00017803
4 부족하는	2	0.00035606
5 줄길	1	0.00017803
6 있어야	2	0.00035606
7 무려	1	0.00017803
8 가깝가	1	0.00017803
9 달려	1	0.00017803
10 생각하게	1	0.00017803
11 인스타그램	1	0.00017803
12 여자친구와	1	0.00017803
13 이어기를	2	0.00035606
14 2015년	2	0.00035606
15 디즈니	2	0.00035606
16 인구	4	0.00071212
17 오이도에서	3	0.00053409
18 또한,	1	0.00017803
19 쉽게	3	0.00053409
20 고칩니다.	1	0.00017803
21 결과가	3	0.00053409
22 대한	13	0.00231440
23 공부에	1	0.00017803
24 여름에	1	0.00017803
25 써서	6	0.00106819
26 노력했다거나	1	0.00017803
27 재미있었요.	1	0.00017803
28 사용해서	2	0.00035606
29 말고	2	0.00035606
30 의견	1	0.00017803
31 분위기에서	1	0.00017803
32 미쳐서	1	0.00017803
33 집에	1	0.00017803
34 가족이라고	2	0.00035606
35 여자와	1	0.00017803
36 없는	15	0.00267046
37 왔던	1	0.00017803
38 산만없지만	1	0.00017803
39 수위의	1	0.00017803
40 불고기를	2	0.00035606
41 글자 </mis>에	1	0.00017803
42 오전	1	0.00017803

음절	빈도수(합계: 18284)	비율
1 툴	24	0.00131262
2 셴	2	0.00010939
3 토	11	0.00060162
4 록	2	0.00010939
5 겐	1	0.00005469
6 이	617	0.03374535
7 뽕	1	0.00005469
8 관	30	0.00164078
9 익	3	0.00016408
10 인	125	0.00683658
11 추	2	0.00010939
12 속	1	0.00005469
13 씨	12	0.00065631
14 께	1	0.00005469
15 일	86	0.00470357
16 알	6	0.00032816
17 퍼	3	0.00016408
18 측	2	0.00010939
19 출	3	0.00016408
20 따	32	0.00175016
21 씌	1	0.00005469
22 면	13	0.00071100
23 섰	1	0.00005469
24 일	1	0.00005469
25 인	6	0.00032816
26 인	21	0.00114855
27 아	67	0.00364441
28 약	2	0.00010939
29 광	5	0.00027346
30 네	6	0.00032816
31 줄	2	0.00010939
32 있	200	0.01093853
33 뽕	17	0.00092977
34 출	1	0.00005469
35 있	1	0.00005469
36 솟	1	0.00005469
37 씨	1	0.00005469
38 솟	1	0.00005469
39 즙	9	0.00049223
40 자	109	0.00596150
41 작	4	0.00021877
42 이	1	0.00005469

자소	빈도수(합계: 39920)	비율
1 ㄱ_초성	2,538	0.06357715
2 ㄴ_초성	69	0.00172846
3 ㄷ_초성	1,240	0.03106212
4 ㄹ_초성	1,538	0.03852705
5 ㄴ_초성	111	0.00278056
6 ㄷ_초성	907	0.02272044
7 ㄹ_초성	787	0.01971443
8 ㅁ_초성	565	0.01415331
9 ㅂ_초성	25	0.00062625
10 ㅅ_초성	1,634	0.04093186
11 ㅆ_초성	39	0.00097695
12 ㅇ_초성	3,693	0.09251002
13 ㅈ_초성	1,259	0.03153808
14 ㅊ_초성	15	0.00037575
15 ㅊ_초성	234	0.00586172
16 ㅋ_초성	53	0.00132766
17 ㅌ_초성	149	0.00373246
18 ㅍ_초성	102	0.00255511
19 ㅎ_초성	1,433	0.03589679
20 ㅈ_중성	4,014	0.10055110
21 ㅊ_중성	870	0.02179359
22 ㅌ_중성	107	0.00268036
23 ㅍ_중성	1	0.00002505
24 ㅈ_중성	1,523	0.03815130
25 ㅊ_중성	841	0.02106713
26 ㅋ_중성	671	0.01680862
27 ㅌ_중성	38	0.00095190
28 ㅍ_중성	1,538	0.03852705
29 ㅈ_중성	330	0.00826653
30 ㅊ_중성	11	0.00027555
31 ㅊ_중성	103	0.00258016
32 ㅆ_중성	256	0.00641283
33 ㅈ_중성	1,090	0.02730461
34 ㅊ_중성	75	0.00187876
35 ㅆ_중성	2	0.00005010
36 ㅈ_중성	59	0.00147796
37 ㅊ_중성	33	0.00082665
38 ㅈ_중성	1,904	0.04769539
39 ㅊ_중성	162	0.00405812
40 ㅈ_중성	2,763	0.06921343
41 ㅊ_중성	677	0.01695892
42 ㅈ_중성	5	0.00012525

<그림 87> 한마루 2.0의 어절·음절·자소 통계

이러한 통계 결과는 메뉴바의 ‘파일’ - ‘저장’ - ‘통계 결과 저장’ 메뉴를 통해 저장할 수 있다.

1.3.4. 용례 기본 검색

메뉴바의 ‘검색’ - ‘어절 검색’을 선택하면 대화상자가 뜨는데, 여기서 ‘기본 검색’을 선택하고, 찾을 문자열(예를 들어 ‘하고’)을 입력하면 다음과 같은 결과가 나온다.

화면 하단에 말뭉치 탑재 소요 시간, 검색어, 검색 대상 말뭉치 크기(문장 수, 어절 수), 검색된 문장·어절 수, 검색 소요 시간이 나온다. 그 위에는 22개의 용례가 좌측(이전) 문맥, 검색어, 우측(이후) 문맥, 파일명, 파일번호의 순으로 제시된다. 이렇게 검색어가 좌우 문맥 사이에 나오는 형식의 콘코던스를 KWIC(keyword in context)이라고 부른다. 앞뒤의 문맥을 한눈에 볼 수 있어 편리하다. 검색 문자열의 앞뒤에 어떤 것들이 주로 나오는지를 콘코던스를 보고 파악할 수 있다.



<그림 88> 한마루 2.0 기본 검색

메뉴바의 ‘파일’ - ‘저장’ - ‘검색 결과 저장’을 선택하여 검색 결과를 파일로 저장할 수 있다. 확장자가 txt인 파일과 view인 파일로 저장되는데, 전자는 문장만 포함한 단순 텍스트 파일이고, 후자는 나중에 한마루에서 ‘파일’ - ‘열기’ - ‘검색 결과 열기’로 다시 열어 볼 수 있는 형태의 파일이다.

1.3.5. 아무개 문자 검색

그런데 <그림 88>의 검색 결과에서 유의할 점은, ‘하고’가 하나의 어절 전체를 이루는 경우만 추출되었고, ‘하고’의 어절의 일부인 경우는 추출되지 않았다는 점이다. 즉 한마루의 ‘어절 검색’의 ‘기본 검색’은 사용자가 입력한 문자열이 온전한 어절을 이루는 경우만 찾아주는 것이다. 입력한 문자열이 어절의 일부인 경우를 찾으려면 ‘어절 검색’의 ‘음절 검색’을 선택하여 아무개 문자(wild card)를 사용하여 한다. 아무개 문자에는 다음의 4가지 종류가 있다.

% : 0개 또는 1개의 문자(음절)

? : 1개의 문자(음절)

* : 0개 이상의 문자(음절)

+ : 1개 이상의 문자(음절)

예컨대 ‘*하*’는 어절 속에 ‘하’라는 음절이 들어 있는데, 그 앞이나 뒤에 임의의 길이의 문자열이 있을 수도 있고 없을 수도 있다는 뜻이 된다. 검색 결과는 다음과 같다. 267개 문장에서 371개의 용례가 추출되었다.

검색 결과 및 일반 텍스트 편집		어절 목록	검색어	이후 문맥	좌우문맥
1		타인서 이말한 이기두유를	해결하기	유해 정보와 사생활을 같이 이야기를 해야 한다.	문어, 일본, 속이다. 6.62, .00
2			변반하고	실은 경우에도 한 번 상대방의 의견을	문어, 일본, 권한다. 6.62, .00
3			변반하고	실은 경우에도 한 번 상대방의 의견을	문어, 일본, 권한다. 6.62, .00
4		변반하고 실은 경우에도 한 번 상대방의 의견을	나서 자기의 의견을 말하면 좀 더 서로가 기분이 좋아질 수가 있다.	문어, 일본, 권한다. 6.62, .00	
5		이런 사실 문제를	개선하려면	좀 더 서로가 기분이 좋아질 수가 있다.	문어, 일본, 권한다. 6.62, .00
6		이말한 참으로서 군교에 사는 주민에는 다음 지역보다 세금을 적게	하는	무엇을 할바론 인터넷 사용 태도를 취해야 한다.	문어, 일본, 권한다. 6.62, .00
7		직접 만나지 않으므로 해의를 자치지 않고 침부로	말하는	것이 좋은 방법이라고 할 수 있다.	문어, 일본, 속이다. 6.62, .00
8		그러므로 인터넷에 정보를 올리고 싶다면 그 정보가 정확한 것인지 틀은 것인지 먼저 확인해야	하는	사람이 많다.	문어, 일본, 권한다. 6.62, .00
9		그로 인해 몸속 일일생물에서 말할 수 없는 말	특히는	필요가 있다.	문어, 일본, 권한다. 6.62, .00
10		이런 사람들이 대부분 남의 의견을	종종하지	말 바꾸는 말은 흔히 볼 수 있다.	문어, 일본, 권한다. 6.62, .00
11		앞으로도 여러 곳에서 이런 문제가 생길까 생각 하지만 자신만	생각하지	알고 자신의 의견과 조금씩 다르면 바로 견해를 사용해서 다른 사람에게 심하게 준다.	문어, 일본, 속이다. 6.62, .00
12		앞으로도 여러 곳에서 이런 문제가 생길까 생각 하지만 알고 세계, 지구환경까지 생각하게 되면 세상 좋은 방법으로 해결할 것이라고 생각한다.	생각하게	되고 세계, 지구환경까지 생각하게 되면 세상 좋은 방법으로 해결할 것이라고 생각한다.	문어, 일본, 속이다. 6.62, .00
13		어디나든 자주 상담에	기다하러	되면 제일 좋은 방법으로 해결할 것이라고 생각한다.	문어, 일본, 속이다. 6.62, .00
14		그래서 상대방의 의견	종종하는	게 후회가 커져야 가장 기본적인 태도이라고 생각한다.	문어, 일본, 권한다. 6.62, .00
15			왜냐하면	인터넷을 지금까지 PC나 인터넷도 휴대전화로 간편히 사용할 수 있어서 아직 내이가 어떤 초고속생도 인터넷을	문어, 일본, 권한다. 6.62, .00
16		어디에서 아직 내이가 어떤 초고속생도 인터넷을 사용할 기회가 충분히 있고, 인터넷을 통해서 발생한 사건도	종종하고	있기 때문이다.	문어, 일본, 권한다. 6.62, .00
17		상대방의 의견이 다음 때 그 의견을	인정하여	남자에게 자신의 의견을 표현하는 게 상대방을 존중하는 방식이다.	문어, 일본, 권한다. 6.62, .00
18		상대방의 의견이 다음 때 그 의견을 인정하여	남성에게	자신의 의견을 표현하는 게 상대방을 존중하는 방식이다.	문어, 일본, 권한다. 6.62, .00
19		상대방의 의견이 다음 때 그 의견을 인정하여 남성에게 자신의 의견을	표현하여	게 상대방을 존중하는 방식이다.	문어, 일본, 권한다. 6.62, .00
20		상대방의 의견이 다음 때 그 의견을 인정하여 남성에게 자신의 의견을 표현하는 게 상대방을	존중하는	방식이다.	문어, 일본, 권한다. 6.62, .00
21		부서 가장 중요한 것이 무엇인지 20, 30대 남자와 40, 50대 남자에게 조사한 결과 20, 30대 남자는 가족이 가장	중요하다고	나타났다.	문어, 일본, 속이다. 6.62, .00
22			중요하다고	생각하는 젊은이들은 인터넷 사용 태도는 먼저 타인의 사생활 보호하기이다.	문어, 일본, 권한다. 6.62, .00
23		이론 세상에서 중요하다고	생각하는	젊은이들은 인터넷 사용 태도는 먼저 타인의 사생활 보호하기이다.	문어, 일본, 권한다. 6.62, .00
24		이론 세상에서 중요하다고 생각하는 젊은이들은 인터넷 사용 태도는 먼저 타인의 사생활	보호하기이다		문어, 일본, 권한다. 6.62, .00
25			그러고	같이 행동해 갔다.	문어, 중국, 국인다. 6.62, .00
26		또한, 의견이 원리도 객관 여부를 사용하지 않고 이성적인 방식으로 바른 언어를	사용하지	말고 이성적인 방식으로 바른 언어를 사용하면 좋겠다.	문어, 일본, 권한다. 6.62, .00
27		또한, 의견이 원리도 객관 여부를 사용하지 않고 이성적인 방식으로 바른 언어를	사용하면	좋겠다.	문어, 일본, 권한다. 6.62, .00
28		그 반면 40, 50대 남자는 돈이 가장	중요하다고	여자는 가족이 가장 중요하다고 답하였다.	문어, 일본, 권한다. 6.62, .00
29		그 반면 40, 50대 남자는 돈이 가장 중요하다고 여자는 가족이 가장	중요하다고	답하였다.	문어, 일본, 속이다. 6.62, .00
30		그 반면 40, 50대 남자는 돈이 가장 중요하다고 여자는 가족이 가장 중요하다고	답하였다	.	문어, 일본, 속이다. 6.62, .00
31			명동에서	문동화를 입는다.	문어, 중국, 국인다. 6.62, .00
32			사건하고	같이 일정한 기간 정보를 보고 스포츠 행위를 할 사람들도 있기 때문에 인터넷 상에 있는 타인의 사생활을 보호	문어, 일본, 권한다. 6.62, .00
33			다시	시내에 복귀를 하기 위해 우리는 올바른 인터넷 사용 태도를 취해야 한다.	문어, 일본, 권한다. 6.62, .00
34			생각하기에는	오늘의 20, 30대 남자는 별로 돈이나 탐욕에 대해 관심이 없는 것 같다.	문어, 일본, 속이다. 6.62, .00
35			다음은 의견 차이를	인정하기이다	문어, 일본, 권한다. 6.62, .00
36			그게 바로 상대방의 의견을	존중하고	문어, 일본, 권한다. 6.62, .00
37			표현하는	것이다.	문어, 일본, 권한다. 6.62, .00
38		인터넷 상에서는 말로 할 때보다 자기 생각을 쉽게 보여 줄 수 있지만 의견은 사람마다 다르기 때문에	차이하고	의견이 다른 사람이 있어도 어느 정도 인정해야 한다.	문어, 일본, 권한다. 6.62, .00

결과 및 일반 텍스트 편집

검색어 : 한*

검색 결과 항목들 크기 : 전체 문장(756문장), 전체 (애니.65178)

검색된 문장 및 (애니 개수 : 261문장(371개))

문장 길이 범위 : 0.000000 ~ 99.999999

문장 길이 범위 : 0.000000 ~ 99.999999

<그림 89> 한마루 2.0 아무개 문자 검색 결과

1.3.6. 자소 검색

초성·중성·종성의 세 부분으로 이루어져 있는 한글의 특성을 반영하여, 한마루는 자소 단위 검색 기능을 제공한다. 메뉴바의 ‘검색’ - ‘어절 검색’ - ‘자소 검색’을 선택하여 이용할 수 있다. 예컨대 “[??.ㅅ]다”를 검색하면 ‘다’라는 음절 앞에 초성·중성은 아무거나 팬잡고 종성은 ‘ㅅ’인 음절이 있는 경우만을 추출해 준다. 학습자 말뭉치에서는 56개의 용례가 추출되었다.

검색 결과 창 - 일반 텍스트 파일		이전 문맥	검색어	이후 문맥
1	이것은 바로 님비 현상의 현상이라고 할 수	있다	.	
2	요즘 인터넷 폭력으로 인한 사회 문제가 날마다 늘어나고	있다	.	
3	실은 경우에도 한 번 상대방의 의견을 인정하고 나서 자기의 의견을 말하면 좀 더 서로가 기분이 좋아질 수가	있다	.	
4	해결방법의 경우 세금의 감소를 할 수	있다	.	
5	이렇한 혐오시설 근교에 사는 주민에는 다들 지역보다 세금을 적게 하는 것이 좋은 방법이라고 할 수	있다	.	
6	그러므로 인터넷에 정보를 올리고 싶다면 그 정보가 정확한 것인지 틀린 것인지 먼저 확인해야 하는 필요가	있다	.	
7	그로 인해 평소 일상생활에서 말할 수 없는 말, 혹은 할, 배우는 말 등을 흔히 볼 수	있다	.	
8	경제·문화론 인터넷 사용 태도'라는 것은 어떤 것인가 생각할 필요가 있음 많아지고	있다	.	
9	나도 이런 조사를 받으면 아마 가장 중요한 것은 가족을 행복할 수	있다	.	
10	요즘 일본에서도 이런 현상이	있다	.	
11	가치관은 많은 구분으로 나눌 수	있다	.	
12	그것 때문에 회사나 기업에서 신입사원을 채용할 때 이력서 사진을 중시하는 경향이	있다	.	일
13	이 결과를 통해 남자의 인생관이 세대에 따라 변화해 온 것을 알 수	있다	.	
14	람도 인사할 수 없지만 태도나 성격도 좋지않은 시황을 보는 기회도 없는 사람도 많은 경향이	있다	.	일
15	그리고 인구 집중 때문에 대도시가 시골보다 소인원을 많이 들을 수	있다	.	일
16	첫째, 수도의 정치사업은 다른 지방으로 이상해서 다른 지방을 개발할 수 있고 수도의 인구 집중을 개선할 수	있다	.	일
17	청부는 진화경으로 공원이나 나무를 많이 지었어 시민들이 좋은 환경이 행복하게 살 수	있다	.	일
18	현대 도시는 청부와 시민들이 같이 노력하고 이러한 문제점들이 해결될 수	있다	.	
19	현대 세계적으로 환경문제가 대두되고	있다	.	
20	그래서 최근에 세계적으로 환경보호하는 운동이 많아지고	있다	.	
21	예를 들면 물을 아껴 쓰는 것도 있고 각 나라의 대표적인 건물이 한 시간 동안 불 끄는 것도	있다	.	
22	선진국들이 환경보호를 중시하는데 개발도상국이 경제 발전이라는 미명아래 환경을 계속 파괴하고	있다	.	
23	만 20<punctuation>가문밋집</punctuation>30대는 인생에서 주변 사람들이 중요하다고 생각하는 것을 알 수	있다	.	
24	최근 활다에서 비롯된 자살이 사회적인 문제가 되고	있다	.	
25	같이라고 차지했지만 생각 대해서 20<punctuation>가문밋집</punctuation>30대와 차이가 있는 것을 알려주고	있다	.	
26	<punctuation>가문밋집</punctuation>30대에 2위로 차지한 친구보다 전집이 중요하다고 생각하는 것을 알 수	있다	.	
27	들수록 자신의 책임감이나 자신의 생각 자신의 의견이 달라지는 법하고 변하는 다양한 것이라는 사실을 알 수	있다	.	
28	그것에서 벗어나고 편해지고 싶다는 마음으로 활다를 시키는 경향이	있다	.	
29	나는 대련에 여행할 가	했다	.	
30	물론 당하기 쉬운 사람은	있다	.	
31	나는 친구와 같이 대련에 가	했다	.	
32	주위 사람들이 활다 극복은 피해자만 아니라 가해자에게도 얼마나 관심을 가질 수 있는지에 달려	있다	.	
33	그래서 수업을 못	갔다	.	
34	우리는 놀이공원에 가	했다	.	
35	요즘 비행 혐소년들에서 기인한 문제가 많이 나타나고	있다	.	
36	예를 들면 폭력이나 절도 등이	있다	.	
37	요즘은 좋은 학교를 졸업하거나 좋은 성적을 받아야 성공할 수	있다	.	
38	그리고 공부에 대한 의욕도 살아날 수도	있다	.	

장제 정보
 검색어: [??.ㅅ]다
 검색 대상 말뭉치 크기: 전체 문장(756문장), 전체 어절(5,617개)
 검색된 문장 및 어절 개수: 56문장(56개)
 말뭉치 검색: 0.016000초 걸렸습니다.
 검색 결과 출력: 0.016000초 걸렸습니다.

<그림 90> 한마루 2.0의 자소 검색 결과

1.3.7. 범위 검색

두 어절이 일정한 범위 내에서 함께 나타난 경우를 찾아 주는 기능이다. 검색식은 다음과 같은 구조로 되어 있다.

기준 어절+공백+숫자(좌측 범위값)@숫자(우측 범위값)+공백+검색어 어절

기준 어절을 중심으로 하여 검색어 어절이 두 숫자 사이의 범위에 나타나는 경우를 찾아 준다. 기준 어절과 검색어 어절에 아무개 문자를 이용할 수도 있다. 예컨대 “수 1@3 있*”이라고 입력하면 ‘수’라는 기준 어절을 중심으로 ‘있*’(‘있’ 뒤

에 임의의 문자열이 있는 어절)이 바로 오른쪽에 있거나 오른쪽 2번째 어절에 있거나 오른쪽 3번째 어절에 있는 경우를 찾아 준다.

검색 결과창 - 일반 텍스트 파일				
	이전 문맥	검색어	이후 문맥	
1	이것은 바로 넓어 현상의 현상이라고 할	수 있다	.	
2	얼굴을 볼	수 없는 쪽이 있기	.	때문에 직접 버전을 만들 때보다 상처가 더 많은 것이다.
3	해결방법의 경우 세금의 감소를 통	수 있다	.	
4	이렇한 철도시설 근교에 사는 주민들은 다음 지옥보다 세금을 적게 하는 것이 좋은 방법이라고 할	수 있다	.	
5	그로 인해 철도 시설상황에서 말할 수 없는 말, 육하는 말, 바꾸는 말, 줄 줄 줄 줄 줄	수 있다	.	
6	왜냐하면 인터넷은 지금이야 PC나 인터넷도 휴대전화로 가능한 사용법	수 있어서	.	아직 나이가 어린 초등학교도 인터넷을 사용법 기회가 충분하고, 인터넷을 통해서 발생한 사건도 증가하고, 방법이라고 생각한다.
7	그래야 부활요한 사도와 사탄 사건을 풀일	수 있다	.	
8	나도 이런 조사를 받으면 아마 가장 중요한 것은 가족을 풀을	수 있다	.	
9	인터넷 상에서는 말로 할 때보다 자기 생각을 쉽게 보여 줄	수 있지만	.	외관은 사람마다 다르기 때문에 자신하고 의견이 다른 사람이 있어도 어느 정도 인정해야 한다.
10	정확하지 않은 정보를 인터넷 상에 올리면 잘못된 정보가 확산되고 그것은 본 사람들에겐 피해를 줄	수 있고	.	외출할 수 없는 사태를 일으킬 수도 모르고 때문이다.
11	커뮤니티는 법에도 적시할	수 있습니다	.	
12	고향에서 고향친구도 만날	수 있습니다	.	
13	보무님도 만날	수 있습니다	.	
14	가치관만 많은 구문으로 나눌	수 있다	.	
15	이 결과를 통해 남자의 인생관이 세대에 따라 변화해 온 것을 알	수 있다	.	
16	그리고 인구 집중 때문에 대도시가 서울보다 소인을 많이 풀을	수 있다	.	
17	사람은 목표가 있으면 노력하는 것에 대해서 할일	수 있다는 생각이 있으니까	.	
18	이런한 문제들이 어떻게 해결할	수 있을까요	?	
19	첫째, 수도의 정치사범은 다른 지방으로 이상해서 다른 지방을 개발할	수 있고	.	수도의 인구 집중은 개선할 수 있다.
20	첫째, 수도의 정치사범은 다른 지방으로 이상해서 다른 지방을 개발할	수 있다	.	
21	정부는 전술중으로 공헌이나 나무를 많이 지었어 시민들이 좋은 물량이 풍부하게 살	수 있다	.	
22	현대 도시는 경제와 시민들이 많이 노력하고 이러한 문제점들이 해결될	수 있다	.	
23	보면 20<punctuation>가문어점 <punctuation>30대는 인생에서 주변 사람들이 중요하다고 생각하는 것을 알	수 있다	.	
24	20<punctuation>가문어점 <punctuation>30대 2위로 차지한 친구보다 전기는 중요하다고 생각하는 것을 알	수 있다	.	
25	먹을수록 자신의 책임감이나 자신의 생각 자신의 여건에 따라지는 변화는 다양한 것이라는 사실을 알	수 있다	.	
26	자신의 인생이 어떻게 보면 행복해지는지 풀일	수 있는지에	.	대해서는 사람마다 다르지만 다시 자신의 인생에 대해서 생각할 필요가 있다고 느꼈다.
27	의 환인을 살펴 보았는데 월이는 학교나 가정에서 관심과 사랑이 부족한 데에서 비롯된 현상이라는 것을 알	수 있었다	.	
28	주위 사람들이 왔다 국력은 민해지만 아니라 가해자에게도 얼마나 관심을 가질	수 있는데도 말려 있다	.	
29	요즘은 좋은 학교를 졸업하거나 좋은 성적을 받아야 성공할	수 있다	.	
30	지 환경을 의형시킨 것은 주로 선진국 때문이고 우리도 경제 개발을 해야 한다는 개발도상국의 의견도 이해할	수 있지만	.	선진국이나 개발도상국이나 그런 것에 구애받지 않고 인간이라는 하나의 공통체로서 어떻게 해야 하는지 같
31	열심히 공부해야 좋은 대학교에서 입학할	수 있다	.	
32	그리고 후손을 위해서도 어둡게 하면 인간과 지구가 오래 갈 어둡힐	수 있는지	.	심사숙고해야 한다.
33	이 먼저 환경에 대한 대책을 해야 되고 개발도상국은 환경을污染시키게 아니라 환경도 유지하면서 개발할	수 있는	.	"지속 가능한 개발"을 해야 한다.
34	그것은 인간의 유리가 지금까지 편하게 쉽게 생활할	수 있게	.	자연물 소중하게 생각하지 않고 우리 손으로 그렇게 만드는 것이다.
35	를 늦은 감이 있지만 그래도 알아들	수 있어서	.	대행이다.
36	하지만 선진국이 그렇게 만들었으니까 선진국이 환경을 보호하기 위해 힘을 써야 되고 개발도상국도 알 살	수 있게	.	도약해야 한다고 생각한다.
37	이렇게 살으면 정말 행복할	수 있다	.	
38	그리고 선진국은 이미 개발되었고 높은 기층이 있으니까 그 기층을 몰음하면 환경 보호하면서 경제 개발도 할	수 있을	.	것이다.

<그림 91> 한마루 2.0 범위 검색 결과

1.3.8. 불리언 검색

2개의 논리 연산자 | (or) 및 & (and)를 이용하여 검색 조건을 결합하여 찾아 주는 기능이다. |는 3개 이상의 항을 연결할 수 있으나, &는 2개의 항을 연결하는 데만 쓸 수 있다. 그리고 |로 연결된 항들은 순서와 상관없으나, &로 연결된 항들은 제시된 순서대로 나오는 것만 찾아 준다. 각 항에 아무개 문자를 쓸 수 있다. “매우|아주|너무|가장+좋*”이라고 하면 4개의 정도 부사와 ‘좋-’이 공기한 용례들을 찾아 준다.

검색 결과 창 - 일반 텍스트 파일					
	이전 문맥	검색어	이후 문맥	파일명	일 번
1	한국 생활이	너무	재미있는 바람습니다.	일반_문어_중국_경희대_1급_0069.txt	46
2		아주	바람습니다.	일반_문어_베트남_한양대_1급_0036.txt	10
3	통해서 사생활을 쉽게 세상에 공개해 버린 사람들이	아주	많다.	일반_문어_일본_경희대_6급_0019.txt	23
4	가족들	너무	보고 싶습니다.	일반_문어_베트남_한양대_1급_0036.txt	10
5	저는 특히 출대에	너무	좋아합니다.	일반_문어_대만_국민대_1급_0011.txt	1
6		너무	힘들었었어요.	일반_문어_중국_국민대_1급_0007.txt	48
7		너무	피곤해요.	일반_문어_중국_국민대_1급_0007.txt	48
8		너무	실심했습니다.	일반_문어_중국_국민대_1급_0007.txt	48
9	요즘은 외모가 사람을 평가할 때	매우	중요하다고 생각하는 '외모지상주의'가 익숙한 말이 되었다.	일반_문어_일본_총남대_5급_0072-02.txt	25
10	그래서 주말에	너무	중심입니다.	일반_문어_일본_한양대_1급_0037.txt	26
11	한국의 발수가	너무	맛있습니다.	일반_문어_일본_한양대_1급_0037.txt	26
12	저는 아이들이	너무	좋아해서 한국의 아이들과 같이 놓고 싶습니다.	일반_문어_일본_한양대_1급_0037.txt	26
13	저는 웃하고 화장품	아주	좋아해요.	일반_문어_스웨덴_경희대_1급_0040.txt	14
14	대련의 바다	너무	크다.	일반_문어_중국_국민대_2급_0053.txt	51
15		아주	오랜만이었어요.	일반_문어_스웨덴_경희대_1급_0040.txt	14
16	제 여자친구는	아주	예뻐요.	일반_문어_스웨덴_총남대_1급_0029-02.txt	15
17	대련의 사람들	너무	좋다.	일반_문어_중국_국민대_2급_0053.txt	51
18	이번 여름에 저는 한국어를	아주	배우고 싶어요.	일반_문어_스웨덴_총남대_1급_0029-02.txt	15
19	한국 음식이	아주	맛있습니다.	일반_문어_중국_경희대_1급_0054.txt	40
20	구경할 때 기분이	너무	좋아.	일반_문어_중국_국민대_2급_0054.txt	52
21	맛있기 때문에	너무	비싸요.	일반_문어_중국_국민대_2급_0054.txt	52
22	저는 등산하고 산책이	아주	좋아합니다.	일반_문어_일본_경희대_1급_0003.txt	17
23	그리고 제주도에서는 굽이	아주	유연합니다.	일반_문어_일본_경희대_1급_0003.txt	17
24	일본도 생선이	아주	맛있습니다.	일반_문어_일본_경희대_1급_0003.txt	17
25		아주	좋아합니다.	일반_문어_일본_경희대_1급_0003.txt	17
26	그리고 여행이	너무	좋아합니다.	일반_문어_일본_경희대_1급_0028.txt	19
27	한국 생활이	너무	재미있습니다.	일반_문어_중국_경희대_1급_0058.txt	43
28	저는 한국이	아주	좋아해요.	일반_문어_일본_경희대_1급_0035.txt	20
29		너무	재미있습니다.	일반_문어_중국_경희대_1급_0069.txt	46
30		아주	맛있습니다.	일반_문어_중국_경희대_1급_0069.txt	46

상세 정보
 검색어: 매우|아주|너무|가장&중+
 검색 대상 말뭉치 크기: 전체 문장(758문장), 전체 어절(5,617개)
 검색된 문장 및 어절 개수: 30문장(30개)
 말뭉치 검색: 0.015000초 걸렸습니다.
 검색 결과 출력: 0.016000초 걸렸습니다.

<그림 92> 한마루 2.0 불리언 검색 결과

1.3.9. 형태 분석 말뭉치

학습자 형태 분석 말뭉치는 형태 분석 말뭉치이기는 하지만 세종 형태 분석 말뭉치의 포맷을 그대로 따르지는 않았다. 한마루는 세종 형태 분석 말뭉치의 포맷을 그대로 따른 말뭉치에 대해서는 ‘검색’ 메뉴에서 ‘형태소 검색’이라는 특화된 검색 기능을 제공하나, 학습자 형태 분석 말뭉치는 포맷이 다르기 때문에 아쉽게도 이 기능을 이용할 수 없다. 따라서 말뭉치 파일을 열 때, ‘현대 문어/구어 형태 분석 말뭉치’가 아니라 ‘일반 텍스트 파일’을 선택해야 한다.

그렇기는 하지만 형태소 단위로 품사가 부착되어 있다는 점을 활용하면 원시 말뭉치에 비해 훨씬 더 정교한 검색 또는 훨씬 더 포괄적인 검색이 가능하다. 더 정교한 검색의 예로는 지정사 ‘-이-’의 검색을 들 수 있다. 그냥 문자열로서의 ‘이’는 주격 조사도 있고 지정사도 있고 그 외에 다른 형태소의 일부인 경우도 있다. 품사가 부착되어 있지 않다면 이들을 구별하여 검색할 수가 없는 것이다. 형태 분석 말뭉치에서는 ‘*이/VCP*’이라고 함으로써 지정사 ‘-이-’만을 검색할 수 있다(<그림 93>).

검색 결과 창 - 일반 텍스트 파일						
	이전 문맥	검색어	이후 문맥	파일명	파일 번호	인덱스 번호
1	것이다.°	것/NNB+이/VCP	+다/EF+./SF	일반.문어_대만_한양대_6급_0004.pos	3	521
2	보인다.°	보이/VW	+나.다/EF+./SF	일반.문어_일본_경희대_6급_0018.pos	22	2,085
3	것이라고°	것/NNB+이/VCP	+라고/EC	일반.문어_일본_총남대_5급_0072-02.pos	25	2,609
4	마련이다.°	마련/NNB+이/VCP	+다/EF+./SF	일반.문어_일본_경희대_6급_0018.pos	22	2,094
5	폭력이라고°	폭력/NNG+이/VCP	+라고/EC	일반.문어_대만_경희대_6급_0003.pos	0	15
6	것이지만°	것/NNB+이/VCP	+지만/EC	일반.문어_일본_한양대_6급_0006.pos	29	3,143
7	것이다.°	것/NNB+이/VCP	+다/EF+./SF	일반.문어_일본_홍익대_6급_0014.pos	36	4,190
8	납비현상이라고°	납비/NNG+현상/NNG+이/VCP	+라고/EC	일반.문어_일본_홍익대_6급_0001.pos	33	3,671
9	것이다.°	것/NNB+이/VCP	+다/EF+./SF	일반.문어_일본_경희대_6급_0018.pos	22	2,114
10	사장이니다.°	사장/NNG+이/VCP	+습니다/EF+./SF	일반.문어_중국_경희대_1급_0053.pos	39	4,721
11	가치관이라고°	가치관/NNG+이/VCP	+라고/EC	일반.문어_일본_홍익대_6급_0014.pos	36	4,202
12	때문이다°	때문/NNB+이/VCP	+다/EF	일반.문어_일본_한양대_6급_0006.pos	29	3,161
13	요소이다.°	요소/NNG+이/VCP	+다/EF+./SF	일반.문어_일본_경희대_6급_0018.pos	22	2,125
14	생각이라면°	생각/NNG+이/VCP	+라면/EC	일반.문어_일본_홍익대_6급_0014.pos	36	4,214
15	오랜만이었어요°	오랜만/NNG+이/VCP	+있/EP+어요/EF	일반.문어_스웨덴_경희대_1급_0040.pos	14	1,609
16	예외이다.°	예외/NNG+이/VCP	+다/EF+./SF	일반.문어_미국_경희대_6급_0020.pos	8	1,088
17	생각이라면°	생각/NNG+이/VCP	+라면/EC	일반.문어_일본_홍익대_6급_0014.pos	36	4,218
18	피근네요.°	피근/NA+이/VCP	+네요/EF+./SF	일반.문어_중국_국민대_1급_0007.pos	48	5,260
19	직업관이라고°	직업관/NNG+이/VCP	+라고/EC	일반.문어_일본_홍익대_6급_0014.pos	36	4,219
20	인타넷이라는°	인타넷/NA+이/VCP	+라는/ETM	일반.문어_일본_경희대_6급_0018.pos	22	2,135
21	누군지도°	누구/NP+이/VCP	+나.지/EC+도/JX	일반.문어_일본_경희대_6급_0018.pos	22	2,137
22	혼자입니다.°	혼자/NNG+이/VCP	+습니다/EF+./SF	일반.문어_중국_국민대_1급_0007.pos	48	5,269
23	것이다.°	것/NNB+이/VCP	+다/EF+./SF	일반.문어_일본_한양대_6급_0007.pos	30	3,187
24	겁니다.°	것/NNB+이/VCP	+습니다/EF+./SF	일반.문어_러시아_국민대_1급_0024.pos	4	587
25	무엇인지°	무엇/NP+이/VCP	+나.지/EC	일반.문어_일본_홍익대_6급_0014.pos	36	4,237
26	무엇인지°	무엇/NP+이/VCP	+나.지/EC	일반.문어_일본_홍익대_6급_0014.pos	36	4,246
27	폭력)이라는°	폭력/NNG+/SS+/SS+이/VCP	+라는/ETM	일반.문어_미국_경희대_6급_0020.pos	8	1,120
28	사회적인°	사회적/NNG+이/VCP	+나.지/ETM	일반.문어_일본_한양대_5급_0004.pos	27	2,686
29	것이다.°	것/NNB+이/VCP	+다/EF+./SF	일반.문어_일본_경희대_6급_0018.pos	22	2,165
30	기본적인°	기본적/NNG+이/VCP	+나.지/ETM	일반.문어_대만_경희대_6급_0003.pos	0	82
31	태도이라고°	태도/NNG+이/VCP	+라고/EC	일반.문어_대만_경희대_6급_0003.pos	0	83
32	것이다.°	것/NNB+이/VCP	+다/EF+./SF	일반.문어_일본_한양대_6급_0007.pos	30	3,211
33	세계적인°	세계적/NNG+이/VCP	+나.지/ETM	일반.문어_일본_한양대_5급_0004.pos	27	2,692
34	문재인°	문재/NNG+이/VCP	+나.지/ETM	일반.문어_일본_한양대_5급_0004.pos	27	2,693
35	돈이라는°	돈/NNG+이/VCP	+라는/ETM	일반.문어_일본_홍익대_6급_0014.pos	36	4,257
36	대행이다.°	대행/NNG+이/VCP	+다/EF+./SF	일반.문어_일본_한양대_6급_0007.pos	30	3,220
37	거고°	거/NNB+이/VCP	+고/EC	일반.문어_중국_경희대_1급_0054.pos	40	4,784
38	보이지°	보이/VW	+지/EC	일반.문어_일본_홍익대_6급_0014.pos	36	4,263
39	무역개°	무역/NP+이/VCP	+가.자/EF+2/SF	일반.문어_일본_한양대_5급_0004.pos	27	2,700
상세 정보						
검색어: *이/V+ 검색 대상 말뭉치 크기: 전체 문장(5,628문장), 전체 어절(5,628개) 검색된 문장 및 어절 개수: 190문장(190개) 말뭉치 검색: 0.031000초 걸렸습니다. 검색 결과 출력: 0.031000초 걸렸습니다.						

<그림 93> 한마루 2.0으로 형태분석 말뭉치에서 지정사를 검색한 결과

더 포괄적인 검색의 예로는 품사 태그를 이용한 검색을 들 수 있다. 예를 들어 입력식을 ‘*/JKO’와 같이 아무개 문자를 이용하여 입력하면, 대격 조사로 끝나는 모든 어절을 다 추출해 준다(<그림 94>). 그 대격 조사가 ‘을’이든 ‘를’이든 ‘르’이든 또 다른 형태이든 상관하지 않는 것이다.

검색 결과 창 - 일반 텍스트 파일						
	이전 문맥	검색어	이후 문맥	파일명	파일 번호	인덱스 번호
1	영화를	영화/NNG+를/JKO		일반_문어_중국_경희대_1급_0053.pos	39	4,690
2	의견을	의견/NNG+을/JKO		일반_문어_미국_경희대_6급_0020.pos	8	1,043
3	밥을	밥/NNG+을/JKO		일반_문어_중국_국민대_1급_0004.pos	47	5,212
4	감정을	감정/NNG+을/JKO		일반_문어_일본_경희대_6급_0018.pos	22	2,086
5	발전소를	발전소/NNG+를/JKO		일반_문어_일본_홍익대_6급_0001.pos	33	3,649
6	사생활을	사생활/NNG+을/JKO		일반_문어_미국_경희대_6급_0020.pos	8	1,049
7	컴퓨터를	컴퓨터/NNG+를/JKO		일반_문어_중국_국민대_1급_0004.pos	47	5,217
8	음식을	음식/NNG+을/JKO		일반_문어_중국_경희대_1급_0053.pos	39	4,697
9	계발*을	계발/NNG+*/SS+을/JKO		일반_문어_일본_한양대_6급_0006.pos	29	3,134
10	쇼핑을	쇼핑/NNG+을/JKO		일반_문어_일본_한양대_1급_0037.pos	26	2,613
11	참을	참/NA+을/JKO		일반_문어_중국_국민대_1급_0004.pos	47	5,221
12	의견을	의견/NNG+을/JKO		일반_문어_일본_경희대_6급_0018.pos	22	2,096
13	대책을	대책/NNG+을/JKO		일반_문어_일본_한양대_6급_0006.pos	29	3,139
14	경험을	경험/NNG+을/JKO		일반_문어_일본_홍익대_6급_0014.pos	36	4,183
15	쇼핑을	쇼핑/NNG+을/JKO		일반_문어_일본_한양대_1급_0037.pos	26	2,620
16	친구를	친구/NNG+를/JKO		일반_문어_중국_경희대_1급_0053.pos	39	4,705
17	패션을	패션/NNG+을/JKO		일반_문어_스웨덴_경희대_1급_0040.pos	14	1,580
18	텔레비전을	텔레비전/NNG+을/JKO		일반_문어_중국_경희대_1급_0053.pos	39	4,708
19	심리를	심리/NNG+를/JKO		일반_문어_일본_홍익대_6급_0001.pos	33	3,666
20	문제를	문제/NNG+를/JKO		일반_문어_대만_경희대_6급_0003.pos	0	19
21	환경을	환경/NNG+을/JKO		일반_문어_일본_한양대_6급_0006.pos	29	3,146
22	상대방을	상대방/NNG+을/JKO		일반_문어_일본_경희대_6급_0018.pos	22	2,106
23	문화를	문화/NNG+를/JKO		일반_문어_중국_경희대_1급_0053.pos	39	4,712
24	지구환경을	지구/NNG+환경/NNG+을/JKO		일반_문어_일본_한양대_6급_0006.pos	29	3,150
25	태도를	태도/NNG+를/JKO		일반_문어_대만_경희대_6급_0003.pos	0	25
26	인터넷을	인터넷/NNG+을/JKO		일반_문어_일본_경희대_6급_0018.pos	22	2,110
27	밥을	밥/NNG+을/JKO		일반_문어_중국_국민대_1급_0007.pos	48	5,236
28	정보를	정보/NNG+를/JKO		일반_문어_미국_경희대_6급_0020.pos	8	1,072
29	남기를	남/VV+기/ETN+를/JKO		일반_문어_대만_한양대_6급_0004.pos	3	552
30	예의를	예의/NNG+를/JKO		일반_문어_대만_경희대_6급_0003.pos	0	31

상세 정보

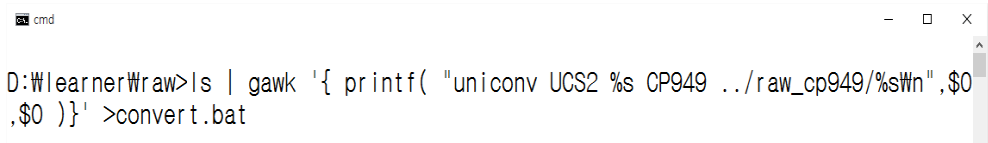
말뭉치 열기: 0.016000초 걸렸습니다.
 검색어: */JKO
 검색 대상 말뭉치 크기: 전체 문장(5,628문장), 전체 어절(5,628개)
 검색된 문장 및 어절 개수: 455문장(455개)
 말뭉치 검색: 0.062000초 걸렸습니다.
 검색 결과 출력: 0.031000초 걸렸습니다.

<그림 94> 한마루 2.0으로 형태분석 말뭉치에서 품사 태그 및 아무개 문자를 이용하여
 검색한 결과

1.4. 지능형 형태소 분석기를 이용한 말뭉치 활용

학습자 말뭉치는 원시 말뭉치와 형태 분석 말뭉치가 함께 결과물로서 제공되기 때문에, 형태소 분석기를 이용할 필요는 그리 많지 않다. 그러나 어떤 이유에서든 학습자 원시 말뭉치에 자동 형태소 분석기를 적용하여 볼 일이 있을 수 있다. 이 때 세종계획에서 개발한 ‘지능형 형태소 분석기’가 유용하다. 국립국어원의 ‘언어정보나눔터’ 홈페이지에서 회원 가입 후 내려받을 수 있다.

지능형 형태소 분석기는 input 파일이 CP949일 것을 전제로 한다. 따라서 UTF-8으로 되어 있는 학습자 원시 말뭉치는 CP949로 인코딩을 변환해야 한다. 한마루의 ‘변환’ 메뉴는 output이 UTF-16LE인 경우만을 위한 것이므로 여기에는 쓸 수 없다. 1.5.1.에서 언급할 EncodeAnt는 output이 UTF-8인 경우만을 위한 것이므로 역시 여기에는 쓸 수 없다. 1.3.1.에서 말한 UniConv와 배치 파일을 이용해야 한다.

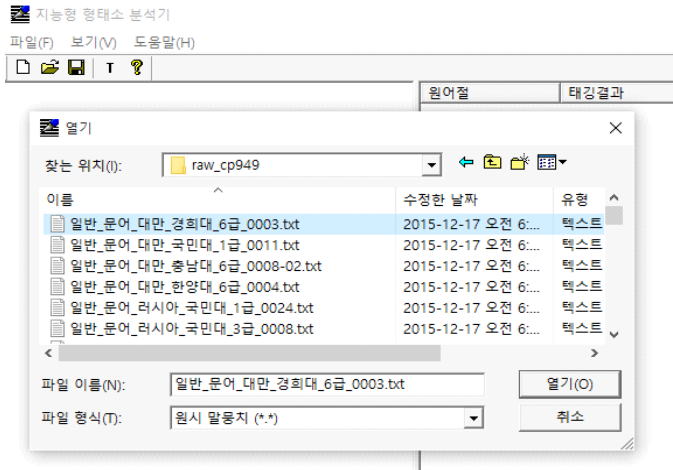


```
D:\Wlearner\Wraw>ls | gawk '{ printf( "uniconv UCS2 %s CP949 ../raw_cp949/%s\\n", $0, $0 )}' >convert.bat
```

<그림 95> ls와 gawk를 이용한 코드 변환: UTF-16LE → CP949

<그림 95>와 같이 하여 배치 파일을 만든 뒤, 첫째 행을 삭제하고, 슬래시를 역슬래시로 바꾸고 CP949로 저장하면서 줄바꿈 문자도 CR-LF로 지정한다. 그 다음에 convert라고 타이핑하고 엔터를 쳐서 배치 파일을 실행하면, raw_cp949 디렉토리에 CP949로 변환된 파일들이 저장된다. 이 파일들을 지능형 형태소 분석기의 input으로 삼으면 된다.

지능형 형태소 분석기는 한 번에 하나의 파일만 처리할 수 있다.



<그림 96> 지능형 형태소 분석기에서 input 파일 열기

input 파일을 열면 좌측 창에 파일이 열려서 보이고, 툴바의 “T” 아이콘을 클릭하면 태깅이 이루어져 우측 창에 제시된다. 지능형 형태소 분석기에서 사용하는 태그셋은 기본적으로 세종 형태 분석 말뭉치의 태그셋과 대동소이하다. 태깅된 결과는 툴바의 저장 아이콘 또는 메뉴바의 ‘파일’ - ‘작업 저장’ 메뉴를 통해 파일로 저장할 수 있다.



<그림 97> 지능형 형태소 분석기로 학습자 원시 말뭉치의 한 파일을 태깅한 결과

1.5. AntConc를 이용한 말뭉치 활용

말뭉치 분석 도구, 특히 콘코던스 소프트웨어로서 세계적으로 널리 알려진 것으로 Wordsmith, MonoConc, AntConc 등이 있다. 최근에는 완전히 무료이면서도 매우 다양한 기능을 제공하며 유니코드를 지원하는 AntConc가 가장 각광을 받고 있다. 공개/무료 소프트웨어는 상용 소프트웨어에 비해 관리가 잘 안 되는 일이 많은데, AntConc는 공개/무료 소프트웨어이면서도 사용자로부터 피드백을 받아 비교적 자주 업데이트되는 등, 전반적으로 잘 관리되고 있다는 점이 큰 장점이다.

서로 비슷비슷한 여러 소프트웨어들을 알고 넓게 소개하기보다는, 우수한 하나의 소프트웨어에 초점을 맞추어 자세히 소개하는 것이 더 유용하리라 생각되어, 여기서는 AntConc에 초점을 맞추어 살펴보겠다.

AntConc는 <http://www.laurenceanthony.net/software/antconc/>에서 다운로드할 수 있으며, 이 사이트에서 AntConc에 관한 자세한 설명을 볼 수 있다. 이 소프트웨어는 Laurence Anthony가 만들었는데, 개발 언어는 Perl이며 여러 컴파일 도구를 이용하여 각 운영체제에서 사용할 수 있는 실행파일을 만들었다. 윈도우, 매킨토시, 리눅스 운영체제가 모두 지원된다.

1.5.1. 말뭉치 파일 탑재를 위한 사전 작업: 인코딩과 파일 이름

검색을 하기 전에 우선 말뭉치 파일들을 탑재(load)해야 하는데, 각 파일의 인코딩이 어떻게 되어 있는지를 점검해야 한다. AntConc의 디폴트 파일 인코딩은 UTF-8인데 이를 변경하려면 Global Settings 메뉴로 들어가서 Character Encoding을 변경하면 된다. 만약 한국어, 영어, 중국어, 일본어 등의 다국어 말뭉치를 사용한다거나, 여러 언어의 말뭉치를 수시로 바꿔 가며 사용할 것 같으면, 디폴트 인코딩을 UTF-8으로 두는 것이 나을 수 있다.

디폴트 인코딩을 이렇게 놔둔 상태에서는 load되는 모든 파일이 UTF-8으로 되어 있어야 한다. 한국어 인코딩(EUC-KR, CP949)으로 되어 있는 파일은 UTF-8으로 변환해야 한다.

이런 인코딩 변환 작업은 1.3.1.에서 소개한 UniConv와 배치 파일을 이용하여 할 수도 있고, Laurence Anthony가 만든 EncodeAnt라는 소프트웨어를 이용하여 할 수도 있다. (<http://www.laurenceanthony.net/software/encodeant/>)

학습자 원시 말뭉치는 UTF-8으로 되어 인코딩 상으로는 AntConc에서 사용하기에 문제가 없다. 다만 파일 이름에 아스키 문자(로마자, 숫자, 키보드상에서 입

력 가능한 기본 기호들)가 아닌 것(예를 들면 한글)이 들어 있으면 문제가 생긴다. 따라서 파일 이름을 모두 아스키 문자로 바꾼 뒤에 이용해야 한다. rename이라는 명령어를 이용하되, 앞서 1.3.1.에서 설명한 것과 비슷한 방식으로 배치 파일을 만들어서 일괄적으로 처리할 수 있다.

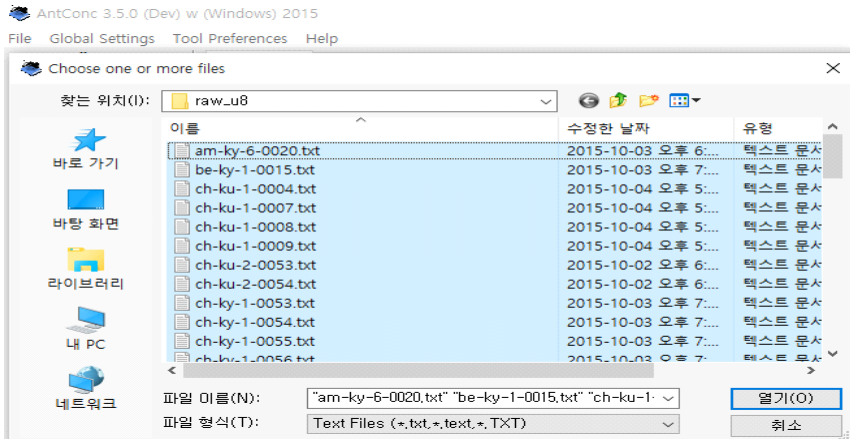
```

1 rename 일반_문어_대만_경희대_6급_0003.txt ta-ky-6-0003.txt
2 rename 일반_문어_대만_국민대_1급_0011.txt ta-ku-1-0011.txt
3 rename 일반_문어_대만_충남대_6급_0008-02.txt ta-cn-6-0008-02.txt
4 rename 일반_문어_대만_한양대_6급_0004.txt ta-hy-6-0004.txt
5 rename 일반_문어_러시아_국민대_1급_0024.txt ru-ku-1-0024.txt
6 rename 일반_문어_러시아_국민대_3급_0008.txt ru-ku-3-0008.txt
7 rename 일반_문어_러시아_국민대_3급_0028.txt ru-ku-3-0028.txt
8 rename 일반_문어_몽골_국민대_1급_0033.txt mo-ku-1-0033.txt
9 rename 일반_문어_미국_경희대_6급_0020.txt 미국-ky-6-0020.txt
10 rename 일반_문어_미얀마_경희대_6급_0007.txt my-ky-6-0007.txt
11 rename 일반_문어_베트남_한양대_1급_0036.txt vi-hy-1-0036.txt
12 rename 일반_문어_베트남_호남대_1급_0009.txt vi-hn-1-0009.txt
13 rename 일반_문어_벨기에_경희대_1급_0015.txt be-ky-1-0015.txt
14 rename 일반_문어_스웨덴_경희대_1급_0007.txt sw-ky-1-0007.txt
15 rename 일반_문어_스웨덴_경희대_1급_0040.txt sw-ky-1-0040.txt
16 rename 일반_문어_스페인_충남대_1급_0029-02.txt sp-cn-1-0029-02.txt
17 rename 일반_문어_인도네시아_충남대_1급_0050-02.txt in-cn-1-0050-02.txt
18 rename 일반_문어_일본_경희대_1급_0003.txt ja-ky-1-0003.txt
19 rename 일반_문어_일본_경희대_1급_0010.txt ja-ky-1-0010.txt
20 rename 일반_문어_일본_경희대_1급_0028.txt ja-ky-1-0028.txt
21 rename 일반_문어_일본_경희대_1급_0035.txt ja-ky-1-0035.txt
22 rename 일반_문어_일본_경희대_1급_0037.txt ja-ky-1-0037.txt
23 rename 일반_문어_일본_경희대_6급_0018.txt ja-ky-6-0018.txt
24 rename 일반_문어_일본_경희대_6급_0019.txt ja-ky-6-0019.txt
25 rename 일반_문어_일본_국민대_1급_0001.txt ja-ku-1-0001.txt
26 rename 일반_문어_일본_충남대_5급_0072-02.txt ja-cn-5-0072-02.txt
27 rename 일반_문어_일본_한양대_1급_0037.txt ja-hy-1-0037.txt
28 rename 일반_문어_일본_한양대_5급_0004.txt ja-hy-5-0004.txt
29 rename 일반_문어_일본_한양대_5급_0005.txt ja-hy-5-0005.txt
30 rename 일반_문어_일본_한양대_6급_0006.txt ja-hy-6-0006.txt
31 rename 일반_문어_일본_한양대_6급_0007.txt ja-hy-6-0007.txt
32 rename 일반_문어_일본_한양대_6급_0008.txt ja-hy-6-0008.txt
33 rename 일반_문어_일본_한양대_6급_0009.txt ja-hy-6-0009.txt
34 rename 일반_문어_일본_홍익대_6급_0001.txt ja-ho-6-0001.txt
35 rename 일반_문어_일본_홍익대_6급_0002.txt ja-ho-6-0002.txt
36 rename 일반_문어_일본_홍익대_6급_0007.txt ja-ho-6-0007.txt
37 rename 일반_문어_일본_홍익대_6급_0014.txt ja-ho-6-0014.txt
38 rename 일반_문어_일본_홍익대_6급_0016.txt ja-ho-6-0016.txt
39 rename 일반_문어_일본_홍익대_6급_0017.txt ja-ho-6-0017.txt
40 rename 일반_문어_중국_경희대_1급_0053.txt ch-ky-1-0053.txt
41 rename 일반_문어_중국_경희대_1급_0054.txt ch-ky-1-0054.txt

```

<그림 98> 파일 이름 바꾸기 배치 파일

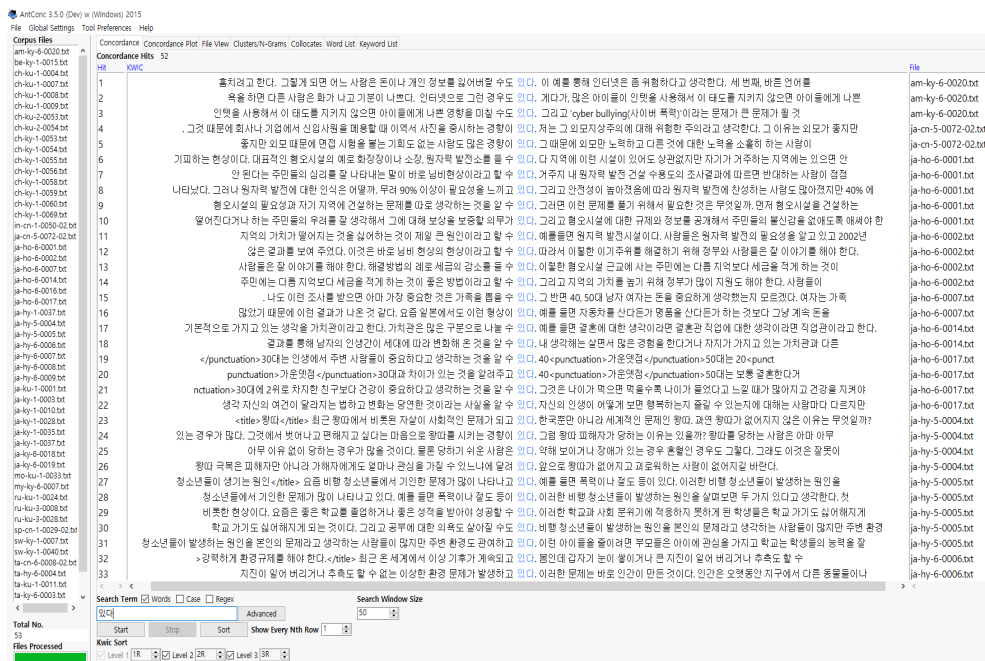
말뭉치 파일을 탑재하기 위해서는 메뉴바의 'File' - 'Open File(s)' 메뉴를 선택한 뒤 말뭉치 파일들을 선택하여 '열기' 버튼을 누르면 된다(<그림 99>).



<그림 99> AntConc에서 말뭉치 파일 열기

1.5.2. 단순 용례 검색

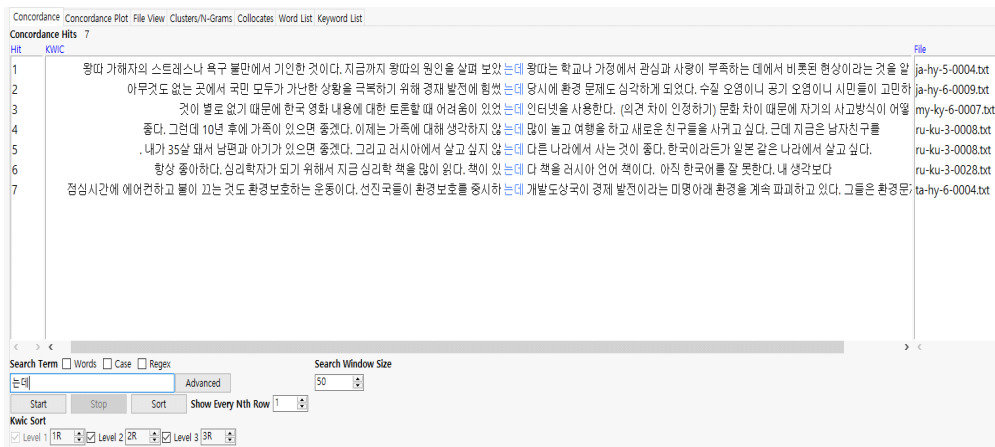
상단의 탭에서 Concordance가 선택되어 있는 상태에서 Search Term에 ‘있다’라는 문자열을 입력하고 Start 버튼을 누르면 모두 52개의 용례가 검색되어 나온다.



<그림 100> AntConc에서 ‘있다’ 검색 결과 (어절 단위)

1.5.3. 어절의 일부 검색

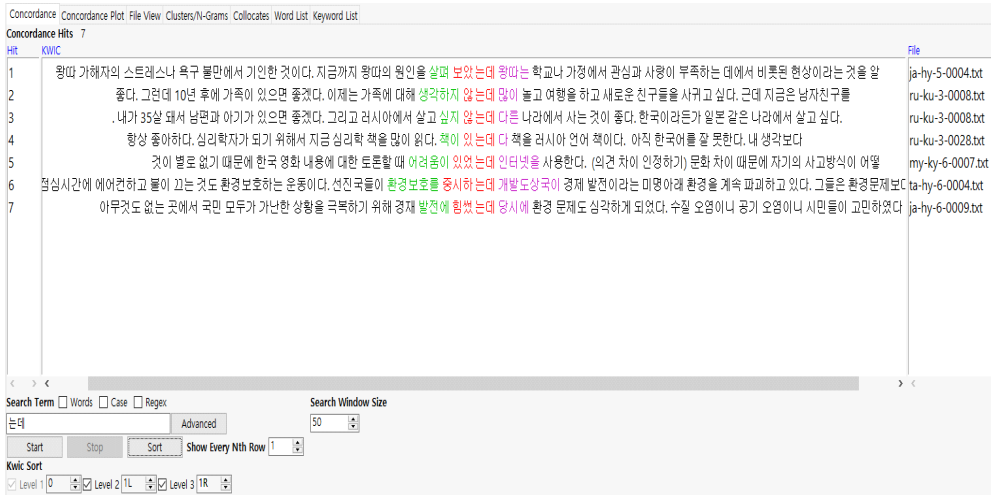
한마루와 마찬가지로 AntConc도 기본 검색은 어절을 단위로 한다. 즉 입력한 검색어가 온전한 어절을 이루는 경우만을 찾아 준다. 온전한 어절이 아니라 어절의 일부를 검색하려면 Search Term 오른쪽의 Words의 선택을 해제해야 한다. 예컨대 어미 ‘-는데’를 검색하고 싶다면, ‘는데’는 그 자체로서 하나의 어절을 이루지는 못하고 어절의 일부이므로 Words의 선택을 해제한 상태에서 검색한다. 그러면 7개의 용례가 검색되어 나온다(<그림 101>).



<그림 101> AntConc에서 ‘-는데’ 검색 결과 (어절 조건 해제)

1.5.4. 검색 결과의 정렬

위의 ‘-는데’ 검색 결과에서, 어미 ‘-는데’가 어떤 용언 뒤에 많이 붙었는지 알고 싶을 수 있다. 이것은 ‘-는데’를 포함한 어절 자체를 기준으로 소팅을 하면 어느 정도 알 수 있다. AntConc는 검색 결과를 소팅할 때 정렬 기준을 세 단계까지 명시할 수 있게 해 준다. 화면 하단의 Kwic Sort 아래에 Level 1, Level 2, Level 3가 있는데, 0은 Search Term 바로 그 어절을 의미하고, 1L은 왼쪽 첫째 어절, 2L은 왼쪽 둘째 어절, ……을 의미하고, 1R은 오른쪽 첫째 어절, 2R은 오른쪽 둘째 어절, ……을 의미한다. Level 1을 0으로 설정하면 ‘-는데’를 포함한 바로 그 어절이 첫째 정렬 기준이 된다. 그러면 정렬 결과가 다음과 같다 나온다(<그림 102>).



<그림 102> AntConc에서 검색 결과를 해당 어절 기준으로 소팅한 결과

1.5.5. 아무개 문자

동사 ‘먹-’ 앞에 어떤 단어(대개 목적어)가 주로 오는지 알고 싶다고 하자. 그럼 우선 동사 ‘먹-’을 검색해야 할 것이다. Words 선택을 해제하고 ‘먹’을 검색할 수도 있겠지만, 그러면 그냥 ‘먹’ 하나의 음절로 이루어진 어절도 검색 결과에 포함될 텐데, 이것은 동사 ‘먹-’의 용례가 아닐 것이다(아마도 명사 ‘먹’의 예일 듯). 따라서 Words를 선택한 상태에서 아무개 문자(wild card)를 사용하는 게 더 좋다. 아무개 문자에는 ?(1개의 임의의 문자), *(0개 이상의 임의의 문자), +(1개 이상의 임의의 문자) 등이 있다. 동사 ‘먹-’의 활용형들을 검색하려면 ‘먹+’라고 하면 된다(<그림 103>). 즉 이 어절은 ‘먹’이라는 문자로 시작되고 그 뒤에는 1개 이상의 임의의 문자가 오면 된다. 물론 이럴 경우 명사 ‘먹’에 조사가 붙은 ‘먹이’, ‘먹을’, ‘먹도’, ‘먹만’ 등도 함께 검색되어 나오겠지만, 품사 구별이 되어 있는 형태분석 말뭉치가 아닌 이상 이 정도의 noise는 감내할 수밖에 없다.

Concordance Hits 21			File
Hit	KWIC		File
1	합니다. 저는 아홉 시에 일어나거나 세수합니다. 세수한 후에 아침에 먹습니다. 밥을 먹은 후에 청소해요. 방에서 조금 어려웠습니다. 그렇지만 가족사 친구가 아침에 친구를 만나		ch-ku-1-0007.txt
2	네 시 반에 우리도 버거 고릅니다. 그래서 밥을 먹으려 갔습니다. 밥을 먹은 후에 집에 갔습니다.		ch-ku-1-0007.txt
3	>아해서 대련을 좋아한다. 대련의 바다 너무 크다. 해물을 좋아한다. 매일 해물을 먹고 싶었다. 나는 친구와 같이 대련에 가 봤다. 그래서 수렁하고 싶었다. 그날대		ch-ku-2-0053.txt
4	섯이 많이 새 먹었어. 오아도에서 해물 유명해. 그래서 퐁메이트와 같이 해물이 먹어 , 맛있기 때문에 너무 비싸요. 그렇지만 바다 근처에 살고 싶어. 마음에 들어요.		ch-ku-2-0054.txt
5	보고 싶어합니다. 한국 영화를 재미있습니다. 저는 중국 친구 같이 한국 음식을 먹고 싶어합니다. 한국에 선생님은 친절합니다. 주말에 공항 친구를 만납니다. 집에 텔레비전		ch-ky-1-0053.txt
6	.명통에 가고 싶습니다. 저는 가방하고 머리가 다 사고 싶습니다. 저는 한국 음식을 먹고 싶습니다. 한국 음식이 아주 맛있었습니다. 2015년 8월 29일부터 저는 집에 갈		ch-ky-1-0054.txt
7	입력함 <note>를 만나고 싶습니다. 저는 올해에 한국 음식을 많이 먹고 싶습니다. 저는 불고기하고 비빔밥하고 순두부찌개하고 갈비를 먹고 싶습니다.		ch-ky-1-0055.txt
8	한국 음식을 많이 먹고 싶습니다. 저는 불고기하고 비빔밥하고 순두부찌개하고 갈비를 먹고 싶습니다.		ch-ky-1-0055.txt
9	올해 한국어를 배우고 싶습니다. 저는 한국에 살고 싶습니다. 한국에서 맛있 음식 먹고 싶습니다. 그리고, 예쁜 옷들 사고 싶습니다. 그래서, 한국어를 배우고 싶습니다. 저		ch-ky-1-0059.txt
10	통대문하고 부산으로 갈 겁니다. 저는 한국 문화를 공부하고 많은 한국 음식을 먹고 싶습니다. 내년 저는 대학원에 들어가고 싶습니다. 저는 지금 한국어하고 영어를 공부합		ch-ky-1-0060.txt
11	산책하고 싶습니다. 그리고 제주도에서는 글이 아주 유원합니다. 그래서 맛있는 글을 많이 먹고 싶습니다. 그리고 제주도에서 생선을 먹고 싶습니다. 일본도 생선이 아주 맛있습니		ja-ky-1-0003.txt
12	아주 유원합니다. 그래서 맛있는 글을 많이 먹고 싶습니다. 그리고 제주도에서 생선을 먹고 싶습니다. 일본도 생선이 아주 맛있었습니다. 자주 먹습니다. 아주 좋아합니다. 그래서		ja-ky-1-0003.txt
13	생선이 아주 맛있습니. 자주 먹습니. 아주 좋아합니다. 그래서 제주도에서 생선을 먹고 싶습니다. 제주도에서 맛있는 곳을 먹고 아름다운 경치를 보고 싶습니다.		ja-ky-1-0003.txt
14	.아주 좋아합니다. 그래서 제주도에서 생선을 먹고 싶습니다. 제주도에서 맛있는 곳을 먹고 아름다운 경치를 보고 싶습니다.		ja-ky-1-0003.txt
15	불고 싶습니다. 저는 많이 여행을 가고 싶습니다. 저는 많이 한국 음식을 먹고 싶습니다. 저는 많이 한국 문화를 보고 싶습니다.		ja-ky-1-0010.txt
16	가고 싶어요. 제 친구들과 같이 어귀나와에도 가고 싶어요. 맛있는 음식을 많이 먹고 싶어요. 외국에도 가고 싶어요. 우리 반 친구들은 여러가지 나라에서 왔어요. 그래서,		ja-ky-1-0035.txt
17	없이 확실하지 않음 <note> 있습니다. 그래서 N서울타워에서 친구하고 같이 스카타티를 먹을 거습니. 다음에 가족사에 갈 거습니. 나는 당신의 주말을 사랑합니다.		ru-ku-1-0024.txt
18	계획을 잡지 않다. 하지만 나는 좋은 대학교 꼭 졸업하고 싶. 다음에 먹는 일이 있고 싶. 아마 그 미래가 좀 어려운 미래. 열심히 공부해요.		ru-ku-3-0028.txt
19	출진만 영화가 재미있어요. 일요일에는 우리 같이 식당에 가고 불고기들 두 그릇 먹고 포도 주스를 마셨어요. 음식을 정말 좋아했어요. 저녁에 우리 같이 공원에서 산책했어요		sp-cn-1-0029-02.txt
20	일본 패션을 정말 좋아해요. 그리고 맛있는 한국 음식이 많아서 여기에서 많이 먹고 싶어요. 그리고 제 다른 나라 친구 다시 만나고 싶어요. 독일 친구하고		sw-ky-1-0040.txt
21	도상국이 경제 발전이라는 미명아래 환경을 계속 파괴하고 있다. 그들은 환경문제보다 자기 먹고 살기가 더 중요하다고 생각한다. 그런데 현재 지구온난화 심하는 미명에 계속 지구를		ta-hy-6-0004.txt

Search Term ☐ Case ☐ Regex

Start

Stop

Sort

Show Every Nth Row

Advanced

50

Kwic: Sort

Level 1

0

Level 2

TL

Level 3

1R

Search Window Size

50

<그림 103> AntConc에서 아무개 문자를 이용한 검색: ‘먹+’

이 검색 결과를 바탕으로 하여, ‘먹-’의 활용형을 분류하고 싶으면 제1 검색 기준을 해당 어절(0)로 하면 되고, ‘먹-’ 앞에 오는 요소를 알아보고 싶으면 제1 검색 기준을 왼쪽 첫째 어절(1L)로 하면 된다(<그림 104>).

Concordance Hits 21				File
Hit	KWIC			
1	한국 음식을 많이 많이 먹고 싶습니다. 저는 불고기하고 비빔밥하고 순두부찌개하고 갈비를 먹고 싶습니다.	ch-ky-1-0055.txt		
2	.아주 좋아합니다. 그래서 제주도에서도 생선을 먹고 싶습니다. 제주도에서 맛있는 곳을 먹고 아름다운 경치를 보고 싶습니다.	ja-ky-1-0003.txt		
3	출진만 영화가 재미있어요. 일요일에는 우리 같이 식당에 가고 불고기들 두 그릇 먹고 포도 주스를 마셨어요. 음식을 정말 좋아했어요. 저녁에 우리 같이 공원에서 산책했어요	sp-cn-1-0029-02.txt		
4	계획을 잡지 않다. 하지만 나는 좋은 대학교 꼭 졸업하고 싶. 다음에 먹는 일이 있고 싶. 아마 그 미래가 좀 어려운 미래. 열심히 공부해요.	ru-ku-3-0028.txt		
5	입력함 <note>를 만나고 싶습니다. 저는 올해에 한국 음식을 많이 먹고 싶습니다. 저는 불고기하고 비빔밥하고 순두부찌개하고 갈비를 먹고 싶습니다.	ch-ky-1-0055.txt		
6	산책하고 싶습니다. 그리고 제주도에서는 글이 아주 유원합니다. 그래서 맛있는 글을 많이 먹고 싶습니다. 그리고 제주도에서 생선을 먹고 싶습니다. 일본도 생선이 아주 맛있습니	ja-ky-1-0003.txt		
7	가고 싶어요. 제 친구들과 같이 어귀나와에도 가고 싶어요. 맛있는 음식을 많이 먹고 싶어요. 외국에도 가고 싶어요. 우리 반 친구들은 여러가지 나라에서 왔어요. 그래서,	ja-ky-1-0035.txt		
8	일본 패션을 정말 좋아해요. 그리고 맛있는 한국 음식이 많아서 여기에서 많이 먹고 싶어요. 그리고 제 다른 나라 친구 다시 만나고 싶어요. 독일 친구하고	sw-ky-1-0040.txt		
9	됩니다. 저는 아홉 시에 일어나거나 세수합니다. 세수한 후에 아침에 먹습니다. 밥을 먹은 후에 청소해요. 방에서 조금 어려웠습니다. 그렇지만 가족사 친구가 아침에 친구를 만나	ch-ku-1-0007.txt		
10	네 시 반에 우리도 버거 고릅니다. 그래서 밥을 먹으려 갔습니다. 밥을 먹은 후에 집에 갔습니다.	ch-ku-1-0007.txt		
11	아주 유원합니다. 그래서 맛있는 글을 많이 먹고 싶습니다. 그리고 제주도에서 생선 을 먹고 싶습니다. 일본도 생선이 아주 맛있습니. 자주 먹습니. 아주 좋아합니다. 그래서	ja-ky-1-0003.txt		
12	생선이 아주 맛있습니. 자주 먹습니. 아주 좋아합니다. 그래서 제주도에서 생선 을 먹고 싶습니다. 제주도에서 맛있는 곳을 먹고 아름다운 경치를 보고 싶습니다.	ja-ky-1-0035.txt		
13	없이 확실하지 않음 <note> 있습니다. 그래서 N서울타워에서 친구하고 같이 스카타티를 먹을 거습니. 다음에 가족사에 갈 거습니. 나는 당신의 주말을 사랑합니다.	ru-ku-1-0024.txt		
14	올해 한국어를 배우고 싶습니다. 저는 한국에 살고 싶습니다. 한국에서 맛있 음식 먹고 싶습니다. 그리고, 예쁜 옷들 사고 싶습니다. 그래서, 한국어를 배우고 싶습니다. 저	ch-ky-1-0059.txt		
15	.명통에 가고 싶습니다. 저는 가방하고 머리가 다 사고 싶습니다. 저는 한국 음식을 먹고 싶습니다. 한국 음식이 아주 맛있었습니다. 2015년 8월 29일부터 저는 집에 갈	ch-ky-1-0054.txt		
16	통대문하고 부산으로 갈 겁니다. 저는 한국 문화를 공부하고 많은 한국 음식을 먹고 싶습니다. 내년 저는 대학원에 들어가고 싶습니다. 저는 지금 한국어하고 영어를 공부합	ch-ky-1-0060.txt		
17	불고 싶습니다. 저는 많이 여행을 가고 싶습니다. 저는 많이 한국 음식을 먹고 싶습니다. 저는 많이 한국 문화를 보고 싶습니다.	ja-ky-1-0010.txt		
18	보고 싶어합니다. 한국 영화를 재미있습니다. 저는 중국 친구 같이 한국 음식을 먹고 싶어합니다. 한국에 선생님은 친절합니다. 주말에 공항 친구를 만납니다. 집에 텔레비전	ch-ky-1-0053.txt		
19	도상국이 경제 발전이라는 미명아래 환경을 계속 파괴하고 있다. 그들은 환경문제보다 자기 먹고 살기가 더 중요하다고 생각한다. 그런데 현재 지구온난화 심하는 미명에 계속 지구를	ta-hy-6-0004.txt		
20	섯이 많이 새 먹었어. 오아도에서 해물 유명해. 그래서 퐁메이트와 같이 해물이 먹어 , 맛있기 때문에 너무 비싸요. 그렇지만 바다 근처에 살고 싶어. 마음에 들어요.	ch-ku-2-0054.txt		
21	>아해서 대련을 좋아한다. 대련의 바다 너무 크다. 해물을 좋아한다. 매일 해물을 먹고 싶었다. 나는 친구와 같이 대련에 가 봤다. 그래서 수렁하고 싶었다. 그날대	ch-ku-2-0053.txt		

Search Term

Words

Case

Regex

Advanced

50

Start

Stop

Show Every Nth Row

Kwic Sort

Level 1

Level 2

Level 3

Search Window Size

50

<그림 104> AntConc에서 ‘먹+’ 검색 결과를 좌측 어절 기준으로 정렬한 결과

1.5.6. 정규표현

어미 ‘-고’가 반복해서 쓰일 때의 패턴을 알아보고 싶다고 치자. 이 경우에 ‘-고’와 ‘-고’ 사이에는 매우 다양한 말들이 올 수 있기 때문에 Search Term을 어떻게 설정해야 할지 막막할 것이다. 이럴 때에는 정규표현(regular expression)을 이용하는 것이 좋다. 정규표현에서 사용되는 특수문자들의 의미는 위키백과 등 인터넷 검색을 통해 쉽게 조사할 수 있으므로 각자 알아보기 바란다. ‘-고’가 반복된 경우를 찾기 위해서는 Search Terms 오른쪽의 Words는 선택을 해제하고 Regex는 선택한 상태에서 ‘고.{0,20}고’를 검색해 보자(<그림 105>). 이 정규표현의 의미는 대략 ‘고’와 ‘고’ 사이에 임의의 문자가 0개~20개 온 경우를 찾으라는 것이다.

Concordance Plot, File View, Clusters/N-Grams, Collocates, Word List, Keyword List

Concordance Hits 124

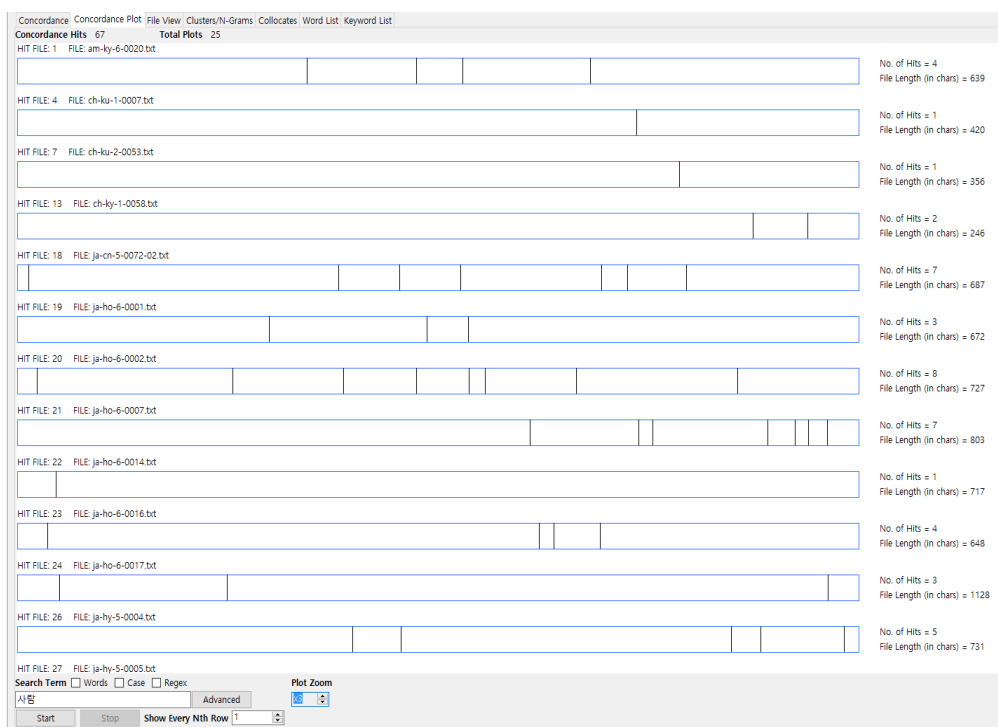
Hit	KWIC	File
1	저는 올해 여자친구들이 부산에 가고 싶습니다. 저는 한국친구를 만나고 싶습니다. 저는 요리를 배우 싶습니다. 저는 온	be-ky-1-0015.txt
2	요리를 배우 싶습니다. 저는 운동을 계속 하고 싶습니다. 저는 영국에서 축구광에 가고 싶습니다. 저는 콘서트에 가고 싶습니다. 저는 벨리기에 가고 싶습니다. 저는 버무	be-ky-1-0015.txt
3	저는 영국에서 축구광에 가고 싶습니다. 저는 콘서트에 가고 싶습니다. 저는 벨리기에 가고 싶습니다. 저는 버무대에 가고 싶습니다. 버무대에 가서 먹습니다.	be-ky-1-0015.txt
4	을 읽습니다. 점심에 밥을 먹으러 식당에 갑니다. 오후에 커피숍에서 친구를 만납니다. 그리고 친구하고 같이 명동에 갑니다. 명동에서 옷하고 운동화를 삽니다. 그리고 기숙사에 갑니다	ch-ku-1-0004.txt
5	. 오후에 커피숍에서 친구를 만납니다. 그리고 친구하고 같이 명동에 갑니다. 명동에서 옷하고 운동화를 삽니다. 그리고 기숙사에 갑니다. 일요일에 도서관에 갑니다. 친구하고 같이 공	ch-ku-1-0004.txt
6	동에서 옷하고 운동화를 삽니다. 그리고 기숙사에 갑니다. 일요일에 도서관에 갑니다. 친구하고 같이 공부합니다. 그리고 한식당에 밥을 먹습니다. 오후에 기숙사에 와서 컴퓨터를 합니	ch-ku-1-0004.txt
7	다. 텔레비전을 켜고. 주말에 날씨가 맑지만 방에서 혼자입니다. 너무 심심했습니다. 월요일에 고향에 가서 명동에서 선물을 사고 싶습니다. 월요일에 고향에 가지고 갑니다. 친구에게서 전화	ch-ku-1-0007.txt
8	받습니다. 너무 심심했습니다. 월요일에 고향에 가서 명동에서 선물을 사고 싶습니다. 월요일에 고향에 가지고 갑니다. 친구에게서 전화 받습니다. 약속은 오후 한 시 명동에 만납니다.	ch-ku-1-0007.txt
9	저는 주말에 여자친구를 만나거나 기숙사에서 컴퓨터를 해요. 보통 주말에 여자친구하고 명동에 가요. 여자친구하고 영화관에서 같이 영화를 봐요. 제 여자친구가 코미디 영화거나	ch-ku-1-0008.txt
10	제가 한국에 온 지 6개월 됐어. 한국 문화 알고 싶었어 바다에 보고 싶어 그래서 홈페이지에 오이도에 가봤어. 오이도에서 좀	ch-ku-2-0054.txt
11	한국 친구 같이 영화를 보고 싶어합니다. 한국 영화를 재미있습니다. 저는 중국 친구 같이 한국 음식을 먹고 싶어합니다. 한국이 선생님은 친절합니다. 주말에 공항 친구를	ch-ky-1-0053.txt
12	저는 2015년에 한국어를 공부 하고 싶습니다. 한국어를 좋아합니다. 그리고 중국에 가고 싶습니다. 우리 어머니하고 아버지가 보고 싶습니다. 명동에 가고 싶습니다.	ch-ky-1-0054.txt
13	를 공부 하고 싶습니다. 한국어를 좋아합니다. 그리고 중국에 가고 싶습니다. 우리 어머니하고 아버지가 보고 싶습니다. 명동에 가고 싶습니다. 저는 가방하고 머리카락 사고 싶습니다.	ch-ky-1-0054.txt
14	고 싶습니다. 우리 어머니하고 아버지가 보고 싶습니다. 명동에 가고 싶습니다. 저는 가방하고 머리카락 사고 싶습니다. 저는 한국 음식을 먹고 싶습니다. 한국 음식이 아주	ch-ky-1-0054.txt
15	일부러 저는 집에 갈 거니까. 2015년 12월에 크리스마스 있습니다. 그래서 저는 친구하고 흥대에 갈 거고 싶습니다. 그리고 영화를도 보고 싶습니다. 노래방도 가고 싶습니다.	ch-ky-1-0054.txt
16	크리스마스 있습니다. 그래서 저는 친구하고 흥대에 갈 거고 싶습니다. 그리고 영화를도 보고 싶습니다. 노래방도 가고 싶습니다. 친구를 노래를 하고 싶습니다. 그래서 우리 노래방에	ch-ky-1-0054.txt
17	저는 올해에 여행을 하고 싶습니다. 저는 미국에 가고 싶습니다. 저는 구 물부터 심 할까지 고향에	ch-ky-1-0055.txt
18	하고 싶습니다. 저는 미국에 가고 싶습니다. 저는 구 물부터 심 할까지 고향에 놀아가고 싶습니다. 아버지하고 어머니를 만나고 싶습니다. 저는 심 할부터 심 일까지	ch-ky-1-0055.txt
19	삽니다. 아버지하고 어머니를 만나고 싶습니다. 저는 심 할부터 심 일까지 학과에 오고 싶습니다. 학과에 조금 2를 공부하고 싶습니다. 저는 방학에 가수를 만나고 싶습니다.	ch-ky-1-0055.txt
20	입력함</note>를 만나고 싶습니다. 저는 올해에 한국 음식을 많이 먹고 싶습니다. 저는 불고기하고 비빔밥하고 순두부찌개하고 갈비를 먹고 싶습니다.	ch-ky-1-0055.txt
21	저는 올해에 한국 음식을 많이 먹고 싶습니다. 저는 불고기하고 비빔밥하고 순두부찌개하고 갈비를 먹고 싶습니다.	ch-ky-1-0055.txt
22	2015년에 저는 한국어를 공부하고 싶습니다. 그리고 한국 문화 알고 싶습니다. 제가 한국 문화를 정말 좋아합니다.	ch-ky-1-0056.txt
23	국어 배우기 위해서 한국 대학교에 다닙니다. 그래서 지금 한국어를 열심히 공부합니다. 그리고 한국 친구를 만나고 싶습니다. 한국에 온 후에 한국 문화에 많이 알립니다.	ch-ky-1-0058.txt
24	저는 올해 한국어를 배우고 싶습니다. 저는 한국에 살고 싶습니다. 한국에서 맛있는 음식 먹고 싶습니다. 그리고	ch-ky-1-0059.txt
25	올해 한국어를 배우고 싶습니다. 저는 한국에 살고 싶습니다. 한국에서 맛있는 음식 먹고 싶습니다. 그리고	ch-ky-1-0059.txt
26	맛있는 음식 먹고 싶습니다. 그리고, 예쁜 옷 사고 싶습니다. 그래서, 한국어를 배우고 싶습니다. 저는 한국 대학교에 가고 싶습니다. <name>대학교는 가고 싶습니다. 저는	ch-ky-1-0059.txt
27	제주도에 갈 겁니다. 방학 후에 저는 조금 2을 돌아갈 겁니다. 저는 명동하고 종대문하고 부산으로 갈 겁니다. 저는 한국 문화를 공부하고 많은 한국 음식을	ch-ky-1-0060.txt
28	돌아갈 겁니다. 저는 명동하고 종대문하고 부산으로 갈 겁니다. 저는 한국 문화를 공부하고 많은 한국 음식을 먹고 싶습니다. 내년 저는 대학원에 들어가고 싶습니다. 저는	ch-ky-1-0060.txt
29	한국 문화를 공부하고 많은 한국 음식을 먹고 싶습니다. 내년 저는 대학원에 들어가고 싶습니다. 저는 지금 한국어하고 영어를 공부합니다. 한국 일상생활이 재미있습니다. 저는	ch-ky-1-0060.txt
30	올해 한국어를 배우고 싶습니다. 그래서 고향에서 수업이 더 열심히 할 겁니다. 그리고 한국 문화도 공부하고 싶습니다. 너무 재미있습니다. 주말에 친구하고 같이 유명한 곳이	ch-ky-1-0069.txt
31	삼히 할 겁니다. 그리고 한국 문화도 공부하고 싶습니다. 너무 재미있습니다. 주말에 친구하고 같이 유명한 곳이고 싶습니다. 그리고 저는 한국 음식을 좋아합니다. 그래서	ch-ky-1-0069.txt
32	주말에는 기숙사에서 청소를 하고 빨래를 하고 요리를 합니다. 식당에 갑니다. 나쁜 친구도 같이 불고기를 먹습니다.	in-cn-1-0050-02.txt

Search Term Words Case Regex
고.{0,20}고 Advanced
Start Stop Sort Show Every Nth Row 1
Kwic Sort
Level 1 Level 2 Level 3

<그림 105> AntConc에서 정규표현 ‘고.{0,20}고’를 검색한 결과

1.5.7. Concordance Plot

어떤 언어요소가 말뭉치의 어느 파일에서, 그리고 각 파일의 어느 부분에서 많이 나타나는지를 알고 싶다고 하자. 이럴 때는 AntConc의 Concordance Plot이 유용하다. 예컨대 53개 파일로 이루어진 학습자 원시 말뭉치에서 ‘사람’을 검색하면 25개 파일에서 67개의 용례가 검색되는데, 각 파일에서 ‘사람’이 몇 번 나오고 그 출현 위치가 어디쯤인지를 바코드 형태로 일목요연하게 보여준다(<그림 106>).



<그림 106> ‘사람’의 Concordance Plot

1.5.8. 단어 빈도 목록

말뭉치에 나타나는 모든 어절의 출현 빈도를 총체적으로 알고 싶을 때에는 Word List를 이용하면 된다. 이 목록을 빈도순으로 소팅할 수도 있고, 자모순으로 소팅할 수도 있다.

AntConc 3.5.0 (Dev) w (Windows) 2015

File Global Settings Tool Preferences Help

Corpus Files

- am-ky-6-0020.txt
- be-ky-1-0015.txt
- ch-ku-1-0004.txt
- ch-ku-1-0007.txt
- ch-ku-1-0008.txt
- ch-ku-1-0009.txt
- ch-ku-2-0053.txt
- ch-ku-2-0054.txt
- ch-ky-1-0053.txt
- ch-ky-1-0054.txt
- ch-ky-1-0055.txt
- ch-ky-1-0056.txt
- ch-ky-1-0058.txt
- ch-ky-1-0059.txt
- ch-ky-1-0060.txt
- ch-ky-1-0069.txt
- in-cn-1-0050-02.txt
- ja-cn-5-0072-02.txt
- ja-ho-6-0001.txt
- ja-ho-6-0002.txt
- ja-ho-6-0007.txt
- ja-ho-6-0014.txt
- ja-ho-6-0016.txt
- ja-ho-6-0017.txt
- ja-hy-1-0037.txt
- ja-hy-5-0004.txt
- ja-hy-5-0005.txt
- ja-hy-6-0006.txt
- ja-hy-6-0007.txt
- ja-hy-6-0008.txt
- ja-hy-6-0009.txt
- ja-ku-1-0001.txt
- ja-ky-1-0003.txt
- ja-ky-1-0010.txt
- ja-ky-1-0028.txt
- ja-ky-1-0035.txt
- ja-ky-1-0037.txt
- ja-ky-6-0018.txt
- ja-ky-6-0019.txt
- mo-ku-1-0033.txt
- my-ky-6-0007.txt
- ru-ku-1-0024.txt
- ru-ku-3-0008.txt
- ru-ku-3-0028.txt
- sp-cn-1-0029-02.txt
- sw-ky-1-0007.txt
- sw-ky-1-0040.txt
- ta-cn-6-0008-02.txt
- ta-hy-6-0004.txt
- ta-ku-1-0011.txt
- ta-ku-6-0002.txt

Total No. 53
Files Processed

Concordance Concordance Plot File View Clusters/N-Grams Collocates Word List Keyword List

Word Types: 2526 Word Tokens: 5708 Search Hits: 0

Rank	Freq	Word	Lemma Word Form(s)
1	98	싶습니다	
2	95	저는	
3	60	수	
4	52	있다	
5	48	그리고	
6	47	많이	
7	46	그래서	
8	43	같이	
9	41	한국	
10	33	생각한다	
11	32	가고	
12	31	그	
13	31	한국어를	
14	29	것이다	
15	26	title	
16	24	한다	
17	24	환경	
18	23	punctuation	
19	23	싫어요	
20	22	하고	
21	21	우리	
22	21	있습니다	
23	19	것이	
24	19	때문에	
25	19	있는	
26	19	주말에	
27	18	갑니다	
28	17	대	
29	17	더	
30	17	보고	
31	17	이런	
32	17	좋아합니다	

Search Term ☐ Words ☐ Case ☐ Regex Hit Location 0

Lemma List ☐ Loaded

Sort by ☐ Invert Order Sort by Freq

<그림 107> 학습자 원시 말뭉치의 어절 빈도 목록 (빈도순)

AntConc 3.5.0 (Dev) w (Windows) 2015

File Global Settings Tool Preferences Help

Corpus Files

- am-ky-6-0020.txt
- be-ky-1-0015.txt
- ch-ku-1-0004.txt
- ch-ku-1-0007.txt
- ch-ku-1-0008.txt
- ch-ku-1-0009.txt
- ch-ku-2-0053.txt
- ch-ku-2-0054.txt
- ch-ky-1-0053.txt
- ch-ky-1-0054.txt
- ch-ky-1-0055.txt
- ch-ky-1-0056.txt
- ch-ky-1-0058.txt
- ch-ky-1-0059.txt
- ch-ky-1-0060.txt
- ch-ky-1-0069.txt
- in-cn-1-0050-02.txt
- ja-cn-5-0072-02.txt
- ja-ho-6-0001.txt
- ja-ho-6-0002.txt
- ja-ho-6-0007.txt
- ja-ho-6-0014.txt
- ja-ho-6-0016.txt
- ja-ho-6-0017.txt
- ja-hy-1-0037.txt
- ja-hy-5-0004.txt
- ja-hy-5-0005.txt
- ja-hy-6-0006.txt
- ja-hy-6-0007.txt
- ja-hy-6-0008.txt
- ja-hy-6-0009.txt
- ja-ku-1-0001.txt
- ja-ky-1-0003.txt
- ja-ky-1-0010.txt
- ja-ky-1-0028.txt
- ja-ky-1-0035.txt
- ja-ky-1-0037.txt
- ja-ky-6-0018.txt
- ja-ky-6-0019.txt
- mo-ku-1-0033.txt
- my-ky-6-0007.txt
- ru-ku-1-0024.txt
- ru-ku-3-0008.txt
- ru-ku-3-0028.txt
- sp-cn-1-0029-02.txt
- sw-ky-1-0007.txt
- sw-ky-1-0040.txt
- ta-cn-6-0008-02.txt
- ta-hy-6-0004.txt
- ta-ku-1-0011.txt
- ta-ku-6-0002.txt

Concordance Concordance Plot File View Clusters/N-Grams Collocates Word List Keyword List

Word Types: 2526 Word Tokens: 5708 Search Hits: 0

Rank	Freq	Word	Lemma	Word Form(s)
1	1	all		
2	1	bullying		
3	1	cyber		
4	1	everland		
5	1	guest		
6	6	misf		
7	1	most		
8	6	name		
9	1	new		
10	14	note		
11	1	n서올다워		
12	1	n서올타워에		
13	1	n서올타워에서		
14	1	of		
15	2	pc방에		
16	1	pc아니어도		
17	23	punctuation		
18	1	punctuation		
19	26	title		
20	1	toeic		
21	1	top		
22	1	층위의		
23	5	가		
24	1	가게		
25	1	가게가		
26	1	가게는		
27	1	가게에서		
28	1	가겠습니다		
29	32	가고		
30	1	가기도		
31	2	가까운		
32	1	가끔		

Search Term ☐ Words ☐ Case ☐ Regex

Hit Location

Advanced Search Only 0

Start Stop Sort

Lemma List ☐ Loaded

Sort by ☐ Invert Order

Sort by Word

Clone Results

Total No. 53

Files Processed

<그림 108> 학습자 원시 말뭉치의 어절 빈도 목록 (자모순)

1.5.9. 연어/공기 분석

우리가 어떤 단어에 관심을 갖고 있다고 할 때(타겟), 이 타겟이 어떤 표현들과 함께 나타나는지 알고 싶을 수 있다. 이럴 때 Collocates 메뉴가 유용하다. 이것은 연어 관계, 공기 관계를 통계적으로 분석해서 타겟과 유의미한 연어/공기 관계에 있는 언어 표현들을 추출해 준다. 이 때 사용되는 통계량에는 상호정보(mutual information, MI), T-Score, 로그우도비(Log-likelihood) 등이 있는데, 상호정보는 전체적인 빈도가 낮은 표현을 선호하는 경향이 있고 t-score는 전체적인 빈도가 높은 표현을 선호하는 경향이 있다. AntConc에서는 디폴트로 MI를 사용하는데, 이는 Top Preferences 메뉴에 들어가서 변경할 수 있다. Top Preferences - Collocates - Selected Collocate Measure를 T-Score로 변경하고, Window Span을 2L부터 2R까지로 설정해서 '많이'를 검색하면, 이에 대한 연어/공기어가 다음과 같이 나온다.

Concordance					Concordance Plot	File View	Clusters/N-Grams	Collocates	Word List	Keyword List
Total No. of Collocate Types: 144					Total No. of Collocate Tokens: 188					
Rank	Freq	Freq(L)	Freq(R)	Stat	Collocate					
1	9	5	4	2.73102	싫습니 다					
2	6	0	6	2.37890	있습니 다					
3	4	1	3	1.90531	싫 어요					
4	4	2	2	1.83120	한 국					
5	4	2	2	1.81062	그래 서					
6	3	0	3	1.65599	먹 고					
7	4	3	1	1.60888	저 는					
8	2	1	1	1.40839	과자 에					
9	2	0	2	1.40257	지었 어					
10	2	1	1	1.40257	도시 가					
11	2	1	1	1.39675	심리 학					
12	2	0	2	1.39675	놀 고					
13	2	1	1	1.39092	시민 들이					
14	2	0	2	1.37346	된 다					
15	2	1	1	1.36763	음식 을					
16	2	2	0	1.33270	note					
17	2	1	1	1.32106	나 는					
18	2	2	0	1.31523	친구 하고					
19	2	0	2	1.31523	좋아합니 다					
20	2	1	1	1.29194	우리					
21	2	0	2	1.28612	하 고					
22	2	0	2	1.22790	가 고					
23	2	2	0	1.13474	그리 고					
24	2	0	2	1.11145	있 다					
25	1	0	1	0.99177	했지 만					
26	1	1	0	0.99177	한편 으로					
27	1	0	1	0.99177	한국 어로					
28	1	1	0	0.99177	학 교에 도					
29	1	1	0	0.99177	하 우스					
30	1	1	0	0.99177	태 도의					
31	1	1	0	0.99177	타이베 이는					
32	1	0	1	0.99177	책이					

Search Term

☐ Words ☐ Case ☐ Regex

많이

Advanced

Start

Stop

Sort

Sort by

☐ Invert Order

Sort by Stat

Window Span

☐ Same

From...

2L

To...

2R

Min. Collocate Frequency

1

<그림 109> '많이'의 연어 분석 결과

1.5.10. N-Gram 분석

우리가 하나의 어절이 아니라 두 어절(또는 그 이상)의 연쇄에 관심을 가질 수도 있다. 예컨대 ‘철수는 밥을 먹고 있고 나는 빵을 먹고 있고 영화는 굶고 있다’라는 문장에는 ‘철수는 밥을’, ‘밥을 먹고’, ‘먹고 있고’, ‘있고 나는’ 등 10개의 bigram token 이 존재하며(일반적으로 n개 어절(unigram)로 이루어진 텍스트에는 n-1개의 bigram 이 존재함), ‘먹고 있고’라는 bigram token이 2개 있으므로 bigram type은 9개가 존재한다. 이런 식의 조사는 bigram뿐 아니라 trigram, four-gram 등 n-gram 일반에 대해서도 얼마든지 할 수 있다. 어떤 말뭉치에 어떤 n-gram들이 존재하는지 알고 싶을 때에는 Clusters/N-Grams 메뉴가 유용하다. 이 메뉴에 들어가서 Search Terms 오른쪽에 있는 N-Grams를 선택하고 N-Gram Size를 Min, Max 둘 다 2로 설정한 상태에서 Start 버튼을 누르면 bigram들의 빈도와 Range가 결과로 나온다(<그림 110>). Range는 해당 bigram이 몇 개의 파일에 출현하는지를 나타낸다. Range 값이 크면, 해당 bigram이 특정 파일에만 몰려 있는 것이 아니라 여러 파일에 폭넓게 분포하고 있다는 뜻으로 이해할 수 있다.

Concordance Plot File View Clusters/N-Grams Collocates Word List Keyword List			
Total No. of N-Gram Types 4907			Total No. of N-Gram Tokens 5655
Rank	Freq	Range	N-gram
1	29	10	싫습니다 저는
2	22	12	가고 싶습니다
3	21	9	수 있다
4	12	2	punctuation 가운데점
5	11	1	가운데점 punctuation
6	10	7	보고 싶습니다
7	10	9	싫습니다 그리고
8	10	8	저는 한국
9	10	7	친구하고 같이
10	10	6	하고 싶습니다
11	10	7	할 수
12	9	6	먹고 싶습니다
13	8	5	만나고 싶습니다
14	8	6	싫습니다 그래서
15	7	3	것 같다
16	7	4	올바른 인터넷
17	7	4	인터넷 사용
18	7	3	친구 같이
19	7	5	한다고 생각한다
20	6	2	가고 싶어요
21	6	3	가장 중요한
22	6	6	공무하고 싶습니다
23	6	4	상대방의 의견을
24	6	4	알 수
25	6	6	저는 올해
26	6	3	저는 한국어를
27	5	3	note 영어로
28	5	4	것을 알
29	5	2	경제 개발
30	5	3	그대로 입력함
31	5	5	많이 있습니다

Search Term ☒ Words ☐ Case ☐ Regex ☒ N-Grams

Min. 2 Max. 2

Start Stop Sort

Sort by ☐ Invert Order Search Term Position

Sort by Freq ☒ On Left ☐ On Right

Min. Freq. 1 Min. Range 1

<그림 110> bigram 분석 결과

1.5.11. 키워드 분석

그 다음으로 AntConc에서 제공하는 유용한 기능으로 Keyword List 메뉴가 있다. 어떤 대규모 말뭉치를 기준으로 해서, 관심의 대상이 되는 특정 말뭉치나 텍스트에 출현하는 단어들 중 기준 말뭉치에 비해 유난히 빈도가 높거나 유난히 빈도가 낮은 단어들이 무엇인지 알아볼 수 있다. 이런 단어들은 Keyword라고 하는데, 해당 말뭉치/텍스트의 특징을 알아보는 데 유용한 단서가 될 수 있다. 예컨대 물리학 텍스트에는 ‘에너지’, ‘운동’, ‘속도’, ‘가속도’, ‘입자’ 등의 단어들이 일반 말뭉치에 비해 월등히 높은 빈도로 나타날 것이고, 법률 텍스트에는 ‘소송’, ‘유죄’, ‘판결’ 등의 단어들이 자주 나타날 것이다. 이러한 키워드 분석을 하기 위해서는 우선 기준이 되는 대규모 말뭉치의 단어 빈도 목록이 필요하다. 영어의 경우 British National Corpus(BNC) 등의 단어 빈도 목록을 위의 AntConc 홈페이지 등에서 쉽게 구할 수 있다. 한국어도 세종 문어 말뭉치 등을 기준 말뭉치로 삼아서, 특정 말뭉치/텍스트의 키워드 분석을 행할 수 있을 것이다. 학습자 말뭉치가 어느 정도 구축되면, 이를 대상으로 해서 키워드 분석을 할 수 있다. 이를 통해 외국인 학습자들이 한국어 네이티브에 비해 어떤 표현을 유난히 자주 사용하는지, 그리고 어떤 표현을 유독 사용하지 않는지 등을 알 수 있을 것이다.

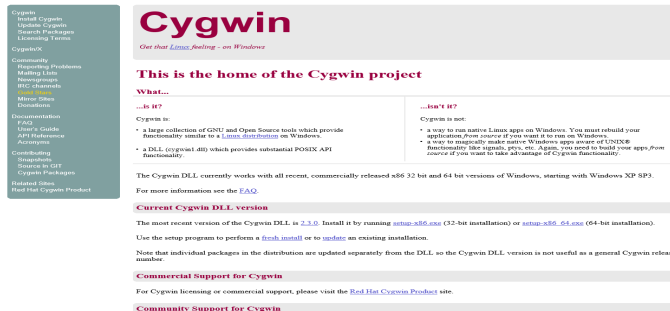
* * *

여기서는 AntConc를 중심으로 말뭉치 활용 방안을 살펴보았다. 일반적인 말뭉치 분석 도구는 제공하는 기능이 거의 비슷하기 때문에, AntConc 외의 다른 소프트웨어를 사용하더라도 사용법은 크게 다르지 않을 것이다. 그리고 이들 소프트웨어는 학습자 말뭉치에 특화된 것은 아니기 때문에, 학습자 말뭉치만을 위한 특별한 기능을 제공하고 있지는 않다. 그러나 학습자 말뭉치도 AntConc의 여러 기능들을 유용하게 이용할 수 있다.

1.6. Cygwin을 이용한 말뭉치 활용

1.6.1. Cygwin 설치

Cygwin은 사실은 하나의 소프트웨어라기보다 많은 소프트웨어들을 묶어 놓은 패키지이다. 콘솔(윈도 운영체제에서는 명령프롬프트라고도 함)에서 키보드를 통해 명령어를 입력하는 방식으로 여러 가지 텍스트 처리를 할 수 있다. Cygwin 패키지는 인터넷에서 무료로 다운로드할 수 있다. 자기 컴퓨터가 32비트인지 64비트인지에 따라 적절한 것을 선택하여 설치하면 된다.

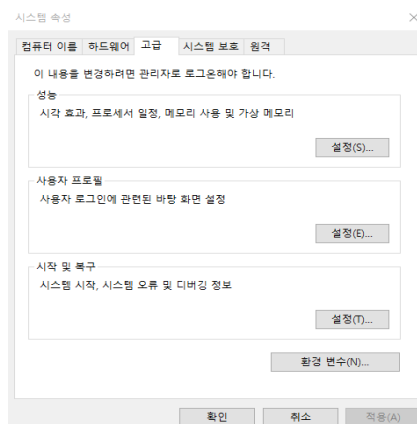


<그림 111> Cygwin 홈페이지(http://cygwin.com) 초기 화면

1.6.2. Cygwin 사용을 위한 설정

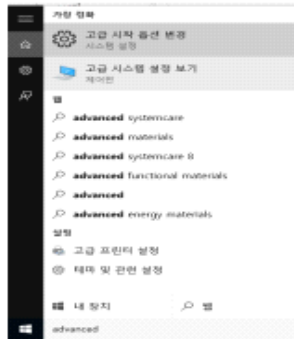
Cygwin에서 문자 인코딩의 default는 UTF-8이다. input 텍스트 파일이 UTF-8으로 되어 있을 것을 전제로 하고, output 텍스트 파일도 UTF-8으로 만들어진다.

Cygwin을 설치할 때 설치 디렉토리를 특별히 다르게 지정하지 않았다면 아마 C:\cygwin(64비트 컴퓨터의 경우는 C:\cygwin64) 폴더에 설치되었을 것이다. 모든 실행 파일들은 그 아래의 bin 폴더에 들어 있다. 명령프롬프트의 어느 디렉토리에서든 Cygwin 실행 파일들을 사용하고자 한다면 path를 설정해야 한다. '설정(또는 제어판)' - '시스템 정보' - '고급 시스템 설정'을 선택하면 다음과 같은 팝업창이 뜬다.



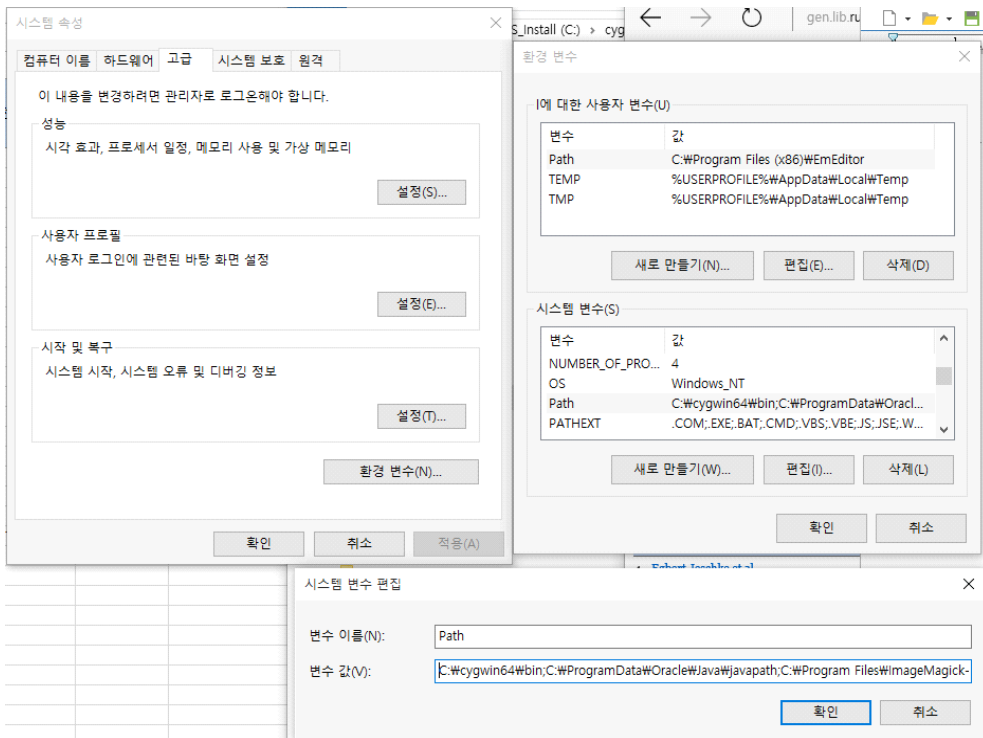
<그림 112> 고급 시스템 설정

윈도우 10의 경우 바탕화면의 좌측 하단에 있는 '웹 및 Windows 검색'에 'advanced'라고 입력하여 뜨는 팝업 메뉴(<그림 112>)에서 둘째 '고급 시스템 설정 보기'를 클릭하면 <그림 113>의 팝업창이 곧바로 뜬다.



<그림 113> 윈도우 10 ‘고급 시스템 설정’ 들어가기

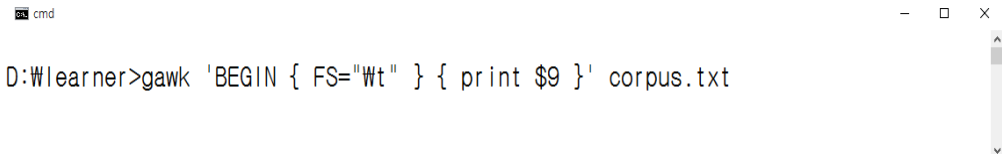
<그림 112>의 ‘시스템 속성’ 팝업창에서 ‘환경 변수’ 버튼을 클릭하면 ‘환경 변수’ 팝업창이 뜨고 여기서 ‘시스템 변수’ - ‘Path’를 선택하여 ‘편집’ 버튼을 클릭하면 ‘시스템 변수 편집’ 팝업창이 뜬다. 여기서 ‘변수 값’ 입력란의 맨 앞에 Cygwin 실행 파일들이 들어 있는 디렉토리(C:\cygwin\bin; 또는 C:\cygwin64\bin;)를 써 주고 계속 ‘확인’ 버튼을 누르면 된다(<그림 114>).



<그림 114> ‘환경 변수’, ‘시스템 변수 편집’ 팝업창

1.6.3. gawk로 특정 필드 추출하기

인코딩이 UTF-8으로 되어 있는 텍스트 파일에서, 오류 층위의 유형별 빈도를 추출하고 싶다고 하자. 우선, 콘솔에서 해당 파일이 들어 있는 디렉토리(D:\learner)로 이동하여(cd D:\learner) 아래와 같이 타이핑하고 엔터를 친다.



<그림 115> 콘솔에서 gawk 실행 방법

gawk에서는 따옴표(‘ ’) 안에 코드를 써 주고 그 뒤에 입력 파일명을 써 준다. 코드 부분은 BEGIN, MAIN, END의 세 부분으로 이루어지는데, BEGIN과 END는 수의적이다. <그림 115>에서는 BEGIN과 MAIN으로만 이루어져 있다. (END 부분이 나오는 예는 뒤에서 살펴본다.) BEGIN 부분에서는 입력 파일을 본격적으로 처리하기 전에, 입력 파일의 포맷에 대해 프로그램에게 알려주고 필요한 세팅을 한다. 입력 파일의 필드 구분자(field separator, FS), 레코드 구분자(record separator, RS)가 무엇인지 알려주는 경우가 많다. FS와 RS를 지정하지 않으면 default로 이해되는데, RS의 default는 줄바꿈문자(“\n”)이고, FS의 default는 줄바꿈문자 이외의 white space(공백 “ ” 또는 탭 “\t”)이다. 우리의 입력 파일의 경우, RS는 default대로 줄바꿈문자이므로 지정해 줄 필요가 없으나, FS는 공백이 아니라 탭이기 때문에 지정해 주어야 한다.(FS=“\t”)

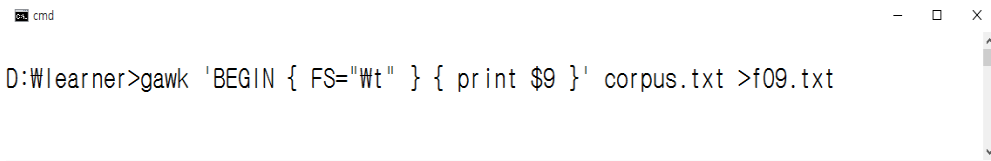
<그림 115>의 MAIN 부분은 “print \$9”라는 하나의 문장으로 이루어져 있다. gawk에서 각 필드는 다음과 같은 약호로 지칭한다.

\$1	첫째 필드
\$2	둘째 필드
.....	
\$n	n번째 필드
\$0	해당 레코드 전체

따라서 “print \$9”는 아홉째 필드를 출력하라는 뜻이다.

1.6.4. redirection으로 출력을 파일로 저장하기

<그림 115>의 명령을 실행하면, 입력 파일의 아홉째 필드가 콘솔 화면으로 출력된다. 이 출력을 파일로 저장하려면 redirection을 이용한다. '>' 뒤에 출력 파일명을 정해서 적어 주면 된다.

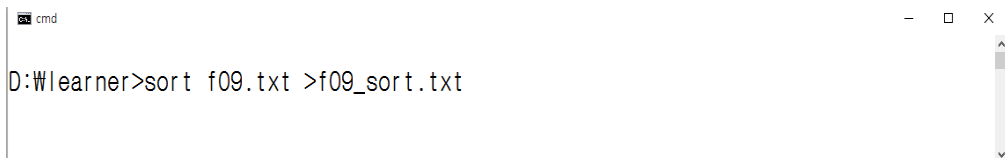
A terminal window titled 'cmd' with a standard Windows command prompt interface. The command entered is 'D:\Wlearner>gawk 'BEGIN { FS="Wt" } { print \$9 }' corpus.txt >f09.txt'. The command is shown on a single line, and the prompt is visible at the end of the line.

```
D:\Wlearner>gawk 'BEGIN { FS="Wt" } { print $9 }' corpus.txt >f09.txt
```

<그림 116> gawk 실행 결과를 파일로 저장

1.6.5. sort

f09.txt를 소트하면 같은 유형의 오류 층위들이 한 곳에 모일 것이다. sort 명령어를 이용하면 되며, 역시 그 결과를 파일로 저장하려면 redirection을 이용한다.

A terminal window titled 'cmd' with a standard Windows command prompt interface. The command entered is 'D:\Wlearner>sort f09.txt >f09_sort.txt'. The command is shown on a single line, and the prompt is visible at the end of the line.

```
D:\Wlearner>sort f09.txt >f09_sort.txt
```

<그림 117> sort 실행 방법

1.6.6. uniq

이제 동일한 라인들이 연달아 나오게 되는데, 이것을 하나로 합치면서(uniq 명령어), 합친 것이 몇 개인지를 count(-c 스위치)해 준다.

```
cmd
D:\Wlearner>uniq -c f09_sort.txt >f09_count.txt
```

<그림 118> uniq 실행 방법

그 결과는 다음과 같다.

u8.txt	f09_count.txt
1	10999 ↓
2	4 DC ↓
3	5 DS ↓
4	1 GC ↓
5	135 GE ↓
6	16 GF ↓
7	4 GN ↓
8	242 GPT ↓
9	105 GS ↓
10	14 GT ↓
11	8 VMA ↓
12	21 VMM ↓
13	166 VNN ↓
14	10 VNP ↓
15	4 VNR ↓
16	57 VV ↓
17	23 VVA ↓
18	10 VVC ↓
19	1 VVV ↓
20	1 오류 총위 ↓
21	←

<그림 119> uniq
실행 결과

아무런 오류 총위 표시가 없는 라인이 10999개이고, 오류 총위들이 알파벳 순서대로 배열되어 있고 각 오류 총위들의 빈도가 나와 있다.

1.6.7. sort의 옵션들

이것을 빈도순대로 소트하려면 sort 명령어의 -g 스위치(문자열이 아니라 숫

자의 크기를 기준으로 소트)와 -r 스위치(오름차순이 아니라 내림차순으로 배열)를 이용하면 된다.

```
cmd
D:\Wlearner>sort -gr f09_count.txt >f09_count_sort.txt
```

<그림 120> 숫자 내림차순 sort

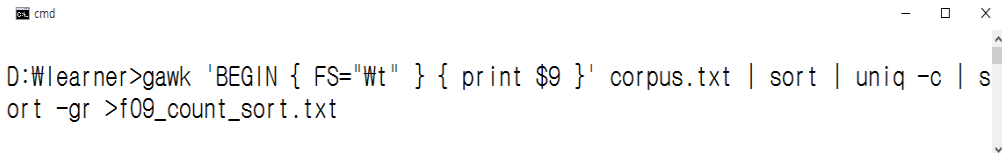
그 결과는 다음과 같다. GPT(조사)가 242개로 가장 많고, VNN 166개, GE(어미) 135개, GS(문장) 105, VV 57개 등이 뒤를 잇고 있다.

	f09_count_sort.txt
1	10999 ↓
2	242 GPT↓
3	166 VNN↓
4	135 GE↓
5	105 GS↓
6	57 VV↓
7	23 VVA↓
8	21 VMM↓
9	16 GF↓
10	14 GT↓
11	10 VVC↓
12	10 VNP↓
13	8 VMA↓
14	5 DS↓
15	4 VNR↓
16	4 GN↓
17	4 DC↓
18	1 오류 총위↓
19	1 VVV↓
20	1 GC↓
21	←

<그림 121> 숫자
내림차순 sort 결과

1.6.8. 파이프를 여러 명령어를 연결하기

지금까지는 각 단계별로 출력 파일을 저장하여 그 다음 단계 작업의 입력으로 삼았는데, 파이프(|)를 이용하면 중간 단계 결과 파일을 저장할 필요 없이 작업을 한 번에 수행할 수 있다. 파이프는 명령어들을 연결하는데, 앞 명령어의 출력을 뒤 명령어의 입력으로 삼으라는 의미이다.



```
cmd
D:\Wlearner>gawk 'BEGIN { FS="Wt" } { print $9 }' corpus.txt | sort | uniq -c | sort -gr >f09_count_sort.txt
```

<그림 122> gawk, sort, uniq, sort를 파이프를 연결하여 실행하는 방법

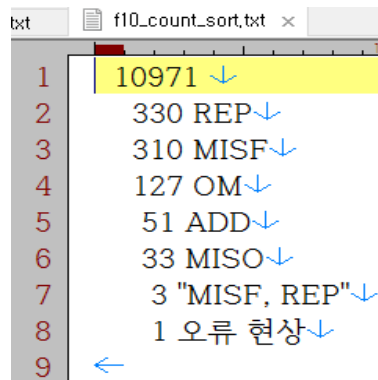
10째 필드에 있는 오류 현상에 대해서도 같은 작업을 같은 방법으로 할 수 있다.



```
cmd
D:\Wlearner>gawk 'BEGIN { FS="Wt" } { print $10 }' corpus.txt | sort | uniq -c | sort -gr >f10_count_sort.txt
```

<그림 123> 오류 현상(10째 필드)에 대한 빈도 통계를 내는 방법

그 결과는 다음과 같다. 오류 현상에 대한 표시가 없는 라인이 10971개이고, REP(대치) 330개, MISF(오형태) 310개, OM(누락) 127개 등으로 뒤를 잇고 있다.



1	10971 ↓
2	330 REP↓
3	310 MISF↓
4	127 OM↓
5	51 ADD↓
6	33 MISO↓
7	3 "MISF, REP"↓
8	1 오류 현상↓
9	←

<그림 124> 오류 현상 통계

1.6.9. 특정 필드에 조건을 지정하여 조건을 충족하는 레코드 추출하기

앞에서 엑셀을 가지고 해 보았던, 오류 층위 GPT(조사 오류)의 하위유형 빈도 조사를 Cygwin으로 해 보자. 올바른 조사는 7째 필드에 나와 있고 학습자가 잘못 쓴 조사는 4째 필드에 나와 있다. 9째 필드가 GPT인 레코드만 뽑아내서 4, 7째 필드만 출력하는 것은 다음과 같이 gawk로 할 수 있다.

```
cmd
D:\Wlearner>gawk 'BEGIN { FS="Wt" ; OFS="Wt" } $9~/GPT/{ print $4,$7 }' corpus.txt >f04_07.txt
```

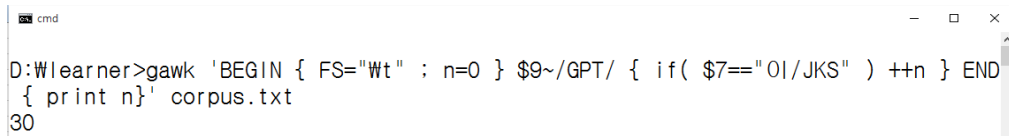
<그림 125> gawk로 조건에 맞는 레코드만 추출하기

그 결과는 다음과 같다. 둘째 필드가 올바른 조사이고, 첫째 필드는 학습자가 잘못 사용한 조사이다.

1	로/JKB→	으로/JKB↓
2	→	을/JKO↓
3	씩/XSN→	이라도/JX↓
4	→	을/JKO↓
5	와/JKB→	과/JKB↓
6	예/JKB→	를/JKO↓
7	→	가/JKS↓
8	에게/JKB→	예/JKB↓
9	때/NNG→	예/JKB↓
10	→	를/JKO↓
11	든/NA→	를/JKO↓
12	가/JKS→	이/VCP↓
13	→	이/ETM↓
14	→	이/JKS↓
15	→	를/JKO↓
16	을/JKO→	이/JKS↓
17	한테/JKB→	의/JKG↓
18	보다/JKB→	가/JKS↓
19	이/JKS→	을/JKO↓
20	은/JX→	을/JKO↓
21	은/JX→	이/JKS↓
22	는/JX→	가/JKS↓
23	이/JKS→	예서/JKB↓
24	예/JKB→	↓
25	→	예서/JKB↓
26	→	가/JKS↓
27	이/JKS→	을/JKO↓
28	→	가/JKS↓
29	→	을/JKO↓
30	→	을/JKO↓
31	이/JKS→	을/JKO↓
32	이/JKS→	은/JX↓
33	→	가/JKS↓

<그림 126> GPT인 레코드만 추출한 결과

앞에서 엑셀로 해 보았던, ‘오류 층위’가 ‘GPT’이고 ‘교정 어절의 형태 주석’이 ‘이/JKS’인 레코드를 세는 작업을 gawk로 해 보자(<그림 127>).

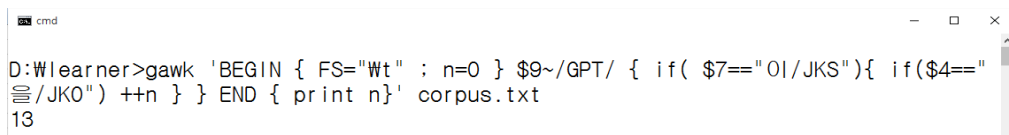


```
cmd
D:\Wlearner>gawk 'BEGIN { FS="Wt " ; n=0 } $9~/GPT/ { if( $7=="이/JKS" ) ++n } END { print n }' corpus.txt
30
```

<그림 127> gawk로 특정 기준을 충족하는 레코드 추출. 기준이 2개인 경우

우선 BEGIN 부분에서 필드 구분자를 탭으로 지정해 주고, 빈도를 센 결과를 저장할 변수 n을 0으로 초기화한다. MAIN 부분에서는 9째 필드(\$9)가 ‘GPT’이고 7째 필드(\$7)가 ‘이/JKS’인 경우 변수 n을 1씩 증가시킨다. 모든 레코드에 대해 이런 일이 반복 실행된 뒤, END 부분에서는 변수 n을 출력한다.

주격조사 ‘이/JKS’를 써야 하는데 대격조사 ‘을/JKO’를 잘못 쓴 조사 오류를 세려면 <그림 128>과 같이 하면 된다. 조건이 2개에서 3개로 늘었을 뿐 프로그램의 구조는 <그림 127>과 동일하다.

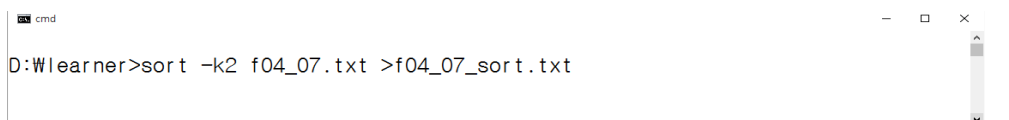


```
cmd
D:\Wlearner>gawk 'BEGIN { FS="Wt " ; n=0 } $9~/GPT/ { if( $7=="이/JKS"){ if($4=="을/JKO") ++n } } END { print n }' corpus.txt
13
```

<그림 128> gawk로 특정 기준을 충족하는 레코드 추출. 기준이 3개인 경우

1.6.10. sort에서 기준 필드 지정하기

우선 올바른 조사를 기준으로 한데 모으고 그 다음에 학습자가 잘못 사용한 조사를 기준으로 소트하면, 한국어의 각 조사를 학습자들이 어떻게 잘못 사용하고 있는지를 일목요연하게 알 수 있을 것이다. 이것은 sort 명령어에서 정렬 기준이 되는 필드를 지정하는 스위치(-k)를 이용하여 수행할 수 있다.



```
cmd
D:\Wlearner>sort -k2 f04_07.txt >f04_07_sort.txt
```

<그림 129> sort 명령어에서 기준 필드 지정

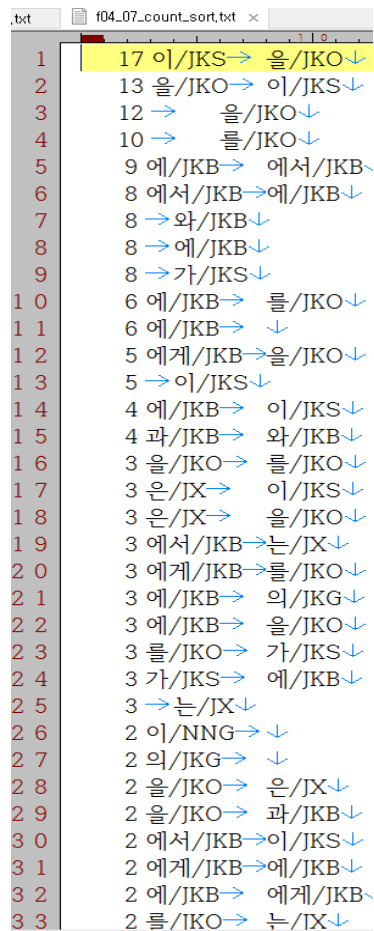
1.6.11. uniq -c를 이용한 통계 추출

이렇게 정렬된 결과로부터 uniq -c를 이용하여 빈도 통계를 낼 수 있다.

```
cmd
D:\Wlearner>uniq -c f04_07_sort.txt | sort -gr >f04_07_count_sort.txt
```

<그림 130> uniq -c와 sort -gr을 이용한 빈도 통계 추출

그 결과는 다음과 같다.



<그림 131> 조사 오류 유형 통계

즉, ‘을’을 ‘이’로 잘못 쓴 것이 17개, ‘이’를 ‘을’로 잘못 쓴 것이 13개, ‘을’을 누락한 것이 12개, ‘를’을 누락한 것이 10개로서, 주격조사와 대격조사의 오용이 1~4위를 차지하고 있다. 그 다음에 ‘에’와 ‘에서’의 혼동이 5, 6위이고, ‘와, 에, 가’의 누락이 7~9위이다. 이러한 통계를 보면, 학습자가 흔히 저지르는 조사 오류의 유형을 알 수 있다.

테이블 구조로 된 학습자 말뭉치에서 다른 필드에 대해서도 비슷한 방식으로 얼마든지 필요한 정보를 추출하고 통계를 낼 수 있을 것이다.

V. 한국어 학습자 말뭉치 구어 전사 지원 도구 개발 기초 연구

1. 연구 개요

1.1. 연구의 목적 및 필요성

1.1.1. 연구의 목적

최근 외국어교육에서 의사소통능력이 강조됨에 따라 실제 발화에서 사용되는 표현이나 변이형, 각종 의사소통 전략 등을 파악하려는 요구가 증가하게 되었고 구어 말뭉치의 필요성 또한 대두되었다. 구어 말뭉치는 녹음 파일을 수집하고 텍스트로 전사해야 한다는 점에서 문어 말뭉치에 비해 구축 과정이 까다롭다. 특히 외국인 학습자를 대상으로 수집한 구어 자료는 모국어 화자의 구어 자료에 비해 전사에 많은 어려움이 따른다. 따라서 작업의 효율을 높이기 위한 방안으로 음성 파일을 자동으로 글말로 변환해 주는 기능, 구어 전사용 주석(마크업) 등의 편집/처리 기능 등을 가진 학습자 말뭉치용 구어 전사 지원 도구의 개발 가능성을 검토하는 데에 연구의 목적이 있다.

1.1.2. 연구의 필요성

학습자 말뭉치 구어 전사 도구의 필요성은 다음과 같이 요약할 수 있다.

첫째로, 외국인 학습자라는 특수성을 고려해야 한다. 현재 국내에는 학습자의 구어 말뭉치를 전사하는 학습자 전용 구어 전사 도구가 전무한 실정이다. 외국인 학습자는 모국어 화자와 달리 발음이 불분명한 경우가 많아 기존의 자동 전사 도구를 사용하여 작업을 진행하기에는 어려움이 따른다. 그 중 가장 큰 문제점은 학습자가 산출한 발음이 불분명하거나 표준어 규정에 맞지 않을 수 있다는 점이다. 음운적 요소뿐 아니라 운율적 요소에서도 예측 불가능한 다양한 변이형으로 나타나므로 전사자들은 자료 입력 단계에서만도 많은 시간과 노력을 들여야 한다. 따라서 학습자 전용 구어 전사 자동/반자동화 시스템이 개발된다면 학습자 구어 말뭉치 구축에 큰 도움이 될 것이다.

둘째로, 현재 사용하고 있는 구어 전사 도구의 작업 과정이 복잡하고 전문적이라는 점을 들 수 있다.

일례로 네덜란드에서 개발한 전사 도구 ELAN(EUDICO Linguistic Annotator)에는 다음과 같은 다양한 음성 처리 관련 기능 등을 포함하고 있다.⁴⁾

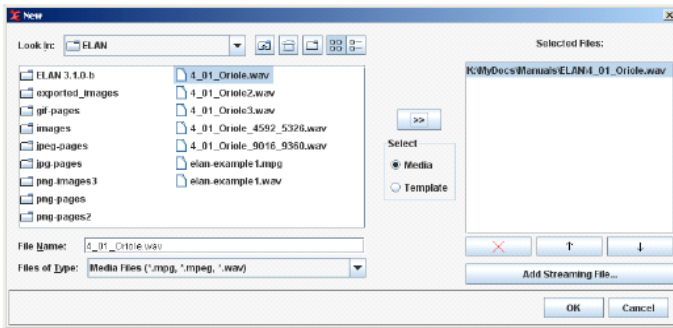


Figure 1: select files in the “New” dialog

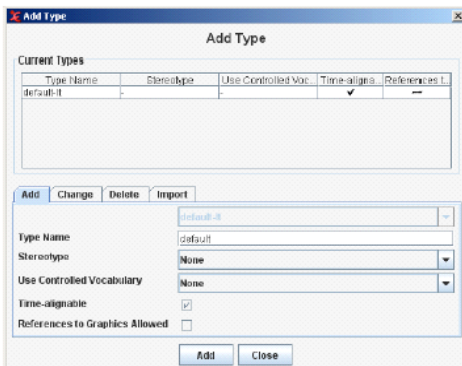


Figure 2: add a linguistic type

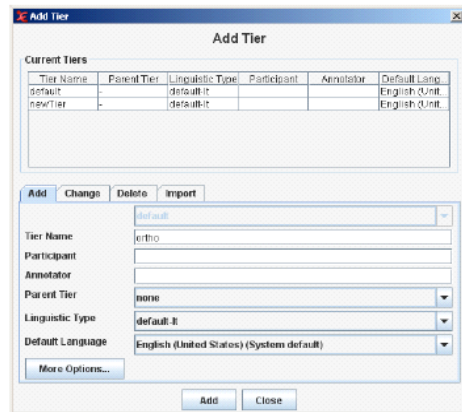


Figure 3: add a tier

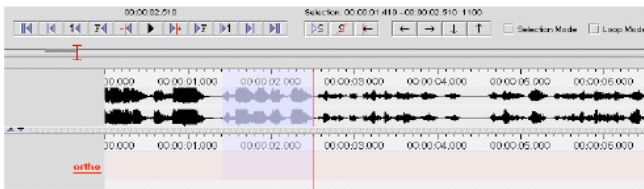


Figure 4: select an interval by dragging in the timeline viewer

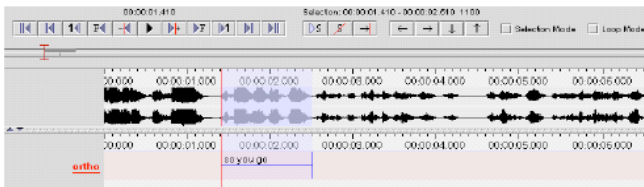


Figure 5: the active annotation

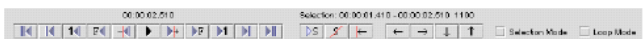


Figure 6: the media player-, selection- and annotation navigation controls

<그림 132> ELAN의 음성 파일 처리 기능의 일부

- 4) 자료의 전체 내용은 ELAN의 공식 홈페이지에서 확인할 수 있다.
(<https://tla.mpi.nl/tools/tla-tools/elan/>)

발화를 분할하여 텍스트화하고 주석을 부가하는 일련의 과정은 복잡하고, 전문성을 요하는 작업이다. 데이터 품질과 작업의 효율성을 개선하기 위해서는 접근성이 좋은 구어 전사 도구의 개발이 요구된다.

셋째로, 구어 전사 자료의 일관성을 위해서이다. 현재는 작업자마다 다른 기준을 가지고 구어 전사 작업을 하고 있기 때문에 전사 결과가 일관성을 가지지 못한다. 예를 들어 학습자가 ‘커피’를 /코피/라고 발음하였을 경우, 어떤 작업자는 이를 ‘커피’라고 전사할 수도 있고, 또 어떤 작업자는 ‘코피’라고 전사할 수도 있다. 작업 숙련도가 높을수록, 외국인 학습자에 대한 노출이 많을수록 발음 용인의 정도는 커진다. 따라서 자동으로 구어 전사 작업이 이루어진다면 결과의 일관성을 높일 수 있고 자료를 처리하는 작업 또한 용이해진다.

이밖에도 현재 사용하고 있는 대부분의 구어 전사 도구는 외국에서 개발한 도구이기 때문에 국내 실정에 맞지 않는다는 점 또한 중요한 요인으로 손꼽힌다. 한국어를 기반으로 하는 구어 전사 도구를 개발한다면 학습자 말뭉치뿐 아니라 향후 다양한 말뭉치 구축에 있어 구어 자료 활용에 기여할 것이다.

1.2. 연구의 방향

본 과제에서는 한국어 학습자 말뭉치 구어 전사 자동/반자동화 도구 개발을 위한 초석으로 기존의 자동 음성 인식 소프트웨어와 자동 음성 전사 도구가 어떤 서비스를 제공하는지 분석할 것이다. 그리고 외국어교육에서 이들 도구가 어떻게 사용되고 있는지 조사하여 한국어 학습자를 위한 음성 인식 도구를 개발할 때 유의할 점이 무엇인지 살펴보고자 하겠다.

또한 기존의 도구를 사용하였을 때 외국인 학습자의 STT(Speech-To-Text)가 어느 정도 가능한지 실험을 통해 살펴보고 도구의 개선, 발전 방향을 제시할 것이다.

마지막으로 실제 구어 전사 작업을 진행할 때 필요한 전사 표지의 개선 방향을 제안하고자 한다.

2. 도구 개발을 위한 기초 연구

2장에서는 공개/상용 음성 인식 도구와 음성 전사 도구에는 어떤 것들이 있는지 알아보겠다. 하지만 국내 및 해외에서 이들 도구를 학습자 말뭉치를 분석하는데 활용한 사례는 찾아볼 수 없었다. 따라서 본 연구에서는 음성 인식 도구를 외국어 교육에 활용한 사례를 몇몇 논문을 통해 살펴보고자 한다. 선행 연구에서 지적인 도구의 한계점을 고려한다면 외국인 학습자를 대상으로 하는 음성 인식/전사 도구를 개발할 때의 요구 사항을 도출할 수 있을 것이다.

2.1. 음성 인식 및 전사 도구

2.1.1. 공개/상용 음성 인식 도구

이름	기술 체계	웹사이트	유형
CMU Sphinx	HMM	http://cmusphinx.sourceforge.net	오픈 소스
HTK	HMM	http://htk.eng.cam.ac.uk/	오픈 소스
Julius	HMM trigrams	http://julius.osdn.jp/en_index.php	오픈 소스
Kaldi	Deep neural net	http://kaldi.sourceforge.net/about.html	오픈 소스
iATROS	LDA (Latent Dirichlet)	https://www.prhlt.upv.es/software/iatros/doc/speech	오픈 소스
RWTH ASR	RWTH Aachen University	http://www-i6.informatik.rwth-aachen.de/rwth-asr	오픈 소스
Simon	Supports Sphinx, HTK, Julius	https://simon.kde.org	오픈 소스
Dragon Dictate	Mac OS	http://www.nuancemobilelife.com/apps/dragon-dictation/	상용 (apple)
MacSpeech Dictate Medical	Medical dictation product	http://www.nuance.com/for-individuals/by-product/dragon-for-mac/dictate-medical/index.htm	상용 (apple)
Macspeech Dictate Legal	Legal-focused dictation	http://www.nuance.com/for-individuals/by-product/dragon-for-mac/mac-speech-dictate-legal/index.htm	상용 (apple)
MacSpeech Scribe	Transcription from recorded text	http://www.nuance.com/macspeech/	상용 (apple)
iListen	PowerPC Macintosh	http://www.macworld.com/article/1	상용

		001725/ilisten.html	(apple)
Speakable items	Included with Mac OS	http://www.macworld.com/article/1165635/get_siri_like_functionality_on_your_mac_with_speakable_items.html	상용 (apple)
ViaVoice	IBM Product. Support ended 2007.	http://www-01.ibm.com/software/pervasive/viavoice.html	상용 (apple)
Vestec Inc.	ASR, NLU, TTS, VSLIC	http://www.vestec.com/	상용 (apple)
Indigo	Virtual Assistant for Android, iOS, and WP, by Artificial Solutions	http://www.hello-indigo.com/	상용 (모바일)
Textshark	Cloud-/API-based speech-to-text transcription	http://www.textshark.com/	상용 (모바일)
Truly Handsfree	Embedded speech recognition for wakeup and command and control	http://www.sensory.com/	상용 (모바일)
TrulyNatural	Embedded large vocabulary speech recognition for natural language	http://www.sensory.com/	상용 (모바일)
S-voice	Samsung Galaxy's Voice based personal assistant	http://www.androidcentral.com/tag/s-voice	상용 (모바일)
Verbio ASR embedded		http://www.verbio.com/es/soluciones/12-speech-recognition.html	상용 (모바일)
Google Now	Android voice search	http://www.androidcentral.com/google-now	상용 (모바일)
Microsoft Cortana	Microsoft voice search	http://www.microsoft.com/en-us/mobile/experiences/cortana/	상용 (모바일)
GoVivace	Cloud-based speech recognition	http://govivace.com/products/	상용 (모바일)
Siri Personal Assistant	Apple's virtual personal assistant	http://www.apple.com/ios/siri/	상용 (모바일)
MeMeMe Mobile	Cloud-based speech recognition	http://www.memememobile.com/	상용 (모바일)
SILVIA	Android and iOS	http://silvia4u.info/	상용 (모바일)
DriveSafe.ly		http://www.drivesafe.ly/	상용 (모바일)
Jeannie	Android	http://android.appstorm.net/reviews/utilities/meet-androids-own-personal-assistant-jeannie/	상용 (모바일)
Speaktoit	Assistant for Android,	https://assistant.ai/	상용

	iOS and Windows Phone		(모바일)
Microsoft Tell me	Windows Phone 7/8	https://www.microsoft.com/en-us/store/apps/tell-me/9wzdncrdhz7b	상용 (모바일)
Ask Ziggy	Windows Phone 7	https://www.microsoft.com/en-us/store/apps/ask-ziggy/9wzdncrfj2lg	상용 (모바일)
Vocre	iOS	https://itunes.apple.com/us/app/vocr-e-translate-voice-text/id454405637?mt=8	상용 (모바일)
Utter!	Voice to speech personal assistant	http://forum.xda-developers.com/showthread.php?t=1508195	상용 (모바일)

<표 14> 국내외 음성인식용 공개/상용 도구

이 외에도 Microsoft의 windows 계열에도 음성 인식 시스템(Microsoft speech recognition engine ver.8)이 내장되어 있으며, 이에 더하여 Voice Attack, VAC, Voice Finger, Dragon Naturally Speaking, Speech Gear Interact, e-Speaking, Microsoft Speech API 등의 시스템을 추가로 활용할 수 있다.

이러한 자동 음성 인식 도구는 음성 다이얼링이 가능한 전화나 가전제품, 음성으로 운행하는 자동차, 음성 인식 휴대 전화 등 음성 명령 제어 분야에서 어느 정도의 성과를 보이고 있다(서상규 · 한영균, 2002). 본 연구에서는 이러한 음성 인식 도구 중 외국어 교육에서 많이 쓰이는 도구 몇 가지를 분석하여 어떤 기능을 지니고 있는지 살펴보고자 하겠다.

2.1.2. 음성 전사 및 분석 도구

(1) Praat

네덜란드에서 개발한 음성 분석용 소프트웨어로 웹사이트에서 무료로 프로그램을 다운받을 수 있다. 스펙트로그램, 피치 등 기본적인 음성 분석 기능을 망라하고 있으며 LPC분석 및 재합성, PSOLA분석 및 재합성, 피치 변경 등을 할 수 있다. 그리고 스크립트 언어에 의해 자체 처리 과정을 정의할 수도 있다. 이외에도 멀티 레벨 레이블링 등 많은 기능을 가지고 있다.

(<http://www.fon.hum.uva.nl/praat>)

(2) Dragon Naturally Speaking

뉘앙스(Nuance) 커뮤니케이션즈의 음성 인식 소프트웨어로 8가지 언어(미국 영어, 영국 영어, 프랑스어, 독일어, 이탈리아어, 스페인어, 네덜란드어, 일본어)를 제공한다.

사용자는 Dragon Naturally Speaking을 이용하여 음성을 텍스트로 또는, 텍스트를 음성으로 변환할 수 있다. 또한 명령어를 입력시켜 프로그램을 실행할 수 있다.

(<http://www.nuance.com/dragon/index.htm>)

(3) Viavoice

IBM에서 개발한 음성 인식 소프트웨어로 2003년에 Nuance가 인수하였다. 사용자는 Viavoice를 이용하여 음성을 텍스트로 전환할 수 있으며, 텍스트의 수정 및 교정, 포맷 또한 가능하다. 사용자의 명령어 인식이 가능하므로 사용자는 손쉽게 문서나 보고서 등을 작성할 수 있다.

(<http://www-01.ibm.com/software/pervasive/viavoice.html>)

(4) Elan

ELAN(EUDICO Linguistic Annotator)은 네덜란드의 Max Planck Institute for Psycholinguistics에서 개발한 소프트웨어이다. 멀티미디어 데이터의 주석과 이용을 위한 기술을 지원하며 사용자는 비디오나 오디오 데이터에 주석을 달아 이를

시각화하고 검색할 수 있다. ELAN은 특히 언어, 수화, 제스처 등의 분석에 맞춰 디자인되었다. 작업은, 언어 타입과 티어를 정의하고, 시간 인터벌을 설정하고, 주석을 다는 단계로 이루어진다.

(<https://tla.mpi.nl/tools/tla-tools/elan>)

(5) Cooledit96

Syntrillium Software에서 제작한 음향 분석, 편집 도구로 신호의 파형, 스펙트로그램, 필터링 등을 수행할 수 있는 소프트웨어이다. 평가판을 무료로 다운로드하여 사용할 수 있으며 주파수 분석, 편집 정도의 분석까지에 적합하다. 현재 cooledit2000, cooledit Pro까지 나와 있다. cooledit Pro를 이용하면 소리 파일을 mp3형태로도 저장이 가능하다.

(<http://syntrillium-software-corporation.software.informer.com>)

(6) WaveSuffer

음성이나 음향 자료를 분석하고 전사할 수 있는 오픈 소스 소프트웨어이다. 사용자 목적에 맞게 변경을 가할 수 있으며 플러그인 방식으로 새로운 기능 추가 가능 등 확장성이 뛰어나다.

(<http://sourceforge.net/projects/wavesurfer>)

(7) EXMARaLDA

음성 및 영상 말뭉치 자료의 전사와 주석 작업을 위한 소프트웨어이다. 전사 및 주석 작업을 위한 툴(Partitur-Editor), 말뭉치 관리를 위한 툴(Corpus-Manager), 검색 및 분석을 위한 툴(EXAKT)로 구성되어 있다.

(<http://www.exmaralda.org/en>)

(8) MAUS

음성 자료에서 음성학적인 주석을 자동화하기 위한 소프트웨어이다. 음성 파일과 그것을 전사한 텍스트 파일을 읽어 들여 음운적 수준의 세그멘테이션과 주석을 가능하게 한다.

(<http://www.phonetik.uni-muenchen.de/forschung/Verbmobil/VM14.7eng.html>)

2.2. 국내외 음성 인식 적용 사례

자동 음성 인식(Automatic Speech Recognition) 소프트웨어가 발전함에 따라 외국어교육에서도 이를 활용하여 발음을 교수하게 되었고 이에 따른 연구도 활발히 진행되고 있다. 컴퓨터가 학습자의 발음을 듣고 점수를 매기거나 피드백을 주는 시스템을 교육 현장에 활용하는 것으로 국외에서는 이미 ASR에 대한 활용법뿐만 아니라 소프트웨어의 신뢰성, 문제점 등에 대한 연구도 적지 않게 이루어졌다.

다음에서는 선행 연구 몇 편을 분석하여 ASR(Automatic Speech Recognition)을 이용한 외국어교육 현황을 살펴보고, 그 한계점 및 극복 방안에 대해 살펴보고자 하겠다.

(1) ASR을 활용한 외국어교육

외국에서는 초등학교에서부터 대학교까지 전반적으로 ASR을 활용한 외국어교육이 곳곳에서 이루어지고 있다. ASR을 활용하여 학습자의 발음을 연습하고 평가하며 피드백을 얻는다.

Elimat(2014)는 요르단의 한 초등학교에서 ASR을 활용한 실험을 하였는데 내용을 정리하면 다음과 같다.

가. 개요

ASR 기술을 EFL 3학년 학생들의 발음 교육에 적용하여 정규 교육과 비교했을 때 어떤 효과가 있는지 알아보았다. 또한 ASR 발음 교육 시 가장 효과적인 교육방법이 어떤 것인지에 대해서도 제시하였다.

나. 활용 도구

Tell Me More Performance English를 사용하였는데 레벨별로 12과제를 시행하였다. 학생들의 발음 향상을 위한 단어 연습, 문장 연습, 일상 회화 연습 등이 포함되며 피드백과 점수가 제시되었다.

요르단의 한 초등학교에서 영어를 배우는 64명의 3학년 학생을 대상으로 실험하였다. 64명을 무작위로 네 그룹으로 나누었는데 세 그룹은 실험군, 한 그룹은 대조군이 되었다. 실험군을 대상으로 Tell Me More Performance English 프로그램을 사용하였으며 대조군에는 같은 선생님이 의한 정규 교육을 적용하였다.

다. 결론

ASR을 사용한 실험군과 교사가 직접 교수한 대조군의 평균 점수는 유의미한 차이를 보였으며 실험군의 점수가 더 우수하였다. 실험군의 점수 중 개별 연습, 짝을 이룬 연습, 그룹 연습 가운데에는 개별 연습의 점수가 가장 우수하였다. 이러한 결과는 ASR을 사용하는 것이 발음을 가르치고 배우는 데 정규 교육보다 더 많은 기회를 제공한다는 것을 시사한다.

교실에서 ASR을 사용하여 영어 발음 교육을 하면 아주 흥미로운 학습 환경이 된다. ASR을 사용한 경우, 개별적으로 발음 연습을 하는 것에 가장 적합하며, ASR은 각자가 자신에게 맞는 학습 속도로 실력을 향상시킬 수 있게 도움을 준다.

Elimat(2014)를 포함한 선행 연구에서 주로 언급한 ASR의 장점은, 학습자에게 많은 연습 기회를 제공한다는 점, 편안한 환경에서 연습할 수 있다는 점, 교수자 없이도 교육이 가능하다는 점이다. 하지만 ASR을 활용하여 외국인 학습자를 교육하고 평가하는 것이 과연 효과적인 교수법인지에 대해서는 아직 의견이 분분한 상태이다.

다음에서는 ASR의 신뢰도와 한계점을 다룬 논문들을 몇 편 살펴보고 이를 바탕으로 학습자 말뭉치의 음성 인식 도구를 설계할 때 유의할 점에 대해 정리하도록 하겠다.

(2) ASR의 신뢰도

ASR의 음성 인식과 평가에 대해 과연 그 결과가 신뢰할 만한 것인가에 대한 비판은 지속적으로 제기되고 있다. 음성 인식률의 정확성을 알아보기 위해서는 모국어 화자를 대상으로 하는 대조군 실험이 필요하며, 평가 결과의 신뢰성을 확인해 보기 위해서는 원어민 교수자의 평가와 ASR 도구의 평가 결과를 비교해 보아야 한다.

In-seok,K.(2006)는 ASR 소프트웨어 중 Fluspeak를 사용하여 그 신뢰성을 검증하였다.

가. 개요

영어 발음을 교수하는 데 있어 Fluspeak의 신뢰성을 검증하였다. ASR이 분석한 학생들의 발음 점수를 NES(Native English Speaking) 교수자의 분석과 비교하였다. ASR에서 역량의 상관 계수는 0에 가까워 인간이 분석한 정확도에 미치지 못하는 못했지만 이 소프트웨어는 학습자들이 발음 연습을 하는 데 매우 유용할

것이다.

나. 활용 도구

Fluspeak(IBM ViaVoice recognizer)를 이용하여 EFL 수업을 듣는 36명의 한국 대학생들을 대상으로 실험하였다. 자동 발음 점수화 시스템은 1,500명의 원어민과 200명의 영어 학습자의 발화를 기반으로 30,000개의 음소를 사용하여 개발되었다. 이는 200,000개 단어의 음파 정보와 발음 변이 양상을 기반으로 한 것이다.

다. 결론

ASR 소프트웨어와 NES가 평가하는 발화 오류율의 상관관계를 알아보고, 그 원인에 대해 고찰하였다. 단어 레벨에서의 상관계수는 0.56 정도였고 억양 레벨은 거의 0이었다. 억양에서 이렇듯 상관계수가 낮게 나온 주요 이유로는 여러 사람이 다양한 높낮이(pitch)를 발화하였다는 점을 들 수 있다. Fluspeak는 L2화자의 억양과 L1화자의 억양의 유사성을 백분율로 측정하기 때문에 높낮이가 달라졌을 경우 70.4라는 발화 오류율을 보였다.(NES는 1.9) 또 다른 이유로 Fluspeak는 억양 정확도만으로 오류를 측정하지만 NES는 학습자의 억양뿐 아니라 발화의 전반적인 유창성을 보고 점수를 측정한다는 점을 들 수 있다. 이를 해결하기 위해서는 NES도 (가능하다면) 학생들의 억양만을 측정하여 비교해 보아야 한다.

이렇듯 ASR 소프트웨어가 한계점을 지니는 것은 분명하지만 원어민 교사가 부재할 경우 ASR은 EFL 학생들에게 발음을 가르치는 귀중한 도구로 사용될 수 있다.

그리고 발음 스코어링 시스템인 FluSpeak나 그 밖의 ASR 소프트웨어에 있어 두 가지의 연구 방향을 생각할 수 있다. 첫 번째는 FluSpeak의 언어 인식 오류율을 시스템의 정확도로 규정한 실험이다. 언어 인식 오류율은 ASR 측정의 가장 중요한 지표 중의 하나로 널리 사용되는데, 예를 들어 대체, 삭제, 삽입된 단어를 실제 정확한 문장의 단어수로 나누는 방식이다. 비율이 낮을수록 언어 인식의 정확도가 높은 것인데 인식 오류율이 10%에 달하면 새로운 알고리즘의 개발이 필요하다. 두 번째로 ASR 소프트웨어가 원어민의 발음을 정확하게 판단하는가를 실험하는 것이 필요하다. FluSpeak의 원어민에 대한 언어 인식률이 80%에 달한다면 시스템이 신뢰할 만한 수준이라고 할 수 있다.

(3) ASR의 한계점

외국인 학습자의 경우 자동 음성 인식 소프트웨어의 음성 인식률은 낮고, 인식 오류율은 높게 나타난다는 한계점이 관찰된다. 또한 학습자들 가운데 소프트웨어

의 속지에 어려움을 느끼는 경우도 있었다.

Eskénazi, M.(1999)은 ASR의 유용성과 한계성에 관해 논의하였는데 내용을 정리하면 다음과 같다.

가. 개요

외국어 교육에서 학습자들이 발음을 향상시키는 데 있어 ASR이 유용한지에 대해 논의하였다. 외국어 교육의 원리와 ASR의 한계점, 그리고 ASR을 어떻게 이용할 것인가를 먼저 이야기하고 Carnegie Mellon FLUENCY 시스템에 어떤 기능이 있는지 알아보았다.

나. 활용 도구

Carnegie Mellon FLUENCY를 사용하였는데 다양한 레벨의 외국인 학습자 20명을 대상으로 하였다.

다. 결론

FLUENCY는 기존 ASR 시스템에서 학습자들이 수동적인 역할만 하여 많은 문장을 생성하지 못한다는 문제점을 보완하였다. 또한 이 도구를 사용하면 학습자들의 오류를 음성 확인 그래프로 확인할 수 있는데 말음절의 t라든지 음절 가운데 있는 r발음의 오류 등이 눈에 띄었다. L2의 음운이 L1의 그것과 유사한 경우 오류율이 적었고, 대부분 학습자의 모국어에 존재하지 않았던 /AE/와 같은 발음은 오류율이 높았다. 이러한 오류는 피드백을 통해 교정할 수 있다. 교실 학습의 보조로 컴퓨터를 사용하고 연습 시간을 늘린다면 외국어 교육에 많은 도움이 될 것이다.

Neri, A., Cucchiarini, C., Strik, H.(2003)는 ASR을 이용한 CALL(Computer assisted language learning)의 한계와 발전 가능성에 대해 논의하였다.

가. 개요

L2 발음을 교육하는 도구로서 ASR의 유용성에 관해 고찰하였다. 최신 기술인 ASR의 실질적인 문제는 신뢰할 만한 도구로 이를 어떻게 발전시키느냐 하는 것이다.

나. 활용도구 - 해당 없음

다. 결론

CALL 시스템은 추가적인 학습 시간을 제공하고 스트레스 없는 환경을 제공함으로써 인지도 있는 L2 발음 교육 도구로 자리매김하였다. 학습자는 컴퓨터와 상호작용을 할 수 있고 피드백 또한 제공받을 수 있어 전반적으로 ASR을 선호한다. 하지만 많은 연구자와 교수자들은 ASR을 통한 L2 교육에 회의적인데 그들이 지적한 ASR의 한계점은 다음과 같다.

ASR에서 가장 중요한 것은 도구가 음성 인식을 정확하게 해야 한다는 것이다. 선행 연구에 따르면 모국어 화자의 음성 인식 정확도가 90%를 기록했던 도구를 외국인 화자의 음성 인식에 사용했더니 인식률이 좋지 않게 나왔다. 일정 수준의 수용할 만한 레벨 이상이 되어야 도구가 이를 인식하기 때문에 외국인 학습자를 대상으로 할 경우 도구를 다르게 고안해야 한다.

음성 인식 정도와 오류 발견 등에 있어 몇 가지 한계가 있긴 하지만 ASR 기술은 외국인 화자의 발화를 인식하고 평가하는 시스템을 발전시키는 도구로 사용 가능하다.

Hamidi, F., Baljko, M.(2013)은 ASR을 이용하여 실험을 진행하고 그 한계와 극복 방안에 대해 논의하였다.

가. 개요

ASR이 아이들의 언어 교육 보조 시스템으로 사용되고 있지만 아직 해결해야 할 몇 가지 난제들이 있다. 그것은 바로 표준어와 일치하지 않는 발화의 입력과 잘 고안된 피드백이 필수라는 점이다.

나. 활용 도구

Dragon Naturally Speaking 프로그램을 이용하여 12살~16살 사이의 인도 중학생 16명을 대상으로 실험하였다. 일주일에 3시간씩 총 5개월간 4명의 관찰자가 이들을 관찰하였다.

다. 결론

ASR 모듈은 화자의 명확한 표준 발음을 기준으로 설계된다. 따라서 주변의 잡음이나 발화의 변형 등은 자동 인식에 어려움을 초래한다.

ASR이 직면한 문제는 ASR의 원래 목적(음성 인식)과 컴퓨터 교육 환경에서 추구하는 목표(발화의 분석과 평가)가 불일치한다는 데서 그 원인을 찾을 수 있다. 또한 ASR의 정확도를 향상시키는 것이 그 효율성을 높이는 효과적인 방법이 될 것이다.

Ono, Y., Ishihara, M., Onishi, A., Yamashiro, M.(2014)은 ASR의 음성 인식

성공률을 높일 수 있는 방안에 대해 이야기하였다.

가. 개요

외국어 교육에서 음성 인식 소프트웨어를 이용하면 학습자의 말하기 능력을 강화시킬 수 있다. 하지만 발화자 주변에서 발생하는 소음(noise)은 음성 인식에 어려움을 초래한다. 이러한 음성 인식을 확인을 위해 개인 PC와 헤드폰, 마이크가 장착된 CALL(Computer-Assisted Language Learning) 교실과 타블릿 PC나 모바일 기기를 사용하는 일반 교실, 두 교실로 상황을 나누어 실험을 진행하였다. 첫 번째 CALL 교실에서는 web application을 사용하고 두 번째 일반 교실에서는 local application을 사용하였다.

나. 활용 도구

일본 GLESA; 외국어 학습을 위한 LMS(Learning Management System) ver.2로 발음을 평가하였다. 평가는 ‘먼저 읽고 녹음, 연습, 모델 발음 듣기, 발음 테스트, 테스트 결과로 피드백 얻기, 다시 읽고 녹음, 돌아보기’의 순서로 진행되었다. 교실 인원은 평균 38.5명이었으며, 발음테스트(PRONTEST)를 통해 리딩시간, 강세, 고저, 유창성 등을 평가할 수 있었다.

다. 결론

헤드폰과 마이크를 사용한 CALL 교실이 상대적으로 소음이 적어 음성 인식 성공률이 높았다. 또한 web application으로 LMS에 직접 연결되므로 보다 좋은 결과를 얻을 수 있었다.

Dodigovic, M.(2012)은 ASR을 이용하여 실험을 진행하고 피험자들이 음성 인식 도구를 어떻게 평가하는지에 대해 논의하였다.

가. 개요

여러 사회, 문화적 집단의 시험 응시자들이 음성 인식 기술을 어떻게 생각하는지에 대해 알아보았다. 음성 인식 도구를 사용하여 사용자들의 외국어 능력을 평가하였는데 특정 나이대의 한국인 응시자들이 언어 인식 기술에 대해 가장 회의적이었다.

나. 활용 도구

The Versant test of English를 이용하여 말하기와 듣기를 포함한 6가지 과제를 테스트하였다. 테스트 내용은 문장 듣고 따라하기, 질문에 대답하기, 짧은 이야기를 듣고 다시 이야기하기 등으로 30분의 시간이 주어졌다.

다. 결론

Versant는 편안한 환경에서 테스트가 진행된다는 장점이 있지만 잘 숙지하여 다루기 어렵다는 단점도 있다.

Versant test of English에는 두 개의 그룹이 포함되었는데 한 그룹은 학생, 한 그룹은 언어 교사였고 두 그룹의 결과에는 확연히 차이가 있었다. 학생들은 Versant를 좀 더 손쉽게 받아들이는 반면 선생님들, 특히 25명의 한국에 거주하는 사람들은 새로운 기술을 받아들이지 못하는 것으로 나타났다. 연구에 포함된 교사들이 대부분 한국에 거주하는 까닭에 결과가 일반화되기는 어렵지만 교사들이 학생들에 비해 통계적으로 유의미한 차이를 보이는 것은 확실하다. 그러므로 언어 인식 기술이 언어 테스트에 있어서 아직 주류에 도달하지는 못한 것으로 보였다.

이상의 선행 연구를 통해 외국인 학습자를 대상으로 기존의 음성 인식 도구를 활용했을 때의 특성을 표로 정리하면 다음과 같다.

	내용
음성 인식률	• 모국어 화자의 발화를 기준으로 설계되었기 때문에 외국인 화자가 사용하였을 경우, 음성 인식률이 낮게 나타난다. • ASR 모듈 설계는 화자의 명확한 표준 발음을 기준으로 하므로 표준어와 일치하지 않는 발화의 경우 이를 입력할 수 없다.
인식 오류율	• 인식 오류율이 높게 나타난다. 특히 학습자의 모국어에 존재하지 않는 발음의 오류율이 높다.
피드백	• 오류가 발생했을 경우, 잘 고안된 피드백이 필수적으로 요구된다.
기타	• ASR 소프트웨어가 원어민 교수보다 학습자의 운율 오류를 발견할 확률이 높다. • 주변에 소음이 있는 경우, 음성 인식률이 낮게 나타난다.

<표 15> 외국인 학습자를 대상으로 ASR 소프트웨어를 활용했을 때의 특성

Neri, A., Cucchiarini, C., Strik, H.(2003)는 모국어 화자의 음성 인식 정확도가 90%를 기록했던 도구를 외국인 화자의 음성 인식에 사용하였으나 인식률이 좋지 않았음을 지적하였다. 그 이유는 일정 수준의 수용할 만한 레벨 이상이 되어야

도구가 이를 인식하기 때문인데 따라서 외국인 학습자를 대상으로 할 경우 음성 인식 도구를 다르게 고안해야 한다. 이러한 문제를 해결하기 위해서는 기존 모델에 학습자의 오류를 미리 예측하여 레이블링하는 작업이 요구된다.



<그림 133> 음성 인식 시스템의 구성 (서상규, 한영균, 2002)

그리고 Eskénazi, M.(1999)에 따르면, 학습자의 음운 오류의 경우, 전체 단어 중 특정 음운에서 오류가 많이 발견되었는데 L2의 음운이 L1의 그것과 유사한 경우 오류율이 적었고, 학습자의 모국어에 존재하지 않았던 발음은 오류율이 높았다. 따라서 한국어 학습자의 발음 오류가 어느 음운에서 빈번히 발생하는지 조사하여 이를 도구 설계에 반영한다면 음성 인식률을 높일 수 있는 좋은 방안이 될 것이다.

또한 In-Seok, K.(2006)에 의하면 도구를 사용했을 경우, 원어민 교수자가 학습자를 평가했을 때보다 학습자 오류를 발견한 확률이 더 높았다. 특히 운율의 경우 이러한 차이가 확연히 나타났는데 이는 다시 말해 소프트웨어에 저장되어 있는 원어민의 표준적인 발화 운율과 외국인 학습자의 운율에 큰 차이를 보인다는 것이다. 따라서 한국어 학습자를 위한 음성 인식 도구를 설계할 때에는 음운뿐 아니라 운율도 고려하여 작업해야 한다.

마지막으로 주변의 소음이 있을 경우 도구가 자동으로 음성을 인식하는 데 어려움을 초래하였다. 따라서 학습자가 조용한 환경에서 발화하고 이를 녹음한다면 음성 인식률을 높일 수 있을 것이다.

요컨대 외국인 학습자의 음성 인식 도구를 설계할 때에는 모국어 화자의 음성 인식 도구를 설계할 때와는 다르게 추가 작업이 이루어져야 한다. 추가 작업 시 고려해야 할 사항은 학습자의 발음 오류를 분석하여 예측하는 것이며 오류에는 음운 오류와 운율 오류가 두루 포함되어야 한다. 그리고 학습자의 발화를 녹음할 때에는 주변의 소음이 없는 조용한 환경에서 작업이 이루어져야 한다.

이러한 결론을 바탕으로 다음에서는 한국어 학습자의 음성 인식 도구를 설계할 때 고려해야 할 사항에 대해 알아보도록 하겠다. 즉, 한국어 학습자들에게서 특히 빈번히 발생하는 발음 오류를 조사하고 분석하겠다.

3. 세부 연구 내용

3.1. 한국어 학습자의 오류 특성

Hamidi, F., Baljko, M.(2013)에서 지적인 바와 같이, 한국어 학습자의 구어 말뭉치를 자동으로 전사하기 위해서는 한국어 학습자의 오류 특성을 먼저 파악하여 음성 인식 도구 설계 시 이를 반영하여야 한다. 기존의 음성 인식 도구는 모국어 화자의 발음을 표준 모델로 하므로 학습자가 기준 모델에 없는 발음 오류를 산출했을 경우 이를 인식하지 못한다. 따라서 학습자에게서 빈번히 발생하는 오류를 미리 예측하여 도구 설계 시 이를 반영해야 한다.

학습자의 발음 오류는 언어권별, 숙달도별 등의 변인에 의해 다르게 나타나므로 이를 고려하여 선행 연구를 통해 오류 유형을 분석해 보고자 한다. 한국어 학습자 말뭉치가 제안하는 언어권별 수집 비율은 다음 표와 같다.

분포	언어권	수집 비율
1위	중국어	±40%
2위	일본어	±10%
3위	베트남어	±10%
4위	스페인어	±10%
5위	영어	±10%
6위	기타	±10%

<표 16> 한국어 학습자 말뭉치의 언어권별 수집 비율

따라서 가장 비율이 높은 중국어권 학습자의 오류를 중심으로 일본어권, 베트남어권, 스페인어권, 영어권 화자의 오류를 함께 수집하는 방식으로 연구를 진행하겠다. 음운 오류뿐 아니라 운율 오류도 함께 알아볼 것이다.

○ 선행 연구

한국어 학습자들의 발음 오류를 관찰하기 위해 주로 사용된 실험은 읽기 목록을 만들어 그들의 발화를 녹음하고 분석하는 방법이었다. 몇 가지 예를 들면 다

음과 같다.

권성미(2007)는 초, 중, 고급 일본어권 학습자 각 16명을 대상으로 모음 음성 산출 실험을 하였다. SM48 마이크를 사용하여 녹음하고 CSL4500을 사용하여 음성 자료를 녹음, 분석하였다. 읽기 목록은 먼저 'k_'의 CV 구성의 1음절어를 만들고, '여기에 _가 있습니다'라는 읽기용 문장에 배치시켰는데, 예를 들면 '여기에 가 있습니다'와 같은 문장을 학습자에게 읽게 하였다.

이혜정(2015)은 수도권에 거주하는 베트남어권 초, 중급 학습자 총 58명을 대상으로 한국어능력시험을 채택하여 시험을 실시했다. 발음 실험 도구는 실험 자음과 음운현상 66개가 들어간 20개의 문장이었다. 발음 실험 문장에 포함된 단어는 자연스러운 발화 상황에서 발생하는 오류를 관찰하기 위해 개별적으로 제시하지 않고 문장 안에 분산 배치하였다. 실험 단어는 한국어능력시험의 초급, 중급 어휘 목록을 토대로 선정하였다. 녹음된 자료는 연구자와 한국어교육 전공생 총 두 명이 청취, 분석하였다.

김소현(2013)은 베트남어권 학습자 중 유학생 26명과 여성 결혼 이민자 24명, 총 50명의 초급 학습자를 대상으로 실험을 실시했다. 음향음성학적 분석 비교를 위해 한국인 표준어 구사 화자 45명의 발화도 녹음하였으며, 실험에 사용된 단어들은 무의미한 단어들로 선정하였다. 이렇게 만들어진 단어들을 동일한 음성 환경을 고려하여 '내가 ____다'라는 읽기용 문장에 넣어서 낭독하게 하였다. 산출 실험 자료의 음향음성학적 분석에는 Praat(version 5.3.41)를 이용하여 파형과 스펙트로그램을 동시에 띄어 놓고 분석했다.

권성미(2011)는 중국인의 억양 특성을 알아보기 위한 실험을 하였다. 중급과 고급 학습자 각 10명씩 총 20명의 학생들을 대상으로 대화 텍스트를 읽게 하였다. 문두 강세구 억양을 조사하기 위해 10개의 대화 텍스트를 준비하고, 경계 성조를 관찰하기 위해 20개의 대화 텍스트를 준비하였다. 750개의 수집 데이터를 Praat 버전 5.1.04를 이용하여 분석하였다.

하세가와 유키코(2005)는 일본인 학습자의 운율적 특징을 분석하기 위해 대학 부설 한국어 교육 기관에서 학습 기간이 9개월 이상인 학습자 32명을 대상으로 실험을 진행하였다. 녹취 내용은 첫 번째, 지명, 인명 등의 고유 명사와 한자어, 외래어를 포함한 일반명사 등 약 100개의 목표어의 단독발화와 문중발화, 두 번째, 8개 무의미어의 단독발화 및 문중발화, 세 번째, 10~50음절로 된 문장 읽기, 네 번째, 한국어에 의한 자유 대화, 다섯 번째, 일본어에 의한 단어 단독 발화와 문중 발화 및 자유대화였다. 한 명씩 두세 번의 발화를 시켰으며 녹음에는 Sony DAT Walkman TCD-D100과 Sony 마이크 ECM-ZS90을 사용하였다. 음성 처리에는 Sound it!을 사용하였다.

다음에서는 선행 연구에서 나타난 중국어권, 일본어권, 베트남어권, 스페인어

권, 영어권 학습자의 발음 오류 양상을 모음과 자음 및 운율로 나누어 정리하도록 하겠다.

3.1.1. 중국어권

1) 모음 ㅜ

선행 연구에서 나타난 중국어권 학습자의 모음 오류는 다음과 같다. 오류 예시는 범류(2010)를 인용하였다.

첫째로, /ㅏ/와 /ㅓ/의 혼동 오류이다. /ㅏ/를 /ㅓ/로 발음하거나 /ㅓ/를 /ㅏ/로 잘못 발음하는 오류가 적지 않게 발견되었다. 이러한 오류는 이중모음 /ㅑ/와 /ㅕ/의 혼동 오류로 이어진다.

(가) 거리 [고리]	어서 [오서]
오소 [어서]	가족 [가쩍]

둘째로, /ㅡ/와 /ㅣ/의 혼동 오류이다. 중국어권 학습자들은 /ㅡ/를 /ㅣ/로 발음하거나 /ㅡ/와 /ㅣ/를 구별하지 않고 [ㅡㅣ]처럼 모두 발음하는 오류를 나타냈다.

(나) 부르다 [부러다]	스승 [서승]
거리 [그어리]	저리 [즈어리]

2) 자음

선행 연구에서 나타난 중국어권 학습자의 자음 오류는 다음과 같다.

첫째로, 평음, 격음, 경음의 혼동 오류이다.

(가) 기 [끼]	곱빼기 [코빼기]	
덜다 [떨다]	당구 [탕구]	
뿌리 [부리]	팡 [방]	(위원장 · 홍미주 2013;170)

둘째로, 유음 ‘ㄹ’의 오류이다. /ㄹ/의 오류에는 ‘유음 성모의 간접 오류(ㄹ[r] → [l])’, ‘(M)VE\$(M)V(E) 연쇄의 간접 오류(ㄹ[r] → [ll] [lr])’, ‘r(권설작용)\$1 연쇄의 간접 오류(ㄹㄹ[ll] → [rl])’ 등이 있다.

(범류 2010;78)

셋째로, /ㅅ/과 /ㅈ/의 혼동 오류이다. 마찰음 /ㅅ/을 파찰음 /ㅈ/과 비슷하게 발음하는 오류인데 중국어 학습자는 실제 한국어의 조음 위치보다 약간 뒤쪽에서 이들 자음을 발음하기 때문에 나타나는 오류로 보인다.

(백소영 2008;81)

넷째로, 권설음의 간섭으로 인한 오류이다. 중국어에는 마찰음과 파찰음 계열에 권설음이 존재하여 한국어에서의 이들 발음도 권설음처럼 발음하는 것이다.

(범류 2010;51)

다섯째로, /ㄴ/과 /ㅇ/을 혼동하는 오류이다. 발음 중 /ㄴ/을 /ㅇ/으로 발음하기도 하고 반대로 /ㅇ/을 /ㄴ/으로 발음하기도 한다.

(위국봉 2014;152)

여섯째로, 성문 마찰음 / h /을 연구개 마찰음 / x /으로 발음하는 오류이다. 중국어의 자음 가운데 / h /에 바로 대응하는 자음은 없어 청취 인상이 비슷한 $\text{h}[\text{x}]$ 로 발음하는 것이다.

(범류 2010;51)

일곱째로, 발화 시 종성을 누락시키는 오류이다.

(양순임 2006;171)

3) 운물

중국어는 성조 언어이기 때문에 학습자들은 모국어의 영향으로 어조나 강세를 넣어 한국어를 발음하는 경향이 있다. 윤영숙(2012)에서 나타난 중국어권 학습자의 운율 오류는 다음과 같다.

첫째로, 초성 성조 오류이다. 한국어 강세구는 초성이 경음과 격음으로 시작되면 HHLH로, 그 외의 분절음으로 시작되면 LHLH로 실현된다. 하지만 이와 달리 초성 약자음에 H가 부과되어 나타나는 오류 유형이다.

둘째로, 피치 곡선의 형태 오류이다. 이는 강세구 내에서 각 음절에 부과되는 L과 H의 부조합에 의해 나타나는 오류이다. 특히 4음절 이상의 강세구에서 둘째 음절에 L이 실리거나 피치곡선의 굴절 횟수가 달라 오류가 나타난다. 한국어 강세구는 둘째 음절이 첫째 음절과 비슷하거나 그보다 높게 실현되어야 한다.

셋째로, 강세구 내부의 음높이 오류이다. 한국어 모국어 화자와 동일한 형태의 피치 패턴이 나타난다 하더라도 중국어권 학습자의 경우 음높이 차이가 크게 벌어져 청각적으로 어색한 강세구가 실현된 유형이다.

3.1.2. 일본어권

1) 모음

선행 연구에서 나타난 일본어권 학습자의 모음 오류는 다음과 같다. 오류 예시는 심보 토모코(2009)를 인용하였다.

첫째로, /ɪ/와 /ɯ/를 혼동하는 오류이다.

(가) 어머니 [오모니]	영당이 [온동이]
머리 [모리]	거짓말 [고짐말]
너구리 [노구리]	어차피 [오차피]

(나) 오히려 [어히려]	오징어 [어징어]
오후 [어후]	조금 [저금]
손님 [선님]	청소년 [천서년]

둘째로, /ɨ/와 /ɯ/를 혼동하는 오류이다.

(다) 가을 [가울]	모든[모든]
무슨 [무순]	요즘[요즘]
가끔 [가꿈]	그릇[구릇]

(라) 겨울 [겨을]	별꿀[별꿀]
물건 [물건]	수박[스박]
운명 [은명]	춤[츄]

이들 오류는 이중모음에서 /ㄱ/와 /ㄲ/, /ㄴ/와 /ㄷ/의 혼동으로 이어진다. 반면 /ㅣ/, /ㅐ/, /ㅓ/, /ㅕ/는 일본인들이 비교적 쉽게 습득하는 모음으로 밝혀졌다.

2) 자음

선행 연구에서 나타난 일본어권 학습자의 자음 오류는 다음과 같다. 오류 예시는 우인혜(1998)를 인용하였다.

첫째로, 평음, 경음, 격음을 혼동하는 오류이다. 한국어에는 다른 언어와 달리 평음, 경음, 격음의 3계열 구분이 확연할 뿐만 아니라 변별적 요소로까지 작용하기 때문에 외국인들이 이 음들을 구별해서 발음하거나 듣는 일은 여간 어렵지 않다.

(가) 운동 [운통]	동전 [통전]
대학 [태학]	잘 [찰]
나빠서 [나파서]	여러 가지 [여러까지]
친구 [진구]	치다 [찌다]
답승 [답승]	아닙니다 [아뽕니다]

둘째로, ‘ㄴ, ㄹ, ㅇ’ 받침을 혼동하는 오류이다. 한국어에서는 ‘ㄴ, ㄹ, ㅇ’이 음절 말에서도 음절초와 마찬가지로 독립된 음소로서 그 기능을 발휘하고 있는데 반해 일본어에서는 음절말에서 오직 일본문자 ‘ん’으로 대표되는 발음 [N]만이 실현되기 때문에 이러한 오류가 발생한다.(박기영 2001;234) 이들 자음 중 ‘ㄹ’은 오류가 적고 ‘ㄴ’과 ‘ㅇ’을 혼동하는 오류가 빈번히 발생한다.

(나) 청소 [천소]	공부 [곤부]
화장품 [화잔품]	항공 [한공]
통장 [톤장]	형제 [현제]

(다) 선생 [성생]	일본말 [일봉말]
-------------	-----------

찬송가 [창송가] 도서관에 [도서광에]
냉면 [냉명] 전화 [정화]

이 외에도 ‘서울’을 ‘서우루’로 발음하는 받침 음절 덧붙임 오류, 무성음 ‘ㄱ, ㄷ, ㅂ, ㅈ’를 발음할 때 단어 첫 글자에까지도 유성음을 사용하는 유성음 오류, [r]로 소리 내야 할 부분을 설측음 [l]로 강하게 발음하여 ‘르르’처럼 소리 내는 설측음 오류 등도 발견된다.

3) 운율

발화 역량은 중고급 레벨이 되어도 쉽게 고쳐지지 않는 오류로 손꼽힌다. 장부연(2002)에 따르면, 일본인 학습자에게는 한국어 단어가 외래어로 인식되어 일본어의 외래어 규칙을 적용한 결과, 2음절어는 /HL/, 3음절어는 /HLL/, 4음절어는 /LHLL/ 또는 /HLLL/과 같은 피치로 실현된다.

하지만 한국어 단어는 단독으로 발음될 때와 문장에서 발음될 때 그 음조가 다른데 이러한 차이점에 유의해서 발음해야 할 것이다. 하세가와 유키코(2005)에 따르면 한국어 단어는 단독으로 발음될 때는 단어가 짧으면 하강조로, 단어가 길면 상승-하강조로 발음되는 것이 일반적이다. 하지만 단어에 조사 등이 붙어 문장에서 발음될 때는 단어가 짧으면 상승조로, 단어가 길면 상승-하강-상승조로 발음되는 것을 기본으로 한다.

3.1.3. 베트남어권

베트남어와 한국어의 단모음 체계는 거의 유사하기 때문에(김지숙 2007:54) 모음에서는 오류가 많이 발생하지 않는다. 따라서 기존의 선행 연구는 자음만을 대상으로 연구한 논문이 주류를 이루었다.

1) 자음

선행 연구에서 나타난 베트남어권 학습자의 자음 오류는 다음과 같다. 오류 예시는 강정희(2011)를 인용하였다.

첫째로, 평음, 격음, 경음의 혼동 오류이다. 베트남어는 한국어와 달리 자음의 삼지적 대립이 없으므로 학습자는 이들 발음의 구분을 어려워한다. 격음 중 /궤/의 발음은 베트남어 자음인 /f/로 대체하여 발음하는 경우가 많다.

(가) 꼬마 [코마] 감기 [감끼, 감키, 캄끼]

떡국 [덕꾹]

도마 [토마]

둘째로, 파찰음 오류이다. 베트남어 자음에는 /ʦ/, /ʦʰ/, /ʦʰ/에 해당하는 자음이 존재하지 않는다. 학습자 중에는 이들 파찰음을 베트남어 경구개 폐쇄음인 /ch/와 치조 폐쇄음인 /tr/로 대체하여 발음하는 경우가 많다.

셋째로, 마찰음 오류이다. 한국어의 /s/, /ss/에 대응되는 베트남어의 마찰음은 /x/ 하나여서 학습자 중에는 이 둘을 구분하지 못하고 /x/로 발음하는 경우가 많다. /x/의 음가는 영어 /s/ 발음과 가까우며 한국어 자음 /s/, /ss/의 중간 세기로 발음된다.

넷째로, 종성에 위치하는 ‘ㄹ’의 오류이다. 학습자 중에는 종성에 위치하는 /ㄹ/을 /ㄴ/으로 발음하는 경우가 많은데 이는 베트남어의 음절 끝에는 /l/이 위치할 수 없다는 규칙에서 연유한다.

(나) 갈비 [간비]

굴에서 [군에서]

술 [쑤]

잘 팔아요 [잔 판아요]

3.1.4. 스페인어권

1) 모음

스페인어는 /a, e, i, o, u/의 다섯 개의 단모음만을 가지고 있기 때문에 학습자들은 한국어 모음을 발음하면서 적지 않은 오류를 보인다.

선행 연구에서 나타난 스페인어권 학습자의 모음 오류는 다음과 같다. 오류 예시는 서경석(2007)을 인용하였다.

첫째로, /ɪ/와 /ɯ/를 혼동하는 오류이다. 모음 /ɯ/는 비교적 정확하게 발음하지만 /ɪ/를 /ɯ/로 발음하는 경우가 많았다.

(가) 더위 [토위]

어머니 [어모니]

커피 [고빠]

무엇입니까 [무오심니까]

공부 [경부]

오빠 [어빠]

이러한 단모음 오류는 이중모음에서 /ɪ/와 /ɯ/의 혼동 오류로 이어진다.

둘째로, /ɨ-/의 오류이다. 학습자들은 /ɨ-/를 /ɪ/, /ɪ/, /ɪ/ 등의 모음으로 대체하여 발음하는 오류를 범했다.

(나) 스키 [수기] 기름 [기름]
 이름이 [이림이] 없습니다 [업칩니다]

2) 자음

선행 연구에서 나타난 스페인어권 학습자의 자음 오류는 다음과 같다. 오류 예시는 서경석(2007)과 박하나(2011)를 인용하였다.

첫째로, 평음, 격음, 경음의 혼동 오류이다.

(가) 가면 [카면] 끄다 [크다] 카드[까드]
 도시 [토시] 토래 [터래] 토요일 [도요일]
 바지 [파지] 포도 [버도] 커피 [고빼]

둘째로, 파찰음 오류이다. 스페인어에 없는 /ʦ/과 /ʦʰ/을 듣고 /ʦ/으로 발음하거나 /ʦʰ/, /ʦʰ/으로 발음하는 경우도 있다.

(나) 종이 [창이] 자꾸 [차꾸]
 준비 [순비] 집 [십]

셋째로, 받침 ‘ㅇ’이 ‘ㄴ’과 혼동되는 오류이다.

(다) 창문 [찬문] 강물 [간물]
 한파 [향파] 부산 [부상]

넷째로, 후음 ‘ㅎ’을 발음하지 않거나 연구개음 /x/로 발음하는 오류이다.

(라) 호랑이 [오라니] 현재 [연재]
 오해 [오해(oxe)] 해매다 [해매다(xemeda)]

다섯째로, 자음을 삽입하는 오류이다. ‘∅-y계 이중모음’ 연속체와 ‘ŋ-∅’ 연속체에 자음이 삽입된다. 전자에는 ‘ʦ’삽입이, 후자는 ‘ㄱ’ 삽입이 두드러지게 관찰되었다.

(마) 여자[저자] 여러[저러]
 영어 [영거] 장애인 [장개인]

3.1.5. 영어권

1) 모음

선행 연구에서 나타난 영어권 학습자의 모음 오류는 다음과 같다. 오류 예시는 전나영(1993)을 인용하였다.

첫째로, /ɪ/와 /ɪ/의 혼동 오류이다. 영어에는 한국어 모음과 같은 /ɪ/ 소리가 없기 때문에 이러한 오류가 발생하는 것으로 보인다.

둘째로, /-/의 오류이다. /-/는 영어에 없는 발음이므로 학습자들이 오류를 빈번히 생성하게 된다. /-/를 /ɪ/로 발음하거나 반대로 /ɪ/를 /-/로 발음하는 오류 양상을 보인다.

(가) 크다 [커다]	있으니까 [이씨니까]
너무 [느무]	거리 [그리]

2) 자음

선행 연구에서 나타난 영어 학습자의 자음 오류는 다음과 같다. 오류 예시는 김정윤(2007)을 인용하였다.

첫째로, 평음, 격음, 경음의 혼동 오류이다.

(가) 조카 [조까]	잠꾸러기 [잠구러기]
대장 [태장]	똥똥하고 [똥통하고]
보리빵 [포리방]	판도 [반도]

둘째로, /ㅅ/과 /ㅆ/을 혼동하는 마찰음 오류이다.

(나) 새우등 [쌔우등]	썰썰하고 [스슬하고]
---------------	-------------

셋째로, 유음인 ‘ㄹ’ 오류이다. 모음 사이의 /ㄹ/을 혀를 말아 발음하는 권설음 /r/로 발음하거나 /ㄹ/이 두 번 반복될 때 한 번만 발음하는 오류가 나타난다.

(다) 몰라요 [모라요]	별로 [벼로]
---------------	---------

이상에서 살펴본 내용을 정리하자면, 한국어 학습자들은 자신의 모국어에 한국어 해당 발음이 없는 경우 오류를 빈번히 생성하였다. 이는 Eskénazi, M.(1999)에서 대부분의 학습자가 모국어에 존재하지 않는 발음에서 높은 오류율을 보인

다고 서술한 내용과 일치한다. 그리고 공통적으로 모음에서는 /ㅏ/, /ㅓ/, /ㅡ/의 발음을 어려워하였고, 자음에서는 평음, 격음, 경음의 분류, 마찰음, 파찰음, 유음의 발음에서 많은 어려움을 나타냈다.

3.2. 구어 전사 도구 개발을 위한 실험

우리는 앞 절에서 각 나라별 한국어 학습자의 오류 패턴을 조사하였다. 이러한 오류 패턴을 고려하여 음성 인식기를 개발한다면 외국인 학습자의 음성 인식률을 높일 수 있을 것으로 기대한다. 다음에서는 이러한 가정을 뒷받침하기 위한 실험으로, 외국인 학습자를 대상으로 현행 음성 전사 도구를 활용하였을 때 그 인식률이 어느 정도로 나타나는지 알아보고자 한다. 모국어 화자에 비해 외국인 학습자의 음성 인식률이 현저히 낮게 나타난다면 도구의 설계가 다르게 이루어져야 함을 시사하는 것이다.

3.2.1. 실험 목적

본 실험의 목적은 크게 두 가지이다.

첫째는 모국어 화자와 외국인 화자의 음성 인식률을 비교하는 것이다. In-seok, K.(2006)는 ASR 소프트웨어가 원어민의 발음을 정확하게 판단하는지 확인하기 위해서는 모국어 화자를 대상으로 하는 대조군 실험이 필요함을 지적한 바 있다. 따라서 모국어 화자도 실험 대상에 포함시켜 음성 인식률을 측정하고 외국인 화자와의 인식률 차이를 분석하고자 한다.

둘째는 외국인 화자의 음성 인식 양상이 어떻게 나타나는지 알아보는 것이다. 인식이 불가능한 경우, 다른 어휘로 인식되는 경우, 해당 어휘로 바르게 인식되는 경우로 나누어 실험 결과를 살펴볼 것이다. 이 중 우리가 주의 깊게 관찰할 항목은 인식이 불가능한 경우인데, Neri, A., Cucchiarini, C., Strik, H.(2003)는 모국어 화자의 음성 인식 정확도가 90%를 기록했던 도구를 외국인 화자의 음성 인식에 사용했더니 인식률이 낮게 나왔다는 사실을 지적하며 외국인 학습자를 대상으로 할 경우 도구를 다르게 고안해야 한다고 제안하였다. 일정 수준의 수용할 만한 레벨 이상이 되어야 도구가 이를 인식하기 때문이다. 따라서 본 연구에서는 다음의 실험 결과를 바탕으로 외국인 학습자를 대상으로 하는 음성 인식기 설계에 반영해야 하는 항목을 알아보고자 한다.

3.2.2. 실험 방법

(1) 실험 대상

본 실험의 피험자는 홍익대학교 국제언어교육원에서 한국어 수업을 받고 있는 중국어권 학습자 37명과 일본어권 학습자 39명이다. 초·중·고급의 학습자가 골고루 실험에 참여하였으며, 대조군으로 한국어 강사 18명도 실험에 참여하여 읽기 데이터를 녹음하였다.

유형 급수	iOS			Android		
	일본어권 학습자	중국어권 학습자	한국인	일본어권 학습자	중국어권 학습자	한국인
2급	12	8		10	10	
4급	8	9		3	4	
5급	.	.		4	4	
6급	.	.		2	2	
계	20	17	8	19	20	10

<표 17> 실험 대상

(2) 음성 자료

음성 자료는 국립국어원의 2012년 한국어 교육 어휘 내용 개발 1단계(연구 책임자: 강현화)에서 채택된 초급 어휘 가운데 한국어의 주요 음운 현상과 관련된 대표 어휘 30개로 선정하였다.

관련 음운현상	단어	관련 음운현상	단어	관련 음운현상	단어
연음	목요일	비음(동)화	정류장	경음화	갑자기
	울음		음료수		결석
	음악		작년		식당
	십이월		십만		젓가락
평폐쇄음화	부엌	유음화	연락	유기음화	그렇지만
	맛		편리		빨갳다
	옆		설날		막히다
구개음화	똑같이	자음군 단순화	값	ㅎ 탈락	문화
	불이다		답		많이
ㄴ 삽입	지하철역		앉다		끓이다

<표 18> 음성 자료

(3) 실험 도구와 실험 절차

실험에 사용한 음성 인식기는 ‘iOS Siri’와 ‘Android Google NOW’ 두 가지로, 두 인식기의 음성 인식률을 비교하였다. 두 가지 도구 모두 일반인을 대상으로 서비스를 제공하며 보편적으로 사용이 가능하다.

각 피험자는 해당 음성 인식기가 설치된 휴대전화를 사용하여 실험지의 어휘 목록을 발화하였다. 밀폐된 공간에서 한국어 교사가 본인과 학습자를 대상으로 실험을 진행하였으며, 교사는 실험지에 음성 인식 여부와 인식 결과를 기록하였다.

음성 인식 결과는 ‘인식 불가’, ‘오인식’, ‘적정 인식’의 세 가지로 나뉜다. ‘인식 불가’는 음성 인식기가 아예 해당 단어를 인식하지 못한 경우이고, ‘오인식’은 실험지에 제시한 단어와 다른 단어로 인식한 경우이다. 그리고 ‘적정 인식’은 제시된 단어와 같은 단어로 바르게 인식한 경우이다. 피험자는 실험에서 같은 어휘를 3회씩 반복 발화하였으며 평균값으로 인식 여부를 판단하였다.⁵⁾

3.2.3. 실험 결과

(1) 실험 분석

실험 결과는 실험에 쓰인 음성 인식 도구의 종류(iOS Siri, Android Google NOW)에 따라 언어권별(중국어권 학습자, 일본어권 학습자, 모국어 화자), 급수별로 분석하였다.

5) 본 실험에서는 학습자의 운율은 분석하지 못하였다. 운율 오류에 대한 분석은 다음을 기약하기로 한다.

첫 번째로, 'iOS Siri'로 실시한 중국어권 학습자의 음성 인식 테스트 결과는 다음과 같다.

(단위: %)

관련 음운현상	단어	2급			4급		
		인식 불가	오인	적정 인식	인식 불가	오인	적정 인식
연음	목요일	25.0	4.2	70.8	7.4	3.7	88.9
	울음	33.3	58.3	8.3	59.3	40.7	0.0
	음악	4.2	12.5	83.3	7.4	18.5	74.1
	십이월	4.2	20.8	75.0	0.0	18.5	81.5
평폐쇄음 화	부엌	25.0	75.0	0.0	37.0	59.3	3.7
	맛	33.3	62.5	4.2	30.8	69.2	0.0
	옆	25.0	75.0	0.0	3.7	96.3	0.0
비음(동) 화	정류장	8.3	25.0	66.7	7.4	22.2	70.4
	음료수	29.2	4.2	66.7	18.5	22.2	59.3
	작년	0.0	66.7	33.3	0.0	66.7	33.3
	십만	12.5	75.0	12.5	0.0	81.5	18.5
유음화	연락	12.5	45.8	41.7	14.8	29.6	55.6
	편리	4.2	83.3	12.5	3.7	70.4	25.9
	설날	20.8	54.2	25.0	7.4	70.4	22.2
경음화	갑자기	0.0	20.8	79.2	0.0	25.9	74.1
	결석	12.5	58.3	29.2	0.0	37.0	63.0
	식당	4.2	16.7	79.2	3.7	25.9	70.4
	젓가락	25.0	50.0	25.0	7.4	55.6	37.0
유기음화	그렇지만	4.2	66.7	29.2	3.7	74.1	22.2
	빨갈다	16.7	83.3	0.0	3.7	92.6	3.7
	막히다	8.3	54.2	37.5	3.7	44.4	51.9
구개음화	똑같이	16.7	16.7	66.7	3.7	33.3	63.0
	불이다	33.3	58.3	8.3	14.8	77.8	7.4
자음군 단순화	값	16.7	83.3	0.0	14.8	85.2	0.0
	닭	0.0	100.0	0.0	25.9	74.1	0.0
	앉다	20.8	79.2	0.0	18.5	81.5	0.0
ㄷ 탈락	문화	25.0	37.5	37.5	25.9	18.5	55.6
	많이	0.0	25.0	75.0	0.0	22.2	77.8
	끓이다	45.8	54.2	0.0	29.6	70.4	0.0
ㄴ 삼입	지하철역	0.0	12.5	87.5	3.7	3.7	92.6

<표 19> 중국어권 학습자 iOS 음성 인식 테스트 결과

두 번째로, ‘iOS Siri’로 실시한 일본어권 학습자의 음성 인식 테스트 결과는 다음과 같다.

(단위: %)

관련 음운현상	단어	2급			4급		
		인식 불가	오인	적정 인식	인식 불가	오인	적정 인식
연음	목요일	16.7	16.7	66.7	20.8	0.0	79.2
	울음	55.6	36.1	8.3	37.5	37.5	25.0
	음악	2.8	8.3	88.9	0.0	0.0	100.0
	십이월	5.6	16.7	77.8	0.0	37.5	62.5
평폐쇄음 화	부엌	50.0	50.0	0.0	45.8	54.2	0.0
	맛	33.3	55.6	11.1	25.0	70.8	4.2
	옆	8.3	88.9	2.8	4.2	95.8	0.0
비음(동) 화	정류장	8.3	11.1	80.6	0.0	4.2	95.8
	음료수	33.3	33.3	33.3	25.0	12.5	62.5
	작년	2.8	80.6	16.7	4.2	62.5	33.3
	십만	8.3	88.9	2.8	0.0	70.8	29.2
유음화	연락	8.3	19.4	72.2	16.7	12.5	70.8
	편리	19.4	72.2	8.3	12.5	62.5	25.0
	설날	8.3	41.7	50.0	8.3	12.5	79.2
경음화	갑자기	2.8	13.9	83.3	0.0	4.2	95.8
	결석	5.6	13.9	80.6	0.0	37.5	62.5
	식당	2.8	2.8	94.4	0.0	4.2	95.8
	젓가락	2.8	33.3	63.9	4.2	25.0	70.8
유기음화	그렇지만	0.0	22.2	77.8	0.0	8.3	91.7
	빨갈다	8.3	88.9	2.8	12.5	87.5	0.0
	막히다	5.6	66.7	27.8	4.2	45.8	50.0
구개음화	똑같이	2.8	86.1	11.1	4.2	70.8	25.0
	붙이다	33.3	58.3	8.3	20.8	75.0	4.2
자음군 단순화	값	22.2	77.8	0.0	29.2	70.8	0.0
	닭	30.6	69.4	0.0	16.7	83.3	0.0
	앉다	22.2	77.8	0.0	8.3	91.7	0.0
ㅎ 탈락	문화	52.8	44.4	2.8	20.8	41.7	37.5
	많이	2.8	5.6	91.7	0.0	12.5	87.5
	끓이다	5.6	88.9	5.6	4.2	95.8	0.0
ㄴ 삽입	지하철역	8.3	13.9	77.8	0.0	12.5	87.5

<표 20> 일본어권 학습자 iOS 음성 인식 테스트 결과

다음으로, ‘Android Google NOW’로 실시한 중국어권 학습자의 음성 인식 테스트 결과와 일본어권 학습자의 음성 인식 테스트 결과는 각각 <표21>, <표22>와 같다.

(단위: %)

관련 음운현상	단어	2급			4급			5급			6급		
		인식불가	오인	적정인식	인식불가	오인	적정인식	인식불가	오인	적정인식	인식불가	오인	적정인식
연음	목요일	16.7	3.3	80.0	0.0	16.7	83.3	0.0	0.0	100.0	0.0	0.0	100.0
	울음	10.0	63.3	26.7	33.3	50.0	16.7	0.0	16.7	83.3	0.0	83.3	16.7
	음악	0.0	16.7	83.3	8.3	75.0	16.7	0.0	16.7	83.3	0.0	50.0	50.0
	십이월	0.0	6.7	93.3	0.0	16.7	83.3	0.0	8.3	91.7	0.0	16.7	83.3
평폐쇄음화	부엌	3.3	40.0	56.7	8.3	41.7	50.0	0.0	75.0	25.0	0.0	50.0	50.0
	맛	20.0	56.7	23.3	41.7	58.3	0.0	0.0	66.7	33.3	0.0	66.7	33.3
	엿	16.7	63.3	20.0	25.0	66.7	8.3	0.0	50.0	50.0	0.0	50.0	50.0
비음(동)화	정류장	6.7	33.3	60.0	0.0	33.3	66.7	0.0	0.0	100.0	0.0	66.7	33.3
	음료수	6.7	16.7	76.7	0.0	0.0	100.0	0.0	0.0	100.0	0.0	16.7	83.3
	작년	0.0	86.7	13.3	0.0	100.0	0.0	0.0	58.3	41.7	0.0	50.0	50.0
	십만	3.3	36.7	60.0	8.3	58.3	33.3	0.0	33.3	66.7	0.0	33.3	66.7
유음화	연락	3.3	26.7	70.0	8.3	50.0	41.7	0.0	25.0	75.0	0.0	50.0	50.0
	편리	3.3	80.0	16.7	8.3	58.3	33.3	0.0	58.3	41.7	0.0	33.3	66.7
	설날	3.3	63.3	33.3	8.3	83.3	8.3	0.0	41.7	58.3	0.0	100.0	0.0
경음화	갑자기	10.0	36.7	53.3	8.3	16.7	75.0	0.0	0.0	100.0	0.0	0.0	100.0
	결석	6.7	36.7	56.7	0.0	50.0	50.0	0.0	0.0	100.0	0.0	50.0	50.0
	식당	3.3	26.7	70.0	0.0	25.0	75.0	0.0	0.0	100.0	0.0	0.0	100.0
	젓가락	6.7	33.3	60.0	0.0	8.3	91.7	0.0	8.3	91.7	0.0	50.0	50.0
유기음화	그렇지만	0.0	53.3	46.7	8.3	41.7	50.0	0.0	33.3	66.7	0.0	33.3	66.7
	빨갳다	3.3	66.7	30.0	0.0	100.0	0.0	0.0	8.3	91.7	0.0	50.0	50.0
	막히다	0.0	40.0	60.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	16.7	83.3
구개음화	똑같이	3.3	20.0	76.7	0.0	8.3	91.7	0.0	8.3	91.7	0.0	0.0	100.0
	불이다	16.7	16.7	66.7	8.3	50.0	41.7	0.0	25.0	75.0	0.0	66.7	33.3
자음군 단순화	값	0.0	96.7	3.3	25.0	75.0	0.0	0.0	100.0	0.0	0.0	66.7	33.3
	닭	6.7	83.3	10.0	33.3	58.3	8.3	0.0	50.0	50.0	0.0	50.0	50.0
	앉다	0.0	80.0	20.0	8.3	91.7	0.0	0.0	100.0	0.0	0.0	100.0	0.0
ㄷ 탈락	문화	0.0	36.7	63.3	33.3	25.0	41.7	0.0	33.3	66.7	0.0	16.7	83.3
	많이	0.0	6.7	93.3	0.0	16.7	83.3	0.0	0.0	100.0	0.0	0.0	100.0
	끓이다	0.0	86.7	13.3	8.3	75.0	16.7	0.0	66.7	33.3	0.0	16.7	83.3
ㄴ 삽입	지하철역	3.3	3.3	93.3	0.0	16.7	83.3	0.0	0.0	100.0	0.0	0.0	100.0

<표 21> 중국어권 학습자 Android 음성 인식 테스트 결과

(단위: %)

관련 음문현상	단어	2급			4급			5급			6급		
		인식불가	오인	적정인식	인식불가	오인	적정인식	인식불가	오인	적정인식	인식불가	오인	적정인식
연음	목요일	20.0	3.3	76.7	0.0	22.2	77.8	0.0	0.0	100.0	0.0	0.0	100.0
	물음	6.7	93.3	0.0	33.3	66.7	0.0	0.0	91.7	8.3	0.0	100.0	0.0
	음악	0.0	6.7	93.3	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0	100.0
	십이월	0.0	0.0	100.0	0.0	0.0	100.0	0.0	8.3	91.7	0.0	0.0	100.0
평폐쇄음화	부엌	30.0	53.3	16.7	55.6	44.4	0.0	0.0	66.7	33.3	0.0	50.0	50.0
	맛	13.3	60.0	26.7	22.2	77.8	0.0	0.0	75.0	25.0	0.0	0.0	100.0
	옆	16.7	60.0	23.3	11.1	33.3	55.6	0.0	66.7	33.3	0.0	0.0	100.0
비음(동)화	정류장	0.0	0.0	100.0	0.0	0.0	100.0	0.0	16.7	83.3	0.0	0.0	100.0
	음료수	0.0	40.0	60.0	0.0	100.0	0.0	0.0	8.3	91.7	0.0	0.0	100.0
	작년	0.0	90.0	10.0	0.0	100.0	0.0	0.0	91.7	8.3	0.0	16.7	83.3
	심만	3.3	63.3	33.3	0.0	66.7	33.3	0.0	41.7	58.3	0.0	16.7	83.3
유음화	연락	6.7	10.0	83.3	0.0	22.2	77.8	0.0	0.0	100.0	0.0	0.0	100.0
	편리	10.0	86.7	3.3	0.0	100.0	0.0	0.0	83.3	16.7	0.0	50.0	50.0
	설날	6.7	60.0	33.3	22.2	44.4	33.3	0.0	16.7	83.3	0.0	0.0	100.0
경음화	갑자기	6.7	6.7	86.7	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0	100.0
	결석	13.3	23.3	63.3	0.0	11.1	88.9	0.0	25.0	75.0	0.0	0.0	100.0
	식당	0.0	3.3	96.7	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0	100.0
	젓가락	6.7	3.3	90.0	0.0	0.0	100.0	0.0	25.0	75.0	0.0	0.0	100.0
유기음화	그렇지만	0.0	30.0	70.0	0.0	33.3	66.7	0.0	0.0	100.0	0.0	16.7	83.3
	빨갈다	0.0	100.0	0.0	0.0	100.0	0.0	0.0	91.7	8.3	0.0	33.3	66.7
	막히다	0.0	73.3	26.7	11.1	88.9	0.0	0.0	83.3	16.7	0.0	16.7	83.3
구개음화	똑같이	0.0	43.3	56.7	0.0	33.3	66.7	0.0	8.3	91.7	0.0	0.0	100.0
	붙이다	0.0	30.0	70.0	0.0	66.7	33.3	0.0	25.0	75.0	0.0	0.0	100.0
자음군 단순화	값	20.0	80.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0	83.3	16.7
	닭	6.7	86.7	6.7	33.3	66.7	0.0	0.0	58.3	41.7	0.0	66.7	33.3
	앞다	0.0	50.0	50.0	0.0	77.8	22.2	0.0	83.3	16.7	0.0	100.0	0.0
ㅎ 탈락	문화	13.3	76.7	10.0	11.1	77.8	11.1	0.0	50.0	50.0	0.0	16.7	83.3
	많이	0.0	6.7	93.3	0.0	11.1	88.9	0.0	0.0	100.0	0.0	0.0	100.0
	끓이다	6.7	93.3	0.0	0.0	100.0	0.0	0.0	83.3	16.7	0.0	16.7	83.3
ㄴ 삽입	지하철역	0.0	6.7	93.3	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0	100.0

<표 22> 일본어권 학습자 Android 음성 인식 테스트 결과

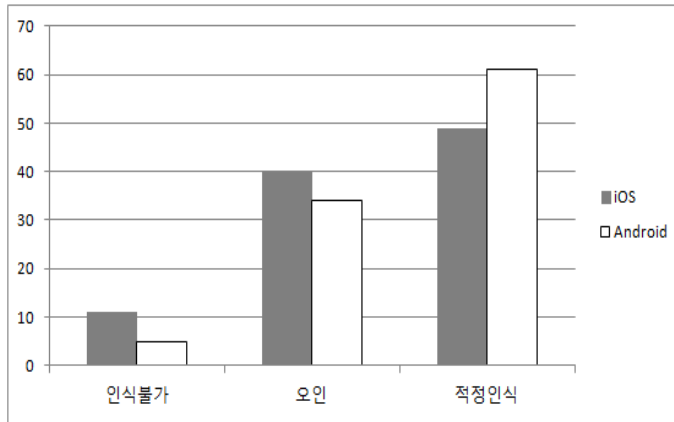
마지막으로, 모국어 화자의 음성 인식 테스트 결과는 다음과 같다.

(단위: %)

관련 음운현상	단어	iOS			Android		
		인식 불가	오인	적정 인식	인식 불가	오인	적정 인식
연음	목요일	0.0	0.0	100.0	0.0	0.0	100.0
	울음	4.2	33.3	62.5	0.0	10.0	90.0
	음악	0.0	0.0	100.0	0.0	0.0	100.0
	십이월	0.0	0.0	100.0	0.0	0.0	100.0
평폐쇄음 화	부엌	12.5	87.5	0.0	0.0	3.3	96.7
	맛	0.0	54.2	45.8	0.0	0.0	100.0
	옆	0.0	100.0	0.0	0.0	3.3	96.7
비음(동) 화	정류장	0.0	0.0	100.0	0.0	0.0	100.0
	음료수	0.0	4.2	95.8	0.0	13.3	86.7
	작년	0.0	16.7	83.3	0.0	13.3	86.7
	심만	0.0	100.0	0.0	0.0	3.3	96.7
유음화	연락	0.0	0.0	100.0	0.0	0.0	100.0
	편리	0.0	29.2	70.8	0.0	20.0	80.0
	설날	0.0	0.0	100.0	0.0	0.0	100.0
경음화	갑자기	0.0	0.0	100.0	0.0	0.0	100.0
	결석	0.0	4.2	95.8	0.0	3.3	96.7
	식당	0.0	4.2	95.8	0.0	3.3	96.7
	젓가락	0.0	4.2	95.8	0.0	0.0	100.0
유기음화	그렇지만	0.0	0.0	100.0	0.0	0.0	100.0
	빨갈다	0.0	100.0	0.0	0.0	3.3	96.7
	막히다	0.0	0.0	100.0	0.0	3.3	96.7
구개음화	똑같이	0.0	16.7	83.3	0.0	0.0	100.0
	붙이다	0.0	91.7	8.3	0.0	0.0	100.0
자음군 단순화	값	0.0	100.0	0.0	0.0	100.0	0.0
	닭	4.2	91.7	4.2	0.0	0.0	100.0
	앉다	0.0	100.0	0.0	0.0	23.3	76.7
ㅎ 탈락	문화	0.0	4.2	95.8	0.0	6.7	93.3
	많이	0.0	0.0	100.0	0.0	0.0	100.0
	끓이다	0.0	100.0	0.0	0.0	26.7	73.3
ㄴ 삽입	지하철역	0.0	0.0	100.0	0.0	0.0	100.0

<표 23> 모국어 화자의 음성 인식 테스트 결과

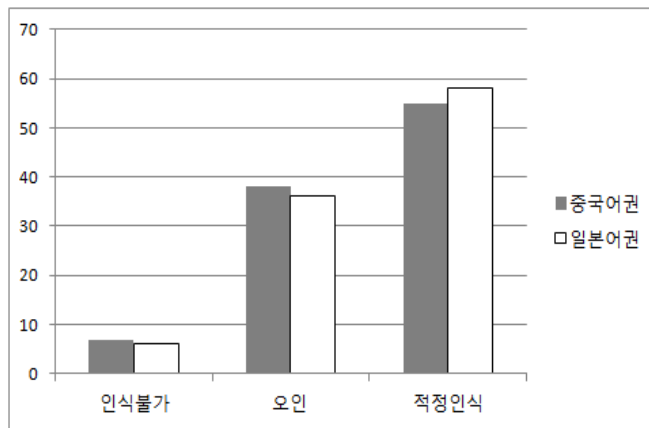
실험 결과, iOS보다 Android의 음성 인식률이 대체적으로 높게 나타났다. iOS를 사용한 피험자는 모국어 화자 외에 2급과 4급 학습자에 한정되었으므로, Android를 사용한 2급과 4급 학습자의 실험 결과만을 비교 대상으로 하여 도구의 인식률을 비교하였다. iOS에서는 인식불가가 11%, 오인식이 40%, 적정인식이 49%였으며, Android에서는 인식불가가 5%, 오인식이 34%, 적정인식이 61%였다.⁶⁾



<그림 134> iOS와 Android의 음성 인식률 비교

특히 모국어 화자의 음성 인식률에서 큰 차이를 보였는데 iOS를 사용하여 실험한 경우, 인식불가가 1%, 오인식이 28%, 적정인식이 71%인데 반해, Android를 사용하여 실험한 경우, 인식불가가 0%, 오인식이 4%, 적정인식이 96%였다.

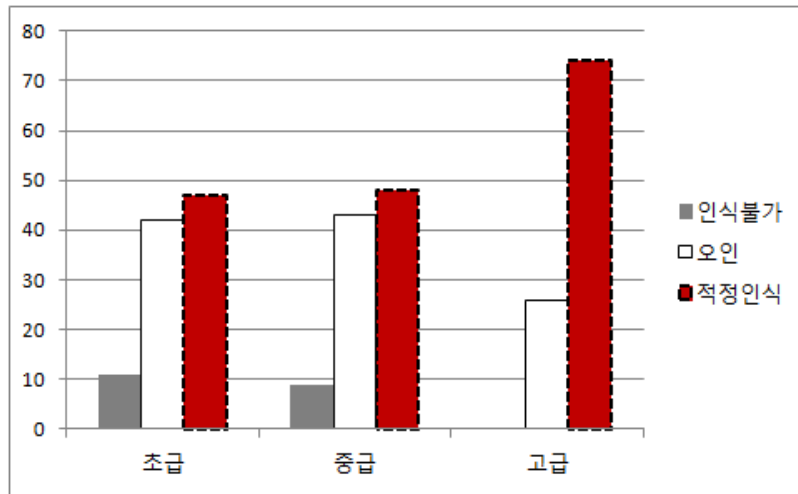
다음으로 학습자의 언어권별 음성 인식률을 살펴보면 중국어권 학습자의 경우, 인식불가가 7%, 오인식이 38%, 적정인식이 55%였고, 일본어권 학습자의 경우, 인식불가가 6%, 오인식이 36%, 적정인식이 58%를 차지했다. 따라서 언어권에 따른 음성 인식률에는 유의미한 차이가 나타나지 않았다.



<그림 135> 언어권별 음성 인식률 비교

- 6) 실험 결과를 분석할 때, 실험지의 문항 중 ‘자음군 단순화’는 제외하였다. 자음군 단순화의 경우 뒤에 모음이 이어지지 않으면 어떤 단어인지 알기 어렵기 때문이다. 예를 들어, ‘값’을 [갑]으로 발음하였을 때 해당 단어가 ‘값’인지 ‘갑’인지 여부를 판단하는 데에는 한계가 있다.

다음으로 학습자의 수준별 음성 인식률을 살펴보면 초급(2급)의 경우, 인식불가가 11%, 오인식이 42%, 적정인식이 47%였고, 중급(4급)의 경우, 인식불가가 9%, 오인식이 43%, 적정인식이 48%였으며, 고급(5급·6급)의 경우, 인식불가가 0%, 오인식이 26%, 적정인식이 74%였다. 이를 통해 학습자의 수준이 높아질수록 음성 인식률 또한 높아진다는 사실을 알 수 있는데, 특히 고급 레벨의 학습자의 경우 초·중급 학습자보다 높은 인식률을 보였다.

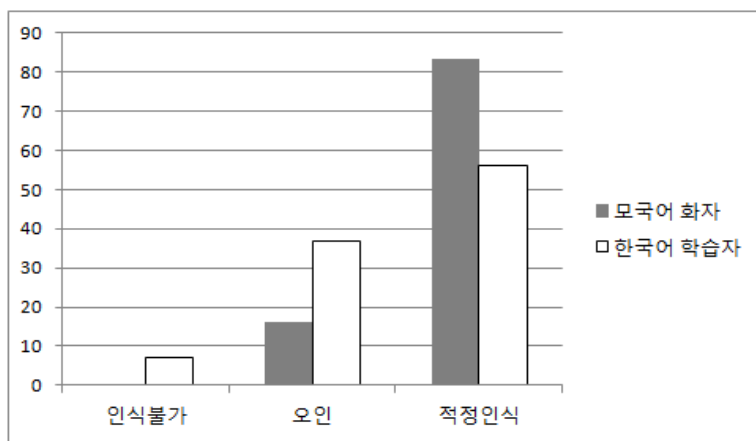


<그림 136> 학습자 수준별 음성 인식률 비교

다음으로 모국어 화자와 한국어 학습자의 음성 인식률을 살펴보겠다. 모국어 화자의 경우, 인식불가가 0.5%, 오인식이 16%, 적정인식이 83.5%였으며, 한국어 학습자의 경우 인식불가가 7%, 오인식이 37%, 적정인식이 56%였다. 한국어 학습자의 적정 인식률이 모국어 화자의 적정 인식률보다 약 28% 가량 낮음을 알 수 있다.

(2) 결과 해석

실험 결과를 정리하면, 1) iOS보다 Android를 사용하였을 때 음성 인식률이 높았고, 2) 학습 레벨이 높아질수록 음성 인식률이 증가했으며, 3) 언어권별로는 큰 차이가 나타나지 않았다. 그리고 4) 선행 연구에서 살펴본 바와 같이 외국인 학습자의 음성 인식률이 모국어 화자의 음성 인식률보다 현저히 낮게 나타났다. 이는 곧, 외국인 학습자를 대상으로 음성 인식 도구를 활용하고자 할 경우, 모국어 화자를 대상으로 하는 도구와 설계를 달리 해야 함을 시사한다.



<그림 137> 모국어 화자와 한국어 학습자의 음성 인식을 비교

이러한 인식률의 한계를 극복하기 위한 방안으로, 인식 불가율이 높게 나타난 단어를 찾아 그 원인을 분석하고 도구의 설계에 이를 반영하는 방법이 있다.

이번 실험에서 음성 인식 불가 비율이 20% 이상 나타난 단어들을 정리하면 다음 표와 같다.⁷⁾

	중국어권 학습자		일본어권 학습자	
	iOS	Android	iOS	Android
1	울음 (46%)	맛 (31%)	부엌 (48%)	부엌 (43%)
2	끓이다 (38%)	울음 (22%)	울음 (47%)	울음 (20%)
3	맛 (32%)	옆 (21%)	문화 (37%)	
4	문화 (26%)		맛 (29%)	
5	끓이다 (24%)		음료수 (29%)	
6	음료수 (23%)		끓이다 (27%)	

<표 24> 음성 인식 불가 단어 목록

이들 단어의 인식 불가 원인은 여러 가지로 상정해 볼 수 있겠지만 모국어 화자의 경우 인식 불가율이 0.5%에 그친 것과 비교해 볼 때, 학습자가 한국어에 없는 음운을 발화하여 음성 인식기가 이를 인식하지 못했을 확률이 높다고 가정할 수 있다. 그리고 이렇게 발화된 음운은 3.1.에서 살펴 본 한국어 학습자의 오류

7) 고급 학습자의 발화는 대부분 음성 인식에 성공하였으므로 고급을 제외한 초·중급 학습자의 실험 결과를 대상으로 하였다.

양상과 관련지어 생각해 볼 수 있다.

예를 들어 일본어권 학습자들이 ‘울음’을 발화했을 때 음성 인식기가 이를 인식하지 못한 것은 일본어권 학습자들의 오류 특성 중 /ㅡ/와 /ㅌ/를 혼동하여 발음하는 것과 관련된다. 일본어의 ‘う’ 발음은 한국어 /ㅡ/와 /ㅌ/의 중간 음으로 발음된다. 따라서 일본어권 학습자들은 ‘ㅌ’를 ‘う’처럼 발화하고, 음성 인식기는 /ㅡ/와 /ㅌ/의 중간음을 한국어로 인식하지 못한 것이다. ‘울음’의 오인식 목록 중 /으름/이 가장 많았다는 사실은 이러한 가정을 뒷받침해 준다.

또 다른 예로 중국어권 학습자들이 ‘끓이다’를 발화하였을 때 음성 인식기의 오인식 결과 중 ‘그리다’가 가장 많이 나타났다는 것은 중국어권 학습자가 한국어의 평음, 격음, 경음을 혼동하는 것과 관련지어 생각할 수 있다. 음성 인식기가 ‘끓이다’를 인식하지 못한 것은 중국어권 학습자가 ‘ㄱ’과 ‘ㄴ’의 중간 음을 발음했기 때문인 것이다.

따라서 이러한 인식 불가 단어들을 분석하여 그 특성을 외국인 학습자를 위한 음성 인식 도구 설계에 반영한다면 자동 구어 전사 성공률을 높이는 데 크게 기여할 것이다. 가령 일본어권 학습자가 모국어의 영향으로 ‘ㅌ’를 ‘う’로 발음했다 하더라도 해당 음운을 한국어의 ‘ㅌ’와 매칭시켜 준다면 인식률을 높일 수 있다.

3.2.4. 결론

이상의 실험 결과로부터, 한국어 음성 인식기는 음성 인식률 면에서, 모국어 화자의 경우 인식 정확률은 약 83.5% 이상 수준인데 반해, 외국인의 경우 56% 수준으로 현저히 낮을 뿐만 아니라, 오인식과 인식불가가 차지하는 비중이 현저히 증가함을 알 수 있다. 또한, 실험에 사용된 두 음성 인식기의 성능 차이가 모국어 화자의 경우 오인식률은 7배(iOS 28%, Android 4%), 적정인식률은 약 0.25배(iOS 71%, Android 96%)로 다소 큰 차이는 보이는 데 반해, 한국어 학습자의 경우 오인식률은 iOS 40%, Android 34%, 적정인식률은 iOS 49%, Android 61%로 다소 작은 차이를 보인다. 즉, 음성 인식기의 모국어 화자 인식률의 성능 차이보다, 발화 주체가 모국어 사용자인지 한국어 학습자인지에 의한 변동 요인이 더 크다고 볼 수 있다.

이와 함께, 한국어 학습자의 음성 인식 결과 내에서도 어휘 유형별 > 수준별 > 언어권별 순으로 유의미한 상관관계를 보임을 알 수 있다. 그러므로 한국어 학습자의 구어 전사를 위한 음성 인식기는, 음성 인식률의 측면에서는 외국인 화자의 발음 특성, 수준을 감안할 수 있도록 설계되어 인식률을 개선해야 하며, 음성 인식 결과의 구분에 있어서는 학습자의 발화와 인식 결과와의 합치 정도를 제시하여 오인식 유무를 구분할 수 있는 기능이 필요하다.

4. 구어 전사 도구 개발안

앞서 실험에서 살펴본 바와 같이, 음성 인식기를 이용한 모국어 화자와 한국어 학습자 간의 음성 인식률에는 상당한 차이가 있다. 이러한 차이를 감수하고 모국어 화자용 음성 인식기를 사용하여 학습자 말뭉치를 구축하는 경우, 데이터 구축 과정에서 심각한 문제를 발생시킬 수 있다. 한국어 학습자 말뭉치에서는 음성 인식기의 오인식에 대한 해석과 처리가 달라져야 하기 때문이다. 예를 들어, 모국어 화자 발화에서 오인식된 ‘울음’, ‘음료수’는 단순히 음성 인식기의 인식 오류로 해석되어도 무방하다. 즉, 오인식 결과의 원인이 발화자의 발음 문제인지, 음성 인식기의 성능 문제인지를 구별하는 것이 어렵지 않을 수 있다. 반면에, 한국어 학습자의 경우, 오인식된 원인을 철저히 검증할 필요가 있다. 학습자 말뭉치의 핵심 가치는 이러한 오인식이 발생하는 부분에 있기 때문이다. 그러므로 한국어 학습자에 대한 음성 인식기의 성능을 향상시키는 것만으로는 한국어 학습자 말뭉치용 도구로 활용하기 어렵다. 더구나, 말뭉치 구축에 있어서 가장 많은 비용과 노력이 요구되는 부분이 이른바 ‘후처리’라고 부르는 검증/교정 과정임을 고려할 경우, 오인식에 대한 처리와 작업 공정이 반영되지 않는다면, 자동 전사로 부터 얻을 수 장점보다 더 큰 비용을 ‘후처리’에서 감당해야 할 수 있다.⁸⁾ 즉, 한국어 학습자 말뭉치 구축 측면에서 볼 때, 음성 인식기의 성능이 충분히 우수하다고 하더라도 단순한 음성 전사 기능만으로는 학습자 말뭉치 구축에 효과적으로 적용될 수 없다.

이러한 문제는 한국어 학습자 구어 말뭉치의 구축 공정 및 주석 체계와 같은 지침과 연계되어야 한다. 그러므로 여기에서는 한국어 학습자의 음성 인식을 위한 성능 개선 방안과 함께 한국어 학습자 구어 말뭉치 구축에 적용되는 전사 표지와 관련 문제들을 함께 검토하기로 한다.

8) 앞서 수행한 실험 결과를 단순 적용하여 10어절 발화의 음성 인식 결과가 인식불가 1어절, 오인식 3어절, 적정인식 6어절이라 하자. 오인식된 3어절을 식별하기 위해서는 10어절을 모두 검토해야 한다. 또한, 오인식된 3어절은 정상적인 한국어 어휘에 가까우므로 문맥을 통한 식별 노력이 필요하다. 이 경우 오인식률이 낮아질수록 오인식된 어절을 식별하기 위한 노력은 배가 된다.

4.1. 음성 인식을 개선 방안

3.2.3.의 실험 결과를 통해 알 수 있듯이 외국인 학습자가 음성 인식기를 사용하여 발화할 경우 다음의 두 가지 사항을 개선해야 한다.

첫째, 음성 인식을 높이기 위한 방안으로 음성 인식이 낮은 단어 즉, 학습자들이 빈번히 오류를 생성하는 단어의 음향 특성을 분석하여 음성 인식기 설계에 이를 반영해야 한다. 이를 통해 외국인 학습자의 음성 인식 실패 비율을 낮출 수 있다.

둘째, 인식 정확도를 높이기 위한 방안으로 음성 인식기가 음절 단위로 발화를 인식하도록 설계해야 한다. 현재의 음성 인식기는 화자의 소리를 내장되어 있는 어휘로 연결시켜 전사하는 데 기능이 치중되어 있다. 따라서 외국인 학습자가 발화한 음성을 그대로 문자화하지 않는다. 3.2.의 실험 결과, 한국어 학습자가 발화 오류를 생성했을 때 음성 인식기가 이를 그대로 문자화하지 않고 비슷한 단어로 매칭시켜 전사한다는 사실을 알 수 있었는데 몇몇 예를 보이면 다음 표와 같다.

제시 단어	인식 단어
똑같이	떡갈비집, 포카칩, 처갓집, 어떡하지, 똥강아지...
음료수	안녕, 전력수급, 이력서, 왜 말이 없쓰, 뉴스, 어벤저스...
설날	수능, 31, 서울말, 사우나, 5월...

<표 25> 오인식 단어 목록

일반 말뭉치와 달리 학습자 말뭉치는 말뭉치를 통해 학습자의 오류를 파악할 수 있어야 하므로 오인식된 음절 그대로가 화면에 출력되어야 한다. 예컨대 ‘똑같이’를 /또카치/로 발음했을 경우, 인식 결과가 내장 어휘인 ‘포카칩’으로 연결되지 않고 ‘또카치’ 그대로 전사되어야 하는 것이다. 이러한 문제를 해결하기 위해서는 음절 단위로 발화를 인식하는 도구 개발이 요구된다.

하지만 이러한 논의가 ‘음성 인식기가 반드시 음절 이하 수준에서만 발화를 인식하고 전사해야 한다’고 규정짓는 것은 아니다. 이상의 작업을 통해 음성 인식기의 발화 인식률을 높인 후, 효율적인 후처리를 위해 단어 수준의 인식 정확도를 높이는 작업 또한 이루어지는 것이 이상적이다.⁹⁾

9) 이러한 작업은 급수별, 언어권별 특성에 따라 다르게 진행되어야 하므로 추후에 다시 논의하기로 한다.

4.2. 전사 표지 개선안

음성 인식기를 통해 인식된 발화는 화면에 자동으로 전사된다. 이를 STT(Speech-To-Text)라 하는데 자동으로 전사된 발화는 전사자에 의해 다시 한번 전사 작업을 거쳐야 한다. 이때 주요하게 다루어지는 것이 전사 표지이므로 전사자는 이에 대한 지침을 미리 마련해야 한다.¹⁰⁾ 특히 구어 말뭉치의 경우 문어 말뭉치와 달리 주석이 많이 요구되는데 이러한 주석 체계를 설정할 때는 작업의 효율성, 기계 가독성, 다양한 데이터 특성을 수용할 수 있는 주석의 구조적 일관성과 발전성, 음성 자료와의 연동성 등을 고려해야 한다.

여기에서는 이 중에서 기계 가독성, 작업의 효율성, 주석의 구조적 일관성과 발전성을 중심으로 살펴보기로 한다.

1) 기계 가독성과 작업의 효율성

최근 접근과 활용이 용이한 학습자 구어 말뭉치로 MICASE(Michigan Corpus of Academic Spoken English)를 들 수 있다. MICASE는 약 180만 단어로 구성된 구어 말뭉치로 강의, 토론, 세미나 등 다양한 분야의 발화를 수집하여 구성하였으며, TEI(Text Encoding Initiative) 문서로 SGML로 마크업되었다.

책임자	Dr. Ute Römer, University of Michigan
기간	1997-2001
언어	영어
대상	88% 모국어 화자, 12% 비모국어 화자
구성	강의, 토론, 세미나, 미팅, 랩, 스터디그룹 등
규모	180만 단어
텍스트 개수	152개 (190시간 이상)
사이트	http://quod.lib.umich.edu/m/micase/

<표 26> MICASE 정보

MICASE의 전사 기호는 XML로의 표현이 용이하도록 되어 있는데 이를 표로 보이면 다음과 같다.

10) 다만, 이번 연구에서는 음성 인식기를 이용한 구어 전사 도구에 적용될 통합된 구어 전사 표지 체계를 제시하지는 않는다. 그 이유로는 첫째, 통합된 구어 전사 표지 체계는 앞서 제시한 음성 인식기를 전제로 하기 때문이다. 제안된 구어 전사 표지 체계를 검증하기 위해서는 한국어 학습자 말뭉치 구축을 위한 음성 인식기가 필요하기 때문이다. 둘째, 본 연구에서 지원하는 한국어 학습자 말뭉치 구축의 진행 상황상 구축되고 있는 구어 말뭉치 자료가 문어 말뭉치와 함께 일관성 있게 다루어지는 것이 우선되기 때문이다. 이는 향후 한국어 학습자 음성 인식 기능이 포함되더라도, 음성 인식과 관련된 사항들을 일관성 있게 처리하는 기반이 될 것으로 기대한다.

A. TRANSCRIPTION AND MARK-UP CONVENTIONS	
MEANING/DESCRIPTION	TRANSCRIPTION SYMBOLS
Speaker: Student (S:), Teacher (T:)	S: at the beginning of each turn or interruption/backchannel.
Two or more speakers, in unison (used mostly for laughter)	ST
PAUSES	
Comma indicates a brief (1-2 second) pause after a clause.	,
Period indicates a brief pause accompanied by an utterance final (falling) intonation contour; not used in a syntactic sense to indicate complete sentences.	.
Question mark indicates intonation contour typical of a question.	?
short pause (up to 1 second)	[]
longer pause	[...] up to 3 seconds [5s], [15s] measured pause (approximate time)
BACKCHANNEL CUES and FAILED INTERRUPTIONS	
BACKCHANNEL	T: the main utterance [Sx: Backchannel] the main utterance continues S: the main utterance [Tx: Backchannel] the main utterance continues
OVERLAP	
	S: [simultaneous utterance] T: [simultaneous utterance] the utterance continues
LAUGHTER	
All laughter is marked. Speaker ID not marked if current speaker laughs.	<LAUGH>, <S LAUGH> <ST LAUGH>, etc.
CONTEXTUAL EVENTS	
Various contextual (non-speech) events are noted, usually only when they affect comprehension of the surrounding discourse.	<WRITING ON BOARD>
READING PASSAGES	
Used when part of an utterance is read verbatim.	<READING>.....</READING>
UNCERTAIN or UNINTELLIGIBLE SPEECH	
Two x's in parentheses indicate one or more words that are completely unintelligible. Words surrounded by parentheses indicate the transcription is uncertain.	(xx)
NAMES-ANONYMISING	
When participants' names occur in a recording, they are changed to <NAME> in the transcript. Names of non-present people referred to in the recording are also changed, unless these people are public persons (authors of books, famous scientists, politicians).	

<표 27> MICASE 전사 기호

하지만 MICASE의 전사 체계는 억양, 불완전한 발화, 방언 등의 다양한 요소를 포함하고 있지 않으므로 구어의 특성을 모두 반영하는 데에는 한계가 있다. 또한 학습자 말뭉치의 경우 화자의 오류 양상을 기술하는 것이 중요하므로 오류 전사에 대한 지침 또한 마련되어야 한다.

현재 국내에서 구어 말뭉치를 전사할 때 많이 참고가 되는 지침은 21세기 세종계획에서 사용한 전사 기호이다. 21세기 세종계획의 전사 기호는 MICASE보다 항목을 상세히 나누어 기술했다는 장점이 있다. 하지만 표현 체계와 작업 편의성이 혼재하고 있다는 단점이 있는데 즉, 전사 작업의 편의성을 위해 약식 기호를 많이 사용하고 있어 기계 가독성이 떨어지는 것이다. 전사 이후의 작업을 위해 텍스트 내용을 XML로 표현하기 위해서는 기계 가독성이 용이해야 하므로 본 연구에서는 이러한 단점을 보완한 새로운 전사 체계를 제시하고자 한다. 전사자가 기호 입력에 대해 가지는 부담감은 UI(User Interface)를 사용하여 줄일 수 있다.

대분류	소분류	현행 기호	개선안	예시
발화자 정보 (person)	발화자 표시	1	<PERSON_ID WHO="P1">	<PERSON_ID WHO="P1">
	분명하지 않을 때	?	<PERSON ID WHO="UNKNOWN">	<PERSON_ID WHO="UNKNOWN">
	필요에 따라	모두/나머지/2,3	<PERSON_ID WHO="ALL"> <PERSON_ID WHO="OTHERS"> <PERSON_ID WHO="P2,P3">	<PERSON_ID WHO="ALL"> <PERSON_ID WHO="OTHERS"> <PERSON_ID WHO="P2,P3">
억양 단위 (intonation)	하강	.	<FALLING> </FALLING>	학생이 돼<FALLING>요</FALLING>.
	상승	?	<RISING> </RISING>	알고 있어<RISING>요</RISING>?
	약한 상승이나 하강	,	<LEVEL> </LEVEL>	아니<LEVEL>면</LEVEL>
	활기, 기운찬 어조	!	<ANIMATED> </ANIMATED>	<ANIMATED>아</ANIMATED>
	길게 발음될 때	::	<LENGTHENING> </LENGTHENING>	모두 <LENGTHENING>다</LENGTHENING>
	하나의 억양 단위가 끼어듦에 의해 끊어진 경우	-	<TRUNC_IN/>	<PERSON_ID WHO="P1"> 그런거<TRUNC_IN/> <PERSON_ID WHO="P2"> 그<RISING>래</RISING>? <PERSON_IDWHO="P1"> <TRUNC_IN/>있었어.
	두 억양 단위가 휴지 없이 이어질 경우	&	<CONTINUED/>	한국사람<CONTINUED/>그 다음
	한 어절 발화 도중 다른 억양 단위로 전사될 때 조사나 어미에	=	<SWITCHING/>	학교<SWITCHING/>를
겹침 현상 (overlap)	겹침 현상	표시하지 않음 ¹¹⁾		

11) ELAN을 사용하여 전사할 경우, 시간 표시 기능을 통해 겹침 현상의 파악이 가능하므로 별도의 표시를 하지 않는다.

잘 들리지 않는 부분 (uncertain)	잘 들리지 않는 부분	<X X>	<UNCERTAIN> </UNCERTAIN>	<UNCERTAIN>거기서</UNCERTAIN>
	전혀 들리지 않는 부분	<note>안 들림</note>	<ABSOLUTELY UNCERTAIN/>	<ABSOLUTELY UNCERTAIN/>
	들리지 않는 음절 수만큼	X	<UNCERTAIN COUNT=" "> </UNCERTAIN COUNT=" ">	지금<UNCERTAIN COUNT="2">XX </UNCERTAIN COUNT="2">에서 삽니다.
발음 오류 (error)	필수적인 음운 규칙을 지키지 않아 발음이 이상한 경우	()	<PRONOUNCIATION_ERROR> <ORIGINAL> </ORIGINAL> <CORRET></CORRECT> </PRONOUNCIATION_ERROR>	<PRONOUNCIATION_ERROR> <ORIGINAL>설날레</ORIGINAL> <CORRET>설날에</CORRECT> </PRONOUNCIATION_ERROR>
전사자의 설명 (researcher's comment)	전사자의 설명	기호 없이 메모	<COMMENT> </COMMENT>	<COMMENT>학생 모두가 ""네""라고 한다.</COMMENT>
준음성 (vocal description)	웃음, 기침, 하품, 재채기, 박수 등과 같은 언어 외적 소리	<vocal desc='웃음'>	<VOCAL DESC=" "/>	<VOCAL DESC="LAUGH"/>
기타 (각각의 이름)	구어적 변이형, 외국인 특유의 발음, 발음 오류	()	<VARIATION> <ORIGINAL> </ORIGINAL> <CORRET> </CORRECT> </VARIATION>	<VARIATION> <ORIGINAL>너</ORIGINAL> <CORRET>넣어</CORRECT> </VARIATION>
	끊어진 단어 (불완전하게 발화된 단어)	=	< T R U N C _ W O R D > </TRUNC_WORD>	<TRUNC_WORD>이약</TRUNC_WORD> 나누면서
	방언형 표시	<dia></dia>	<DIA> </DIA>	차는 <DIA>여일</DIA> 있어
	쉼 표시 (pause)	긴휴지{ }	<PAUSE DUR=" "/>	<PAUSE DUR="6.3"/>
		짧은휴지..	<PAUSE/>	활발한 편인<PAUSE/>다고

인용	<Q Q>	<QUOTATION> </QUOTATION>	<QUOTATION>우물 안 개구리 </QUOTATION>라는
책이나 자료를 보고 읽은 경우	<R R>	<REFERENCE> </REFERENCE>	<REFERENCE>좋은 글이 될 수 있다 </REFERENCE>라고 되어 있어요.
익명성 보장을 위한 마크업 (privacy)	<name>:사람 이름, 단체 이름 등	<PRIVACY> <NAME> </NAME> </PRIVACY>	<PRIVACY> <NAME>제임스</NAME> </PRIVACY> 입니다.
	<id-num>: 주민등록번호, 학번 등 개인식별번호	<PRIVACY> <ID_NUM> </ID_NUM> </PRIVACY>	<PRIVACY> <ID_NUM>900120</ID_NUM> </PRIVACY>
	<card-num> : 신용카드 번호	<PRIVACY> < C A R D _ N U M > </CARD_NUM> </PRIVACY>	<PRIVACY> <CARD_NUM>1234567890</CARD_NUM> </PRIVACY>
	<address> : 주소	<PRIVACY> <ADDRESS> </ADDRESS> </PRIVACY>	<PRIVACY> <ADDRESS>서울 서대문구 신촌동 </ADDRESS> </PRIVACY>
	<tel-num> : 전화번호	<PRIVACY> <TEL_NUM> </TEL_NUM> </PRIVACY>	<PRIVACY> <TEL_NUM>01012345678</TEL_NUM> </PRIVACY>

<표 28> 전사 기호 개선안

2) 주석의 구조적 일관성과 발전성

주석의 구조는 주석의 의미와 기능을 고려하여 설계되어야 한다. 앞서 제시한 새로운 전사 기호는 기계적인 처리를 위한 표현을 중심으로 제시되었다. 여기에서 대분류, 소분류로 구분한 정보들은 기존의 주석 체계들을 토대로 개선 발전시킨 것인데, 주석의 의미와 기능을 나타내고 있다. 이러한 주석의 의미와 기능에 대해서는 기계 가독성이 고려되어 있지 않다. 예를 들어, 억양 단위(intonation)나 잘 들리지 않는 부분(uncertain), 발음 오류 등의 일부는 음성 인식 과정에서 기계적인 처리와 식별이 가능할 수 있다. 또한, 주석 표지가 값을 갖느냐 갖지 않느냐, 상하 관계 유무, 속성 유무 등 기능적 역할에 따라서 적절한 기계적 처리와 해석을 위한 기제를 부여해야 한다.

억양 단위(intonation)나 잘 들리지 않는 부분(uncertain), 발음 오류 등은 처리 도구의 성능과 기능에 따라 결정되거나 전사 작업의 공정과 수준¹²⁾에 따라 달라질 수 있다.

값, 상하 관계, 속성 등의 유무와 관련한 문제는 주석의 표현과도 밀접한 관련을 갖는 것으로 다음과 같은 사례들을 고려하여 설계되어야 한다.

	유형	예시	비고
1	값이 없는 표지	<pause/>	
2	값을 갖는 표지	<name>철수</name>	가장 많은 비중을 차지한다.
3	값과 그 값에 대한 부가 정보를 갖는 표지	<PRONUNCIATION_ERROR> <ORIGINAL>설날레 </ORIGINAL> <CORRET>설날에 </CORRECT> </PRONUNCIATION_ERROR>	하위 주석에 관한 제약을 가져야 한다.
4	값이 유형을 갖는 표지	<VOCAL DESC="LAUGH"/>	
5	상하위 위계가 없는 표지	전사 주석들은 대부분 상하위 관계가 엄격하지 않다. <variant> 밑에 <원어절><교정어절>이 나타날 수도 있고, '잘 들리지 않음' 다음에 나타날 수도 있고, 심지어 오류주석에서도 나타날 수 있는 특성을 가지고 있다.	

<표 29> 주석의 기능/역할에 따른 구조적 유형

12) 품질 수준이 아니라 주석의 상세한 정도를 가리킨다.

이러한 새로운 주석 체계는 기계 가독성뿐 아니라 작업의 편의성, 형식 및 구조의 일관성 향상에도 도움을 줄 것이다.

4.3. 결론

선행 연구와 실험을 통해 외국인 학습자의 음성 인식 도구는 모국어 화자의 음성 인식 도구와는 설계가 다르게 이루어져야 한다는 사실을 확인할 수 있었다. 외국인 학습자가 빈번히 생성하는 오류 음운의 음향 특성을 분석하여 음성 인식기 설계에 반영한다면 음성 인식률을 높일 수 있다. 또한 인식 정확도를 높이기 위한 방안으로 음성 인식기가 음절 단위로 발화를 인식하도록 하는 작업 또한 필요하다.

음성 인식기를 통해 실제 구어 전사를 할 때에는 전사 표지를 미리 설정해야 하는데, 본 연구에서는 작업 편의성, 기계 가독성, 주석 형식 및 구조의 일관성, 음성 데이터와의 연동성 등을 고려해서 새로운 주석 체계를 마련하였다.

학습자 구어 말뭉치 구축 시, 일련의 과정의 통해 음성 인식기를 통한 자동 구어 전사가 가능해진다면 여러 가지 기대 효과를 가질 수 있다. 우선 작업의 시간과 노력을 줄여 전사 작업의 효율성을 높일 수 있다. 뿐만 아니라 전사 결과가 일관성을 갖게 되어 전체 작업의 질을 향상시킬 수 있다. 현재의 전사 작업은 전사자의 숙련도 또는 외국인 화자에 대한 노출 정도에 따라 전사가 다르게 이루어지고 있는 실정이다. 숙련도가 높을수록, 노출 정도가 많을수록 발음 용인이 커지는데 예컨대 한국어 학습자가 ‘울음’을 /우렴/으로 발음했을 경우 이를 ‘울음’으로 인식하고 정확한 형태로 표기하는 오류가 생길 수 있다.

따라서 외국인 학습자들의 구어를 자동으로 인식할 수 있는 학습자 전용 구어 전사 자동/반자동화 시스템이 개발된다면 학습자 구어 말뭉치 구축에 큰 도움이 될 것이다. 이뿐만 아니라 향후 다양한 말뭉치 구축에 있어 구어 자료 활용에 기여할 것이다.

VI. 한국어 학습자 말뭉치 구축 및 지원 도구 개발 설문 조사

설문 목적: 학습자 말뭉치 활용 도구를 개발함에 있어 잠재적 사용자들의 의견을 수렴하고 현 설계에서 개선 사항 여부를 찾아보고자 함

설문 기간: 2015년 11월 10일 - 23일

설문 방식: 설문지 조사 방식

설문 대상자: 한국어교육 관련 연구 혹은 교수 활동 종사자 (총 48명)

1. 설문 응답자별 분류

한국어교수 현장에 있는 응답자가 전체의 84%

✓ 응답한 총 48명 중 한국어교수자가 19명(42%), 한국어교수자 겸 연구자가 19명(42%), 한국어교육 연구자가 4명으로, 대부분이 한국어교수 현장에서 활동하고 있었음. 이 중에서 대학 부설 교육기관 소속이 32명, 대학원 소속이 11명으로 대부분 대학에 소속된 인원으로 나타났음.

✓ 이 중 한국어교수 경력이 5년 이상이 27명, 3년 이상 5년 미만이 6명으로 과반수 이상이 최소 3년 이상의 경험 있는 한국어 교수자로 구성됨.

2. 학습자말뭉치 사용 경험

사용 경험이 없는 응답자들의 학습자 말뭉치에 대한 높은 관심

✓ 한국어 학습자 말뭉치를 사용해 본 경험이 있다고 응답한 사람이 14명(30%), 그렇지 않다고 응답한 인원이 34명(70%)으로 집계됨. 그러나 이 34명 모두가 앞으로 사용 가능한 학습자 말뭉치가 구축되어 공개될 경우 적극적으로 이용할 의사가 있다고 답함.

✓ 학습자말뭉치를 사용해 본 경험이 있는 응답자들은 아직까지는 대부분 (17명 중 12명) 교수 자료가 아닌 연구 자료로만 이용해 본 적이 있다고 응답한 반면, 아직까지 사용해 보지 않았으나 앞으로 사용할 의향이 있는 응답자들의 응답 중 약 50%가 교수 자료로 사용해 보고 싶다고 답하여 학습자 말뭉치의 교수 현장에서의 실제 활용 측면에 관심이 많음을 알 수 있음.

3. 학습자 말뭉치의 사용 목적

학습자 말뭉치의 제한적 이용

✓ 학습자말뭉치를 사용해 본 경험이 있는 응답자들은 대부분 오류 분석 연구, 용례 검색, 빈도 조사, 언어 및 관용구 조사 등과 관련된 연구 목적으로 활용하였다고 답하였으나 <그림 1>에서 볼 수 있듯이 말뭉치 활용을 수업과 연계시켰다고 응답한 경우도 있었음.



<그림 138> 응답자들의 응답 결과의 빈도

학습자 말뭉치의 새로운 활용 방식에 대한 전망

✓ 새롭게 구축 중인 학습자 말뭉치에 대한 적극적인 활용 의사를 보인 응답자들은 (위 2번 항목 참조) 교수 현장에서 활용해 보고 싶다는 의사를 통해, 지금까지의 제한적 활용을 넘어 학습자 말뭉치가 새로운 다양한 방식으로 활용될 수 있을 것으로 예상해 볼 수 있음.

4. 학습자 말뭉치의 습득 경로

공개되어 사용할 수 있는 학습자 말뭉치의 부재

✓ 사용한 말뭉치의 습득 경로로는 직접 구축하거나 지인을 통해 습득하였다고 답한 경우가 약 60%, 공개 자료를 활용했다는 응답이 약 36%로 답함 (공개 자료 활용이라고 답한 응답자도 대부분 세종 말뭉치나 교재 말뭉치를 이용하였다고 응답).

✓ 이는 아직까지 연구나 교수 목적으로 활용할 만한 공개적인 학습지 말뭉치가 구축된 경우가 거의 없기 때문인 것으로 보임. 즉 교수자나 연구자가 적극적으로 활용할 만한 학습지 말뭉치가 구축될 필요가 있음을 다시 한 번 확인시켜 줌.

5. 학습자 말뭉치 활용 도구의 기본 기능에 관한 의견

기존 활용 도구의 기본 기능의 충실한 반영과 기초 통계 정보 추출 기능

✓ 학습자 말뭉치를 활용해 본 경험이 있는 응답자들은 말뭉치 분석 도구로서 MS 엑셀(33.3%), 한마루(18.5%), 깜짝새(15%) 등 말뭉치 활용 도구들을 사용해 본 적이 있다고 답하였음. 엑셀 프로그램의 사용 경험이 있다고 응답한 경우가 전체의 1/3 수준으로 나타난 것은 한마루 등과 같은 검색 도구를 이용하여 정보를 추출하고 엑셀 프로그램을 통해 추출한 정보의 목록 관리, 빈도 정보 계산 등 기초적인 통계 관리를 한 것으로 볼 수 있음.

✓ 이를 통해 말뭉치 활용 도구를 개발할 때 기존 도구들의 검색 기능 외에 기본적인 통계 정보 기능이 필요함을 알 수 있음.

활용 도구가 갖춰야 할 기본 기능은 하위 말뭉치 설정 기능과 기초 통계 관리 기능

✓ 이러한 요구는 말뭉치 활용 도구에 필요한 기능들을 묻는 항목에서 다시 한번 확인됨. 즉 검색 결과의 빈도 정보 제시 등 기초 통계 처리 기능에 대한 요구가 약 20% 수준으로 용례 검색 기능(20.8%), 하위 말뭉치 세부 설정 기능(23%)과 함께 가장 필요한 기능으로 꼽혔음.

✓ 가장 필요하다고 답한 기능이 하위 말뭉치 세부 설정 기능으로 나타났는데, 이는 말뭉치를 활용하여 학습자의 언어 사용 양상을 언어권별, 수준별 등으로 비교하여 분석하고자 하는 목적과 관련됨.

6. 학습자 말뭉치 활용 도구의 세부적 기능에 대한 의견

학습자 언어권별, 학습자 수준별 조건 검색에 대한 요구

✓ 용례 검색의 유형에 대한 질문에서는 학습자 언어권별 혹은 수준별로 나누어 검색이 가능하면 좋겠다는 응답이 가장 많았으며(54%), 여러 패턴을 서로 비교할 수 있게 비교 검색이 가능하면 좋겠다는 의견이 다음으로 많았음(15%). 기타 의견으로는 오류 유형별, 품사별, 철자별, 의미별 검색 등에 대한 요구가 있었음.

✓ 검색된 용례의 정렬 유형에 대한 질문에서도 위와 비슷한 분포의 응답이 나타났다는데, 학습자 언어권별 혹은 수준별 정렬 방식에 대한 응답이 가장 많은 약 43%로 나타났다. 검색 대상 말뭉치(하위말뭉치)를 설정할 수 있는 기능에 대한 요구 또한 있었는데 약 60%가 역시 학습자의 언어권별, 등급별 설정 기능을 요구함. 이는 말뭉치의 구축뿐 아니라 활용 도구 개발 측면에서 가장 우선적으로 반영해야 할 부분이라고 할 수 있음.

7. 학습자 말뭉치의 사용 환경에 대한 의견

일반 텍스트나 엑셀 파일 등의 익숙한 형태의 자료 선호

- ✓ 학습자 말뭉치의 데이터를 제공한다면 어떠한 자료 형식을 선호하는가에 대한 응답으로는 아래아한글 등의 일반 텍스트 파일 형태에 대한 답변이 약 46%로 가장 많았고, 엑셀 파일 형태가 약 29%로 뒤를 이음.
- ✓ 이는 기존에 사용하던 도구들로 다룰 수 있는 자료의 형태를 원하고 있음을 보여줌. 반면 XML이나 HTML 등의 구조화 문서에 대한 요구 또한 약 21% 수준으로 나타났음. 이는 아래의 활용 도구의 환경 문제와 관련됨.

온오프라인 복합 사용 환경에 대한 선호

- ✓ 어떤 매체를 이용하여 자료를 배포하는 것이 좋은가에 대한 응답으로는 웹 사이트라고 답한 응답이 약 52%를 차지하여 구조화된 문서 형태에 대한 선호가 21% 수준으로 나타난 이유를 간접적으로 설명해 주고 있음. 세종말뭉치와 같은 CD 혹은 DVD 형태에 대한 선호도 약 38% 수준으로 나타났음.
- ✓ 이와 관련하여 가장 효율적인 말뭉치 활용 도구의 환경으로는 PC + 웹기반 형태에 대한 응답이 100%로 나타나, PC 혹은 웹 등 한정된 환경이 아닌 언제 어디서든 이용 가능한 사용 환경을 원하는 것으로 나타남.

8. 기타 사항

기타 말뭉치 활용 희망 사항을 묻는 질문에 대한 답변에서도 언어권별, 수준별 오류 양상에 대한 검색 요구가 가장 많이 나타났으며, 특정 문법 현상이나 어휘 사용 양상 등 특정 언어 현상에 대한 검색 요구 역시 응답자들이 요구하는 사항으로 다시 등장하였음.



<그림 139>기타 말뚝치 활용 희망 사항

「한국어 학습자 말뭉치 구축 및 활용 도구」에 관한 설문 조사 설문지

안녕하십니까. 이 설문은 「2015 한국어 학습자 말뭉치 구축 및 활용 도구 개발」(국립국어원) 연구 과제 수행의 일환으로 한국어교육 연구자 및 교사의 의견을 수렴하기 위한 것입니다. 설문 결과는 사용자에게 보다 친숙하고 활용도 높은 말뭉치 구축과 활용 도구 개발을 위한 기초 자료로 사용할 것이며, 그 외에 다른 목적으로는 사용하지 않을 것임을 약속드립니다. 설문에 응해 주셔서 감사합니다.

문의 및 제안: 010-9594-7011, silhuett@iwaz.co.kr (곽용진 (주)아이와즈)
010-8727-9024, anna98kr@yonsei.ac.kr (홍혜란 연세대)

한국어 학습자 말뭉치와 구축 과제 소개

‘한국어 학습자 말뭉치’란 한국어를 학습하는 학습자들이 생산한 언어 자료를 수집하여 컴퓨터로 검색 및 활용할 수 있는 형태로 전산화한 자료를 말합니다. 한국어 학습자 말뭉치는 학습자의 언어 발달과 습득의 특성을 밝히기 위한 연구 자료로 활용될 수 있으며, 그 결과는 교재 개발, 교수법 개발 등의 실제적인 교수·학습에 적용하게 됩니다.

이를 위해 국립국어원에서는 본 과제를 통해 3단계, 6개년 계획으로 국내 한국어교육기관의 학습자(1단계)와 이주민(2단계), 국외 학습자(3단계)가 산출한 작문 및 말하기 자료를 수집할 예정입니다. 2015년 현재 기초 연구와 함께 국내 학습자 자료 구축에 착수하였으며, 향후 웹 기반의 학습자 말뭉치 시스템을 통해 수집된 자료를 한국어교육 연구자, 교사, 학습자에게 폭넓게 보급할 계획입니다.

I. 기본 정보

1. 관련 영역

- ① 한국어교육 관련 연구자(대학원생 포함)

4-2. 사용한다면 어떤 목적으로 사용하고 싶습니까?

- ① 교수 자료로 사용
- ② 연구 자료로 사용
- ③ 기타

4-3. 어떻게 사용하고 싶은지 자유롭게 써 주십시오.

(예) 언어권별, 숙달도별로 특정 문법의 오류를 단순 검색으로 찾고 싶다.
중국인 초급 학생들의 조사 ‘이/가’ 오류 양상을 찾고 싶다.

II. 학습자 말뭉치 구축에 관한 의견

1. 교육 현장에서 수집 가능한 학습자 대상과 언어 자료에는 무엇이 있는지 자유롭게 써 주세요.

(예) [1학년 중도입국자] - 작문 자료
[2학년] -발표 자료
[초등학생] -작문 자료

2. 각 교육기관별로 자료를 수집할 수 있는 적절한 시기가 있다면 알려주십시오.

3. 학습자 말뭉치에는 개인 정보가 들어 있을 수 있고, 연구용이기는 하지만 여러 사람이 공유하게 될 자료이기 때문에 연구 윤리의 차원에서 활용 동의서를 받게 됩니다. 학습자(아동 학습자의 경우 학습자 부모)에게 언어 자료 활용 동의서를 받는 것에 어려움을 없을까요?

※ [동의서]는 한국어 혹은 학습자의 언어로 번역된 것이 제공되며, 학습자의 이름은 이용되지 않습니다.

4. 말하기나 글쓰기 자료를 제공하는 학생이나 그 자료를 수집하는 교사에게 어느 정도의 보상(금전 혹은 물품)이 필요하다고 생각하십니까?

5. 학생들에게서 자료를 수집할 때 특히 고려할 점이 있으면 기술해 주십시오.

(예) 학교의 허락 절차

6. 국외인 경우 교육 현장에서 교포 학습자와 외국인 학습자의 비율은 어떻습니까?

Ⅲ. 학습자 말뭉치 활용 도구에 관한 의견

(말뭉치를 사용해 본 경험이 있는 분만 답하시면 됩니다.)

1. 다음은 국내에서 개발된 대표적인 말뭉치 활용 도구입니다. 활용해 본 적이 있는 도구에 모두 $\checkmark$ 표 해 주세요.

- ① 글잡이 ② 한마루 ③ 깜짝새
④ 기타 ()

2. 말뭉치를 분석하는 데에 주로 사용하는 프로그램은 어떤 것입니까? 사용해 본 적이 있는 도구에 모두 $\checkmark$ 표 해 주세요.

- ① 엑셀 ② SAS ③ SPSS ④ R
⑤ 기타 ()

3. 다음은 학습자 말뭉치 활용 도구에 포함되는 기능들의 예시입니다. 제시된 기능이 필요하다고 생각합니까?

가. 용례 검색 ① 네 ② 아니요

검색 유형: 학습자별, 등급별 검색, 형태소, 품사, 패턴, 오류 유형

 위 검색 유형에서 추가되어야 할 기능 또는 불필요하다고 생각되는 기능이 있으면 써 주세요.

- 1) 추가되어야 할 기능 : ()
- 2) 불필요한 기능 : ()

나. 용례 정렬 ① 네 ② 아니요

정렬 방식: 좌우 품사별, 좌우 패턴별, 오류 유형별, 학습자 변인별

☞ 위 검색 유형에서 추가되어야 할 기능 또는 불필요하다고 생각되는 기능이

있으면 써 주세요.

1) 추가되어야 할 기능 : ()

2) 불필요한 기능 : ()

다. 용례 저장 ① 네 ② 아니요

저장 형태: 검색된 용례를 텍스트 파일 혹은 마이크로소프트 엑셀 파일 형태로 저장

 위 저장 형태 외에 추가되어야 할 저장 형태가 있다고 생각하시면 써 주세요.

추가되어야 할 저장 형태 : ()

라. 하위말뭉치 설정 ① 네 ② 아니요

자료 재구성 방식: 학습자별, 등급별 등의 조건으로 하위 말뭉치를 구성하여 용례 탐색 등을 수행

 위의 재구성 방식에서 추가되어야 할 방식 또는 불필요하다고 생각되는 방식이 있으면 써 주세요.

1) 추가되어야 할 방식 : ()

2) 불필요한 방식 : ()

마. 통계 처리 ① 네 ② 아니요

통계 수치 표시 방식: 검색 조건별, 하위말뭉치 조건별로 표시

 위의 자료에 추가되어야 할 방식 또는 불필요하다고 생각되는 방식이 있으면 써 주세요.

1) 추가되어야 할 방식 : ()

2) 불필요한 방식 : ()

바. 그 외 추가되었으면 하는 기능 ()

4. 효율적인 학습자 말뭉치 활용 도구 사용 매체는 무엇이라고 생각합니까?

- ① PC ② 웹기반 ③ PC + 웹기반

5. 학습자 말뭉치의 바람직한 제공 형식은 어떤 것이라고 생각하십니까?

(복수 답변 가능)

- ① 전자 텍스트 ② 엑셀 ③ 데이터베이스(SQL 등)
④ XML ⑤ 기타 ()

6. 효율적인 학습자 말뭉치 자료 배포 매체는 무엇이라고 생각합니까?

- ① 출판된 책자 ② CD 또는 DVD ③ 웹사이트
④ 기타 ()

7. 한국어 학습자 말뭉치 구축 및 활용 도구와 관련한 의견 또는 제안 사항을 자유롭게 써 주십시오.

□ □ □ 설문에 응해 주셔서 감사합니다. □ □ □

<부록 2> 한국어 학습자 음성 인식 실험지

_____급 _____반 발음 인식 테스트

국적: _____ 이름: _____

테스트 기기: (iOS ‘Siri’ / Android ‘Google Now’)

관련 현상	번 호	단어	1차	2차	3차
연음	1	목요일			
	2	울음			
	3	음악			
	4	십이월			
평폐쇄음 화	5	부엌			
	6	맛			
	7	옆			
비음(동) 화	8	정류장			
	9	음료수			
	10	작년			
	11	십만			
유음화	12	연락			
	13	편리			
	14	설날			
경음화	15	갑자기			
	16	결석			
	17	식당			
	18	젓가락			
유기음화	19	그렇지만			
	20	빨강다			
	21	막히다			
구개음화	22	똑같이			
	23	붙이다			
자음군 단순화	24	값			
	25	닭			
	26	앉다			
ㅎ 탈락	27	문화			
	28	많이			
	29	끓이다			
ㄴ 삽입	30	지하철역			

<참고문헌>

- 강정희(2011), 베트남인 학습자의 발음 오류 분석 및 교육 방안 -자음 대조 및 학습자 자음인식 분석을 중심으로-, <한국문법교육학회 학술발표논문집>2011-2, 한국문법교육학회, 121-130.
- 곽용진(2003), 합목적적 말뭉치(Corpus) 자동 구축, <언어사실과 관점>12, 연세대학교 언어정보연구원, 31-69.
- _____(2008), 말뭉치 관리 시스템(CMS), <언어정보와 사전편찬>17-18, 연세대학교 언어정보연구원, 43-68.
- 권성미(2007), L2 음성 습득에 유사성이 미치는 영향 -일어권 한국어 학습자를 대상으로 한 실험음성학적 연구-, <한국어교육>18-3, 국제한국어교육학회, 1-25.
- _____(2011), 중국인 한국어 학습자의 중간언어에 나타나는 역양의 특성 연구 -문두 강세구와 문말 역양을 중심으로 -, <이중언어학>45, 이중언어학회, 1-25.
- 김규현(2009), 구어 자료의 전사 관행 -담화 및 대화분석의 예를 중심으로-, <언어사실과 관점>23, 연세대학교 언어정보연구원, 77-102.
- 김소현(2013), 베트남인 학습자의 한국어 마찰음과 파찰음 발음 연구, 경희대학교 석사학위논문.
- 김은애(2006), 한국어 학습자의 발음 오류 진단 및 평가에 관한 연구, <한국어교육>17-1, 국제한국어교육학회, 71-97.
- 김정숙, 김유정(2002), 한국어 학습자 말뭉치 구축을 위한 기초 연구 -개인정보 표지 체계와 오류 정보 표지 체계를 중심으로-, <이중언어학>21, 이중언어학회, 98-120.
- 김정윤(2007), 영어권 학습자의 한국어 자음 발음 연구, 한국외국어대학교 교육대학원 석사학위논문.
- 김지숙(2007), 베트남어권 학습자의 오류 연구, <한국어교육>18-1, 국제한국어교육학회, 49-71.
- 김한샘·서상규(1998), 말뭉치의 구성과 가공: 말뭉치의 구축과 활용-연세 말뭉치1의 구성과 실제-, <언어정보개발연구>1, 연세대학교 언어정보연구원, 74-104.
- 김한샘(2000), 말뭉치의 주석과 활용, <언어정보연찬회 발표논문집>14, 연세대학교 언어정보연구원, 17-27.
- 박기영(2001), 일본어 모어 화자의 한국어 발음 오류에 대한 일고찰 -특히 공명자음의 발음을 중심으로-, <관악어문연구>26, 서울대학교 국어국문학회, 225-245.
- 박수연(2007), 한국어 학습자 오류 말뭉치 구축과 그 문제점에 관한 연구, <언어사실과 관점>17, 연세대학교 언어정보개발원, 83-113.

- 박하나(2011), 스페인어 화자의 한국어 발음 오류 분석과 습득 패턴 연구, 한국외국어대학교 석사학위논문.
- 백소영(2008), 중국인 한국어 학습자의 발음 오류와 지도방법, <외국어교육연구>22-1, 한국외국어대학교 외국어교육연구소, 77-88.
- 범류(2010), 중국어권 학습자를 위한 한국어 발음 교육 연구 : ‘ㄹ’ 발음을 중심으로, 연세대학교 대학원 박사학위논문.
- 봉미경(2011), 다면성과 다면 말뭉치에 대하여 -다면 말뭉치 개발을 위한 기초 연구-, <언어사실과 관점>28, 191-215.
- 서경석(2007), 스페인어권 학습자에 대한 한국어 발음 교육 방안, <이베로아메리카>9-2, 부산외국어대학교 중남미지역원, 197-220.
- 서상규(2010), 한국어 학습자 말뭉치 구축 설계, 문화체육관광부.
- 서상규·유현경·남윤진(2002), 한국어 학습자 말뭉치와 한국어교육, <한국어교육>13-1, 국제한국어교육학회, 127-156.
- 서상규·한영균(2002), <국어정보학 입문>, 태학사.
- 심보 토모코·노채환(2009), 일본어권 학습자의 한국어 단모음 습득에 대한 고찰, <언어와 언어학>44, 한국외국어대학교 외국어 종합연구센터 언어연구소, 107-125.
- 양순임(2006), 중국인 학습자언어에 나타나는 한국어 종성에 대한 음향, 청취 음성학적 오류 분석, <한국어교육>17-3, 국제한국어교육학회, 163-183.
- 우인혜(1998), 한일 언어 비교를 통한 발음 교수법, <이중언어학>15, 이중언어학회, 319-347.
- 위국봉(2014), 중국인 학습자의 /-ㄴ, -ㅇ/ 종성 발음 오류 분석 및 교정 방안 연구, <외국어로서의 한국어교육>41, 연세대학교 언어연구교육원 한국어학당, 143-169.
- 위원징·홍미주(2013), 중국인 학습자의 청취, 발음 오류 분석을 통한 한국어 평음, 경음, 격음 교육 방안 연구, <한국어교육>24-4, 국제한국어교육학회, 155-191.
- 유석훈(2001), 외국어로서의 한국어 학습자 말뭉치 구축의 필요성과 자료 분석, <한국어교육>12-1, 국제한국어교육학회, 165-179.
- 윤영숙(2012), 중국인 학습자들의 한국어 강세구 실현양상과 오류진단 및 교정방안 연구, <말소리와 음성과학>, 한국음성학회, 51-59.
- 이선희(2009), 학습자 오류 주석 말뭉치 구축과 지능형 학습 도구 개발 -언어적 지식에 기반한 새로운 한국어교육-, <언어사실과 관점>24, 연세대학교 언어정보연구원, 187-220.
- 이승연(2007), 한국어 교육을 위한 한국어 학습자 말뭉치의 구축과 활용 연구, 고려대학교 대학원 박사학위논문.
- 이혜정(2015), 숙달도와 거주 기간에 따른 베트남인 학습자의 발음 습득 양상 -자음과 음운현상을 중심으로-, <이화어문논집>35, 이화어문학회, 143-172.

- 이화진(2011), 대조언어학적 접근과 언어 보편성을 통한 한국어 단모음 교육 순서: 중국인 학습자를 대상으로, <문법교육>15, 285-305.
- 이화진·하호빈(2012), 음절 구조와 음소 배열 제약을 고려한 한국어 자음 교육, <문법교육>16, 한국문법교육학회, 273-295.
- 장부연(2002), 일본어권 학습자의 한국어교육 개선 방안 -발음편-, <일본어교육연구>3, 한국일어교육학회, 209-222.
- 전나영(1993), 외국인을 위한 한국어 발음지도, <외국어로서의 한국어교육>18, 연세대학교 한국어학당, 151-169.
- 정명숙(2003), 한국어 발음 교육을 위한 음성 DB 구축 방안, <말소리>47, 대한음성학회, 51-72.
- 조철현(2002), 한국어 학습자의 오류 유형 조사 연구, 문화관광부.
- 하세가와유키코(20005), 일본어를 모어로 하는 학습자에 대한 음조 교육의 효과 - 어두 파열음 및 파찰음의 발음을 중심으로-, <한국어교육>16-3, 한국어교육학회, 379-404.
- 허용(2012), 외국인 학습자의 한국어 발음 오류에 대한 음운론적 분석 : 음운현상을 중심으로, <한국학논집>46, 계명대학교 한국학연구원, 201-232.

- Aarts, J., & Granger, S. (1998) "Tag sequences in learner corpora: a key to interlanguage grammar and discourse". In S. Granger (Ed.), *Learner English on computer*, pp. 132-141. London & New York: Addison Wesley Longman.
- Abuhakema, G., Faraj, R., Feldman, A., & Fitzpatrick, E. (2008). "Annotating an Arabic Learner Corpus for Error". In LREC. European Language Resources Association.
- Atkinson, M. and J. Heritage. (eds.) (1984). *Structures of Social Action: Studies in Conversation Analysis*. Cambridge: Cambridge University Press.
- Berglund, Y., & Prütz, K. (2000). "Tagging a learner corpus - a starting-point for quantitative comparative analyses". Presented at the ASLA conference KO RFU, November 11-12, 1999.
- Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.
- Biber, D., Randi, R., Victoria, C. and Jenia, W. (2001). "Representing spoken language in university settings: The design and construction of the spoken component of the T2K-SWAL Corpus." In *Corpus linguistics in North America: Selections from the 1999 symposium*, edited by Rita C. Simpson and John M. Swales, 48-57. Ann Arbor: University of Michigan Press.
- Burnard, L.(2007). "Reference Guide for the British National Corpus (XML E

- dition)." In. Oxford: Research Technologies Service at Oxford University.
- Carrió Pastor, M. L., & Mestre Mestre, E. (2014). "A proposal for the tagging of grammatical and pragmatic errors". *Research in Corpus Linguistics*, 1, pp.7-16.
- Chafe, W. (1988). "Linking intonation units in spoken English". In J. Haiman and S. A. Thompson (Eds.), *Clause Combining in Grammar and Discourse*, pp. 1-27. Amsterdam/Philadelphia: John Benjamins.
- Chafe, W. (1992a). "Intonation units and prominences in English natural discourse". *Proceedings of the IRCS Workshop on Prosody in Natural Speech*, pp.41-52. The Institute for Research in Cognitive Science Report No.92-37. Philadelphia: University of Pennsylvania.
- Chafe, W. J. W. Du Bois, and S. A. Thompson. (1991). "Towards a new corpus of spoken American English". In K. Ajimer and B. Altenberg (Eds.), *English Corpus Linguistics*, pp. 64-82. London & New York: Longman.
- Crowdy, S. (1993). "Spoken corpus design." *Literary and linguistic computing* no.8-4, pp. 259-265.
- De Haan, P. (2000). "Tagging non-native English with the TOSCA-ICLE tagger". In C. Mair & M. Hundt (Eds.), *Corpus linguistics and linguistic theory. Papers from the 20th International Conference on English Language Research on Computerized Corpora (ICAME 20) Freiburg im Breisgau 1999*, pp.69-79. Amsterdam & Atlanta: Rodopi.
- Díaz-Negrillo, A., & Domínguez-Fernández, J. (2006). "Error tagging systems for learner corpora". *Spanish Journal of Applied Linguistics (RESLA)*, 19, pp.83-102.
- Díaz-Negrillo, A., & García-Cumbreras, M. Á. (2007). "A tagging tool for error analysis on learner corpora". *International Computer Archive of Modern and Medieval English (ICAME) Journal* 31, pp.197-203.
- Dodigovic, M.(2012). "Speech Recognition Technology in Second Language Assessment: Social Group Preference". *International Journal of Information and Education Technology* 2-5.
- Du Bois, J. W., Schuetze-Coburn, S., Cumming, S., & N. Paolino. (1993). "Outline of discourse transcription". In J. A. Edwards and M. D. Lampert (Eds.), *Talking Data: Transcription and Coding in Discourse Research*, pp. 45-89. Hillsdale, New Jersey: Lawrence Erlbaum.
- Edwards, J. A. (1993). "Principles and contrasting systems of discourse transcription". In J. A. Edwards and M. D. Lampert (Eds.), *Talking Data: Transcription and Coding in Discourse Research*, pp.3-31. Hillsdale, NJ:Lawr

ence Erlbaum.

- Elimat, Amal Khalil, AbuSeileek, Ali Farhan.(2014). "Automatic speech recognition technology as an effective means for teaching pronunciation". JAL T CALL Journal 10-1, pp.21-47.
- Eskénazi, M.(1999). "Using Automatic Speech Processing for Foreign Language Pronunciation Tutoring: Some Issues and a Prototype". *Language Learning & Technology* 2-2, pp.62-76.
- Hamidi, F., Baljko, M.(2013). "Automatic speech recognition: A shifted role in early speech intervention?" SLPAT 2013. Fourth Workshop On Speech and Language Processing for Assistive Technologies. Grenoble.
- In-Seok, K.(2006). "Automatic Speech Recognition: Reliability and Pedagogical Implications for Teaching Pronunciation". *Educational Technology & Society* 9-1, pp.322-334.
- Neri, A., Cucchiarini, C., Strik, H.(2003). "Automatic Speech Recognition for second language learning: How and why it actually works". *Proceedings of 15th International Congress of Phonetic Sciences*. Barcelona. Spain, pp.1157-1160.
- Nicholls, D. (2003). "The Cambridge learner corpus - error coding and analysis for lexicography and ELT". In D. Archer, P. Rayson, A. Wilson, & T. McEnery (Eds.), pp. 572-581. Presented at the Corpus Linguistics 2003 Conference (CL 2003), Lancaster University: University Centre for Computer Corpus Research on Language.
- Ono, Y., Ishihara, M., Onishi, A., Yamashiro, M.(2014). "Classrooms and voice recognition applications in a foreign language teaching". *Proceedings of 43rd International Congress on Noise Control Engineering*. The Australian Acoustical Society.
- Psathas, G. and T. Anderson. (1990). "The 'practices' of transcription in conversation analysis". *Semiotica* 78-1/2, pp.75-99.
- Ragheb, M., & Dickinson, M. (2013). "Inter-annotator Agreement for Dependency Annotation of Learner". In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 169-179. Atlanta, Georgia.
- Reznicek, M., Lüdeling, A., & Hirschmann, H. (2013). "Competing target hypotheses in the falko corpus: A flexible multi-layer corpus architecture". In N. Ballier, A. Diaz-Negrillo, & P. Thompson (Eds.), *Automatic treatment and anal*

- ysis of learner corpus data, pp. 101-124. Amsterdam & Philadelphia: John Benjamins.
- Stubbs, M. (2001). "Texts, corpora, and problems of interpretation: A response to Widdowson." *Applied Linguistics* no. 22-2, pp.149-172.
- Tono, Y., Satake, Y., & Miura, A. (2014). "The effects of using corpora on revision tasks in L2 writing with coded error feedback". *ReCALL* 26(Special Issue 02), pp.147-162.
- Vaclav, B.(2013). *Compiling and analysing a spoken academic corpus*. Lancaster University.
- Yongjin, K. Pyung, K.(2013). "Quality Management Model and System for Large Corpus", *JNIT(Journal of Next Generation Information Technology)*, Vol 4-8, pp. 517-524. AICIT.

<Abstract>

2015 A Research on Korean Learner's Corpus Building Assistant Toolkit Development

This study aims to present the development plan of tools for the learner corpus and some usage examples whereof in order to give support to the Learner Corpus of Korean Language Project. It is hoped that the collection, management, distribution, and use of the learner corpus data are carried out in a more effective way through this study.

We have designed a user-friendly global system for construction, management, distribution, and use of a corpus data by establishing mid- and long-term development and implementation plan and making a website and tools for the corpus data organization of the year. A user survey on the learner corpus, the analysis of prior learner corpus projects and retrieval systems, domestic and overseas, and the investigation of the needs of working group have been conducted for it. Illustrative examples are also provided where a corpus is retrieved using freely and commercially available tools, and a way forward of developing an effective transcription tool is presented for later consultation.

The details of this study are as follows:

First, it is proposed that the target system is developed in a incremental way to equip the working group with appropriate tools according to the progress of the work, given that both the corpus data collection and the support tools development are conducted at the same time.

Secondly, the development of a global system is presented in this study that helps to collect, manage , distribute, and use corpus data in a more efficient way. The mid- and long-term schedule for the development and implementation of the system has been established based on the requirements analysis of the working group, user surveys of the potential uses of the corpus, and the review of prior learner corpus projects and retrieval systems.

For the year, a tool for data collection and management has been developed. This tool is composed of a web-based module that operates on the metadata in the corpus files, and a PC-based module that manipulates and inspects the corpus files. It takes the original materials from learners (scanned images of writing data and audio data), their metadata (learner profiles), annotations (mark-ups, POS-tagging, error tagging) as separate inputs. It then connects and organize the separate information into a newly generated data form, which helps to improve the data consistency and the data compatibility.

Especially the system adopts the stand-off type of mark-up encoding that has been widely used in large data collection projects in recent times. It incorporates the different mark-ups and annotations within a single format, combined and consistent, making it possible for users to retrieve the corpus to their diverse purposes.

Finally, an initial study of the development of a tool for transcribing spoken data has been carried out in order to save the laborious effort for processing audio data. It takes much more personnel and temporal resource to manage the spoken data than dealing with written data. In the study, some problems resulting from characteristics of learner speech as well as the transcription work itself have been determined. To solve this, it is proposed that the transcription tool should take into consideration the characteristics of learner speech so that it can recognize the speech at least in a phonemic level. In addition to that, it is suggested that the annotation scheme of the global corpus processing system above should be applied to the transcription mark-up syntax too.

연구 책임자: 곽용진((주)아이와즈 책임연구원)

공동 연구원: 박진호(서울대학교)

김한샘(연세대학교)

유현조(서울대학교)

이화진(홍익대학교)

권택수((주)아이와즈 팀장)

오지훈((주)아이와즈 선임연구원)

황용주(국립국어원 한국어진흥과 학예연구관)

연구 보조원: 이지연(연세대학교 대학원 박사과정)

최지명(연세대학교 대학원 박사과정)

담당 연구원: 황용주(국립국어원 한국어진흥과 학예연구관)

2015년 한국어 학습자 말뭉치 구축 지원 도구 개발 연구

발 행 인 송 철 의

발 행 처 국립국어원
서울시 강서구 금남화로 154
전화: 02-2669-9742 전송: 02-2669-9727

인 쇄 일 2015년 12월 11일

발 행 일 2015년 12월 11일

인 쇄 세종기획

※ 이 보고서는 국립국어원 누리집(<http://www.korean.go.kr>)에서도
내려받을 수 있습니다.