해외한국학지원사업(학술연구) 최종보고서

Korean Studies Grant 2008

- Final Report -

과제번호(Grant Number)	AKS-2008-R15
연구과제명 (Title of Project)	한국어 교육자료 공유지원화를 위한 말뭉치 구축: 일본어 모어화자의 작문 데이터베이스화를 중심으로
연구책임자 (Project Director's Name)	임현정(LIM Hyunjung)
소속(Affiliation)	야마구치현립대학
지원기간(Grant Period)	2008년 6월 – 2009년 5월

지원금(Amount of Grant)	US$ 8000 (￥853,500)	2008년6월 24일 환율
지출액(Expenses)	US$ 8000 (￥853,500)	
잔액(Balance)	US$ 0	

1. 연구과제 평가(Appraisal) : 계획대비 결과 비교, 연구결과, 향후 계획 등

(Outline of Activities, Activities completed, Comparison between Plan and Outcome, Results of Project, Future Plans, etc.)

본 연구는 해외, 특히 일본에서 한국어를 학습하는 일본어 모어 화자의 작문에 보여지는 오용 데이터를 토대로한 말뭉치를 구축하는 것에 그 목적이 있다.

최근, 「한류」 「한국 붐」이라는 사회 현상과 함께, 세계적으로 한국어 학습자 수가 급증함에 따라, 한국어 교육에 대한 국내외의 관심이 높아지고 교육 수준의 질적인 향상을 위한 노력이 한층 다양화 되고 있다. 그러나, 이러한 한류붐을 탄 한국어 학습열은 이미 식어가는 추세에 있는 것도 사실이다. 예를 들어, 일본의 4년제 대학에 있어서의 한국어 교육의 실시교의 비율을 보면, 한국어 학습자 수는 1995년도부터 2002-03년도까지의 기간에 전체적으로 25.3%에서 47.7%로 22.4 포인트나 증가했었다(2003 년도, 재단법인 국제 포럼 조사). 그러나, 2005년도 이후 일본 전국의 대학에서의 한국어 학습자의 증가는 정체 상화에 있다. 반면, 학습자의 요구는 확실히 높아져 가고 있어, 한국어 교원에 대해서는 보다 질 높은 교육, 학습자의 필요와 학습 익숙도에 맞춘 교육 컨텐츠의 개발, 효과적인 교수법의 지원등이 요구되고 있다. 한국어 학습자의 수업 개선에, 학습자 코파스를 활용한다, 즉 학습자가 낳은 대량의 데이터를 객관적, 동시에 수량적으로 분석하는 것은, 학습자의 입장으로부터 틀리기 쉬운 한국어의 특성을 파악해, 학습자에게 한국어를 보다 올바르게 이해시켜, 가능한 한 오용을 줄이기 위한 유효한 교수방법이나 교재 개발에 연결될 것이다.

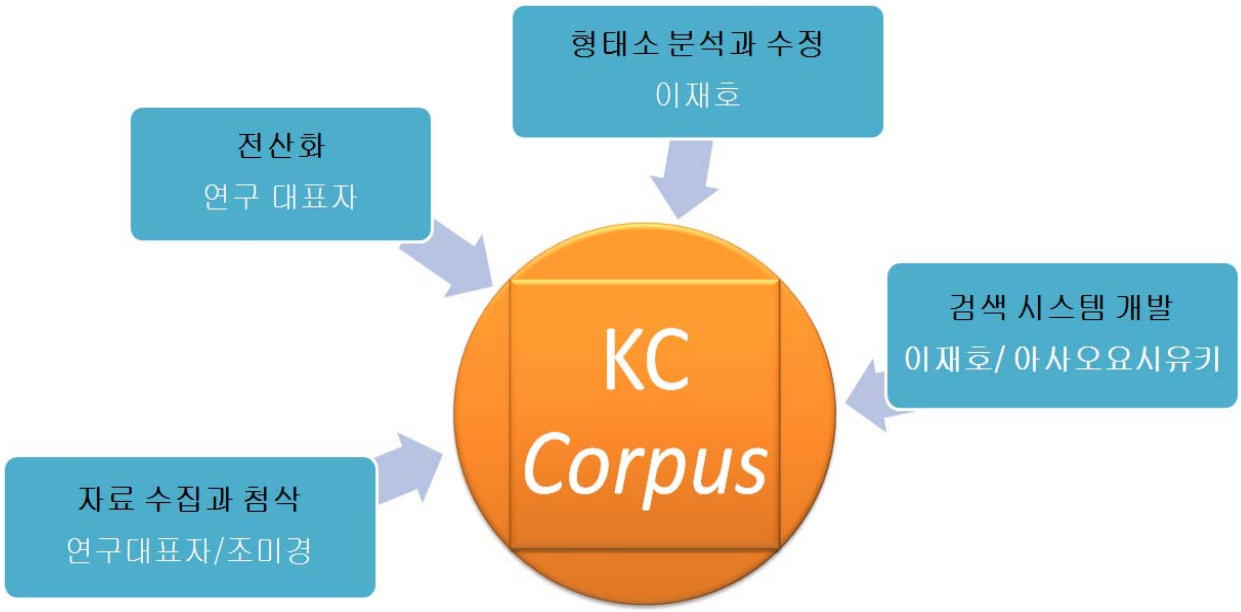
제2 언어교육, 특히 영어 교육의 세계에서는, 1990년대 이후, 학습자에 대한 교육 개선의 고찰 자료로서 학습자가 산출한 데이터 그 자체를 대량으로 모아 놓은 학습자 말뭉치가 주목을 받아 왔다. 학습자 말뭉치란, 제2 언어학습이 산출한 문장·발화 데이터를 대량으로 수집해 필요한 정보를 부가함으로서 다양한 검색이 가능한 형태로 전자화한 것을 말한다. 학습자 말뭉치는 학습자가 산출한 대량의 생 데이터에 대해 객관적· 수량적인 분석을 통해 보다 신뢰성 높은 분석을 가능하게 함과 동시에, 주관적 경험으로부터 얻을 수 있는 지견과는 또 다른 지견을 제공해 준다고 할 수 있다. 이러한 학습자

말뭉치의 구축은 그 활용성이 다방면에서 기대되고 있다. 특히, 학습자 말뭉치는 발달 단계에 있는 학습자의 중간언어를 반영하고 있어, 학습자의 오용 분석을 통해 제 2 언어교육의 효과적인 교수 방안 개발등에도 이용할 수 있다는 점에서 상당한 가치를 지니고 있다.

한편, 한국어에 있어서의 학습자 말뭉치는, 연세 대학교 어학당의 학습자를 대상으로 데이터를 수집한 연세 한국어 학습자 말뭉치(26만 어절, 2002년) 외, 몇개의 학습자 말뭉치가 존재하고 있지만, 일반 공개되고 있는 것은 지극히 적고, 검색 시스템에 대해도 그 사용 방법이 용이하지 않기 때문에, 학습자 말뭉치를 활용한 연구는 아직 불충분하다 (서상규 외, 2002). 또, 학습자의 오용 분석에 있어서는, 학습자가 교실 이외의 장면에서 한국어를 접할 기회가 어느 정도 있는가 등의 학습자의 학습 환경은 매우 중요한 변수 요인이 될 수 있다. 그렇지만, 기존의 한국어 학습자 코파스는 한국 국내에서 한국어를 학습하는 학습자의 데이터를 수집한 것이 대부분이기 때문에, 해외에서, 특히 일본에서 한국어를 학습하는 학습자의 특성을 반영하고 있지 않는 것은 것이 많아, 일본에서 한국어를 학습하는 학습자의 오용 요인을 파악하는데 한계가 있다.

이러한 문제점을 고려하여, 연구대표자인 본인은 공동연구자인 이재호 박사(현재 국제교류기금)와 조미경 교수(규슈대학), 연계 연구자인 아사오 요시히코 씨(교토대학)와 함께 KC(Korean L2 Learners' written Composition) Corpus(이하, KC Corpus)를 구축하였다.

본연구에서는 특히 ①일본의 대학에서의 교실 학습을 주로 하는 학습자의 작문 데이터의 전산화, ②언어 정보(형태소 해석 정보 · 오용 태그 · 첨삭 정보)를 부여한 말뭉치의 제공, ③언어 정보를 용이하게 검색해, 가공·분석할 수 있는 검색 시스템 개발을 목표로 하였다. 연구 조직의 구성과 담당 내용은 각각 다음과 같다.



이하에서는 KC Corpus의 구체적인 구축현황과 성과를 보이겠다.

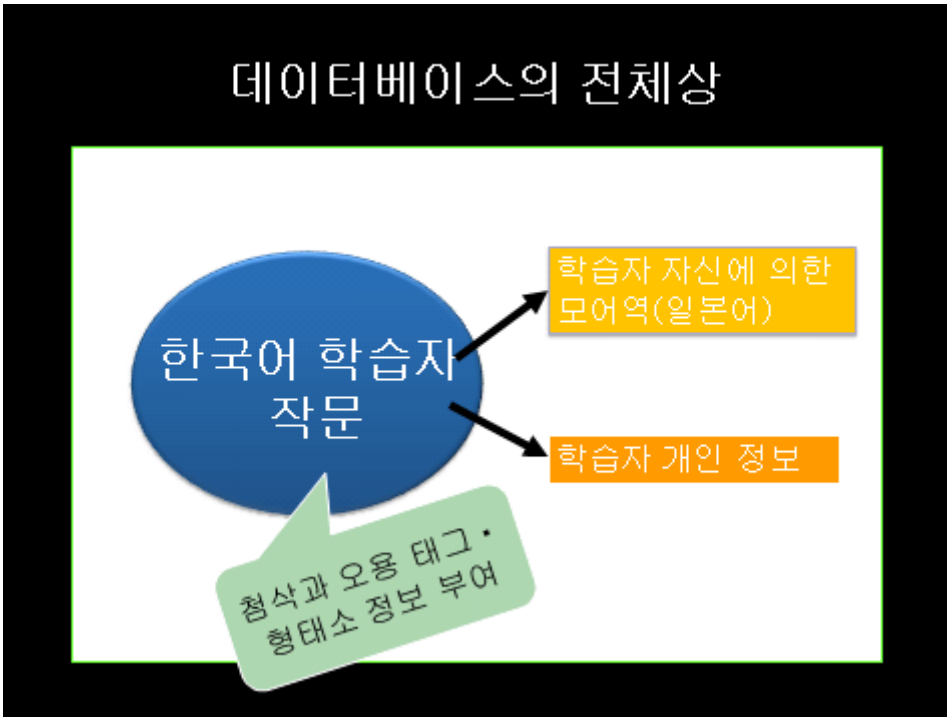
KC Corpus의 구축현황

1. KC Corpus 규모

현재까지, 야마구치현립 대학과 구슈 대학의 한국어 학습자 152명분의 데이터가 수록되어 있다. 학습자가 생산한 표본의 어절은 총 20905이며, 평균어절은 137.532894737이다(자료1 참조). 외국인 학습자가 생산한 표본의 평균 어절이 초급인 경우 60-70어절을 넘기 어렵고, 중/고급 수준의 학생의 경우도 100-120어절 내외의 작문을 하고 있는 실정(고석주외, 2004)을 고려하면, 구축 초년도의 자표 수집 양으로서는 어느 정도 성과를 거두었다고 할 수 있을 것이다. 앞으로 타기관에 협력을 요청해, 규모를 점차적으로 확대시켜 갈 예정이다.

2. KC Corpus 전체상

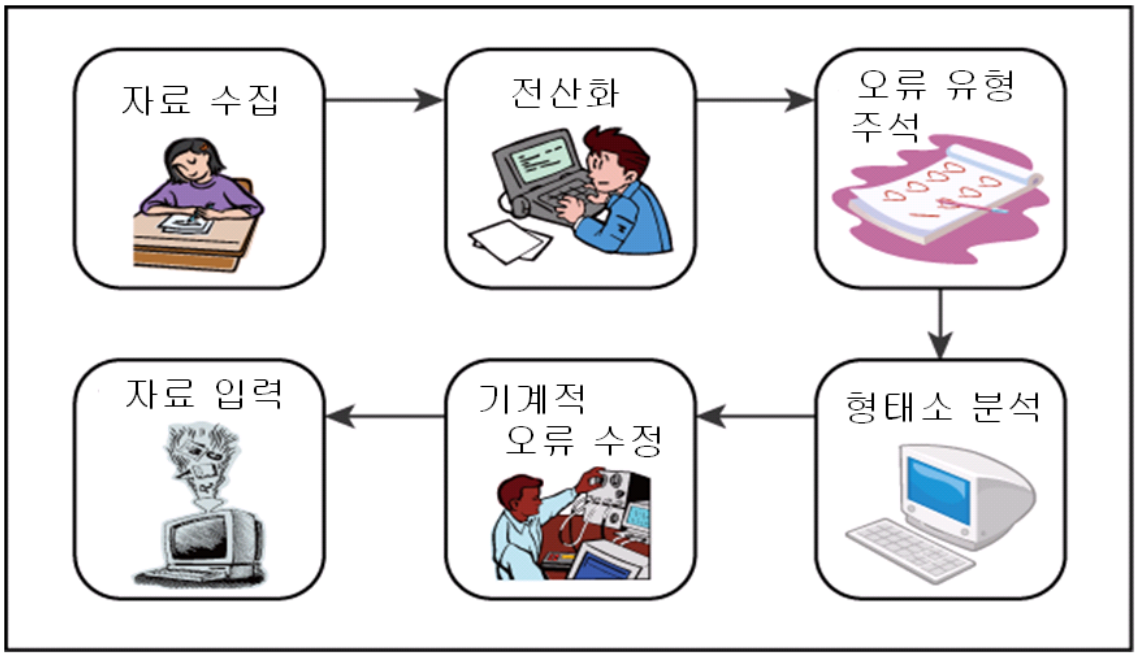
본말뭉치는 세 개의 데이터베이스, ①학습자의 한국어 작문, ②학습자 자신에 의한 모어 번역(일본어 대역), ③학습자 개인 정보로 구성되어 있다.



①한국어 작문은 본말뭉치의 코어 데이터로 한국어 학습자가 산출하는 한국어 작문을 말하며, 각각 첨삭 형태의 오류 주석의 과정을 걸친 데이터이다. ②작문자 자신에 의한 모어 번역은, 교사가 학습자의 작문을 첨삭할 때, 학습자 데이터와의 비교, 즉 학습자의 작문 의도를 추정하기 위해서 학습자 자신이 한국어 작문을 자신의 모어인 일본어로 대역을 쓰게 한 것이다. 사도시마(2001)은 대역 첨삭에 미치는 영향에 대해서, 첨삭자는 대역을 참조함으로 해서 집필자가 말하고 싶었던 것에 보다 가까운 표현을 선택해, 제시할 수 있다고 하고 있다. ③학습자 개인 정보는, 교사의 교육 연구 이용을 목적으로 작성한 것으로, 이하의 항목을 설정하였다. 생년월일, 성별, 한국어 학습력, 주된 학습 장소, 한국어 검정시험 자격의 유무, 한국에서 1개월 이상 생활한 경험의 유무, 집에서 한국어를 하는 사람의 유무, 사전 사용의 유무, 학습 수준, 작성 날짜, 작문 내용, 제목으로 총 12개이다 (자료2 참조).

3. KC Corpus 구축순서

구축 순서는, 아래와 같은 6 단계 작업을 걸쳤다.



3-1. 자료 수집

자료 수집은 2008년 7월-9월, 2009년2월-5월의 한국어 수업(90분) 중에 실시하였다. 자료 수집 방법과 테마를 구체적으로 보면 다음과 같다.

3-1-1. 자료 수집 방법

(1)설명 작문과제의 배포(5분)

①조사 취지에 관한 설명 : 수집한 데이터는 한국어 교육 개선을 위해서만 사용되며 그 외의 목적으로 사용되는 경우는 없음을 설명한다.

②「승낙서」 : 조사에 대해 충분한 설명을 받았으며, 자신이 쓰는 한국어 작문과 그 모어역를 연구·교육 목적을 위해서 공개하는 것을 승낙한다는 취지의 「승낙서」(자료 3 참조)에 서명을 받는다.

③작문 과제 : 작문 과제를 배포.과제는 몇 종류를 준비해, 어느 쪽의 과제에 대해서 쓸까는 학생 자신으로 선택하게 하지만, 수업 운영의 형편상, 클래스 전원에게 같은 과제로 쓰게 하는 것이 좋으면 담당 교원이 판단했을 경우는,

교원이 과제를 선정해도 좋다. 용지는, 400자 원고용지(A4, 가로쓰기)를 사용.
(2) 작문 작성(60 분)
①작문 작성 : 수업 시간내에 한국어로 작문을 작성한다 집필. 시간은 60 분. 문자수는 800 자 정도를 목표로 한다. 단, 학생의 한국어 레벨이나 수업 시간의 길이에 의해서, 수업 시간내에 800 글자를 쓰는 것이 곤란하다고 생각되는 경우는 이것을 밑도는 분량이여도 가능하다.
②사전 사용에 대해 : 작문을 할 때는 사전을 참조해도 지장없지만, 반드시 학습자 한 명의 힘으로 쓰도록 지시한다. 또, 가능한 한 숙제로는 하지 않게 한다.
③학습자 개인 정보 : 학습자에게 개인 정보에 대한 조사표를 기입하게 한다(자료4 참조).
(3) 모어 대역(25 분)
①한국어 작문을 보면서 일본어 역을 작성하도록 지시한다.
(4) 회수
①자료의 번호화 : 수집 후, 각각의 한국어 작문, 모어 번역, 학습자 정보 조사표에 ID번호를 단다.

3-1-2. 주제별 분류

학습자 작문을 주제별로 분류를 하면 다음과 같다.

No	내 용	구 분	표본수	어절	평균
1	일본 문화 소개, 한국과 일본 문화 비교	자유 작문	60	8300	138,33
2	일본에서 유행하는 것	자유 작문	4	476	119,00
3	건강을 위해서 하고 있는 것	자유 작문	32	5653	176,66
4	성형 수술에 대한 당신의 의견은?	찬성과 반대 중 하나를 골라 자신의 견해 쓰기	18	2115	117,50
5	자기 소개	자유 작문	21	2122	101,05
6	미래의 자기에게 보내는 편지	자유 작문	5	766	153,20
7	장래의 꿈	자유 작문	7	759	108,43
99	그 외	자유 작문	5	717	79,67

3-1-3. 학습자 수준

학습자 수준의 레벨 판정은, 학습자의 한국어 관련의 검정시험 자격, 작문 작성시의 학습 레벨과 학습력·유학의 유무와 기간등을 종합 해 6 단계로 정했다. 야마구치대학에서 수집한 자료는 본연구 대표자가, 구슈대학에서 수집한 자료는 구슈대학의 조미경교수가 최종적으로 판정을 하였다.

학습자 수준의 기준은 다음과 같다.

학습자 수준	기 준
레벨 1	90분 수업을 28-30시간 수강하는 가운데, 약 450어의 기초적인 어휘나 기본 문법에 대해 학습 경험을 가진다. 사전을 찾으면서 간단한 문장을 만들 수 있다. 한글 능력 검정시험 5급 이하.
레벨 2	90분 수업을 28-30시간 수강한 정도. 450어 정도의 기초적인 어휘와 기본 문법에 대한 이해가 정착되어 있어, 상투어로서의 인사나 간단한 질의응답을 할 수 있다. 한글 능력 검정시험 5급 정도.
레벨 3	90분 수업을 56-60시간 수강한 정도. 자기 소개·쇼핑, 음식점에서의 주문 등 생활에 필요한 기초적인 언어를 구사할 수 있어 친밀한 화제의 내용을 이해, 표현할 수 있다. 950어 정도의 기초적인 어휘와 기본 문법을 이해할 수 있으며 간단한 문장을 만들 수 있다. 한글 능력 검정 시험 4급 · 한국어 능력 시험 초급(1·2) 정도.
레벨 4	90분 수업을 70-120시간 수강한 정도. 전화나 부탁 정도의 일상생활에 필요한 언어나, 우체국, 은행등의 공공기관으로의 회화를 할 수 있다. 1500-2000어 정도의 어휘를 이용한 문장을 이해하고 사용할 수 있다. 한글 검정 3급·한국어 능력 시험중급(3급) 정도.
레벨 5	90분 수업을 160-200시간 수강한 정도. 일상생활에 지장이 없고, 여러가지 공공 시설의 이용이나 사회적 관계를 유지하기 위한 언어 사용이 가능.문장어와 구어의 기본적인 특성을

	이해할 수 있어 사용이 가능하다. 한글 능력 검정 시험 준 2급 · 한국어 능력 시험중급(4 급) 정도.
레벨 6	사회적 상식의 범위내에 있는 화제를 대부분 이해할 수 있다 .또, 신문의 사설등을 읽어 이해할 수 있다. 한국어로 논리적인 문장이 책, 이야기를 할 수 있다. 한글 검정 2급·한국어 능력 시험 상급(5급·6급) 이상.

3-2. 전산화

자료 수집단계를 걸친 자료들은 각 기관과 생년월일, 성별, 한국어 학습력, 주된 학습 장소, 한국어 검정시험 자격의 유무, 한국에서 1개월 이상 생활한 경험의 유무, 집에서 한국어를 하는 사람의 유무, 사전 사용의 유무, 작문 내용, 제목, 학습수준, 작성날짜 별로 라벨화해 전산화하였다.

데이터의 라벨화

ID	생년월일	성별	한국어 학습력	학습 장소	한국어 검정시험 자격	한국에서 1개월 이상 생활한 경험	집에서 한국어를 하는 사람	사전 사용	작문 내용	제목	학습 수준	작성 날짜,
y10028	1990년 2월	女性	1년	대학	있음(한글능력검정시험 5급)	없음	없음	사용	1		2	2008년 7월-9월
y10029	1989년 7월	女性	1년 1개월	대학	없음	없음	없음	사용	1		2	2008/7月-9月
y10030	[illegible] 7월	[illegible]		대학	있음(한글능력검정시험 5급)	없음	없음	사용	1		2	2008/7月-9月
y10[illegible]				[illegible]학	있음(한글능력검정시험 5급)	없음	없음	사용	1		2	2008/7月-9月
y200[illegible]				[illegible]학	있음(한글능력검정시험 4급)	없음	없음	사용	2	일본의 한류 붐	3	2008/7月-9月
y20004				[illegible]	없음	없음	없음	사용	2	일본에서 유행하는 「아라시」	2	2008/7月-9月
y20011	1987년 8월	女性	4년	대학	있음(한글능력검정시험 3급)	없음	없음	사용	2	일본의 웃는 얼굴	5	2008/7月-9月
y20014	1986년 8월	女性	4년 3개월	대학	있음	있음(2007年3月から2008年2月)	없음	사용	2	일본에서 유행하고 있는 것	5	2008/7月-9月
y30013	1987년 2월	女性		대학	있음(한글능력검정시험 3급·한국어 능력 시험4급)	있음(2007年3月から2008年1月)	없음	사용	3	건강	5	2008/7月-9月

y10029
y (yamaguchi)
1(작문 내용)
0027(학습자)

라벨화가 끝난 자료는 한국어 작문과 일본어 대역을 각각 마이크로

소프트 워드에서 입력을 하였으며, 이 때 일정한 양식을 갖춘 학습자 정보를 표본에 입력하였다. 입력시 특히 주의한 점은 다음과 같다.

①학습자가 생산한 원문을 그대로 반영한다. 그러나, 일본인 화자들의 경우는 모국어인 일본어의 특성상 뛰어 쓰기를 무시하는 경우가 많은데, 이러한 경우, 띄어쓰기 오류에 대해서는 오류에 포함시키지 않고, 입력자가 띄어쓰기 원칙에 맞게 입력을 하였다. 그 이유는 요류 태그를 어절 단위로 잘라서 부착할 필요가 있기 때문이며, 어절 빈도를 산출할 때도 띄어쓰기가 기준이 되기 때문이다.
②한글을 제외한 그 외의 언어는 원문 그대로 입력한다
③맞춤법의 오류로 인해 글자가 깨져서 입력되는 경우, 이를 표시하기 위해서 다음과 같은 방법을 취하였다. [예:아프ㅂ니다]
④입력이 완성된 한글파일은 동일한 파일명으로 텍스트 파일로 저장지킨다. 또한 원문을 스캔한 PDF파일도 동일명으로 저장을 하였다.

3-3. 오류 유형 주석

본연구는 앞서 언급한바와 같이 선행 학습자 말뭉치와의 공유 가능성을 검토하는 것을 하나의 연구 목표로 하고 있다. 따라서 본연구에서의 오형 유형은 선행 학습자 말뭉치 중에서도 가장 규모가 큰 연세 말뭉치를 참고로 하였다. 그러나, 시간과 예산의 제약으로,오류 태그 세트는 구체적으로 작성하되, 이번에는 크게 3종류의 오용 타입만을 설정하여 태그를 부착하기로 하였다. 아래는 연세대학교 언어연구 교육원의 오류 태그 세트를 참고로 하여 본연구에서 작성한 오류 태그 세트(자료 5 참조)와 3종류의 오류 타입이다. 오용의 종류는 다음과 같다.

오용 내용		오용 태그
문법 오용	활용이나 문법적 측면의 실수 (형태·활용, 문법 요소의 탈락등)	G
문체 오용	표현이나 어용론적 측면의 실수 (어순·표현적 오용등)	S
표기 오용	문자 표기의 실수 (한국어에는 없는 문자·어휘·표현등)	C

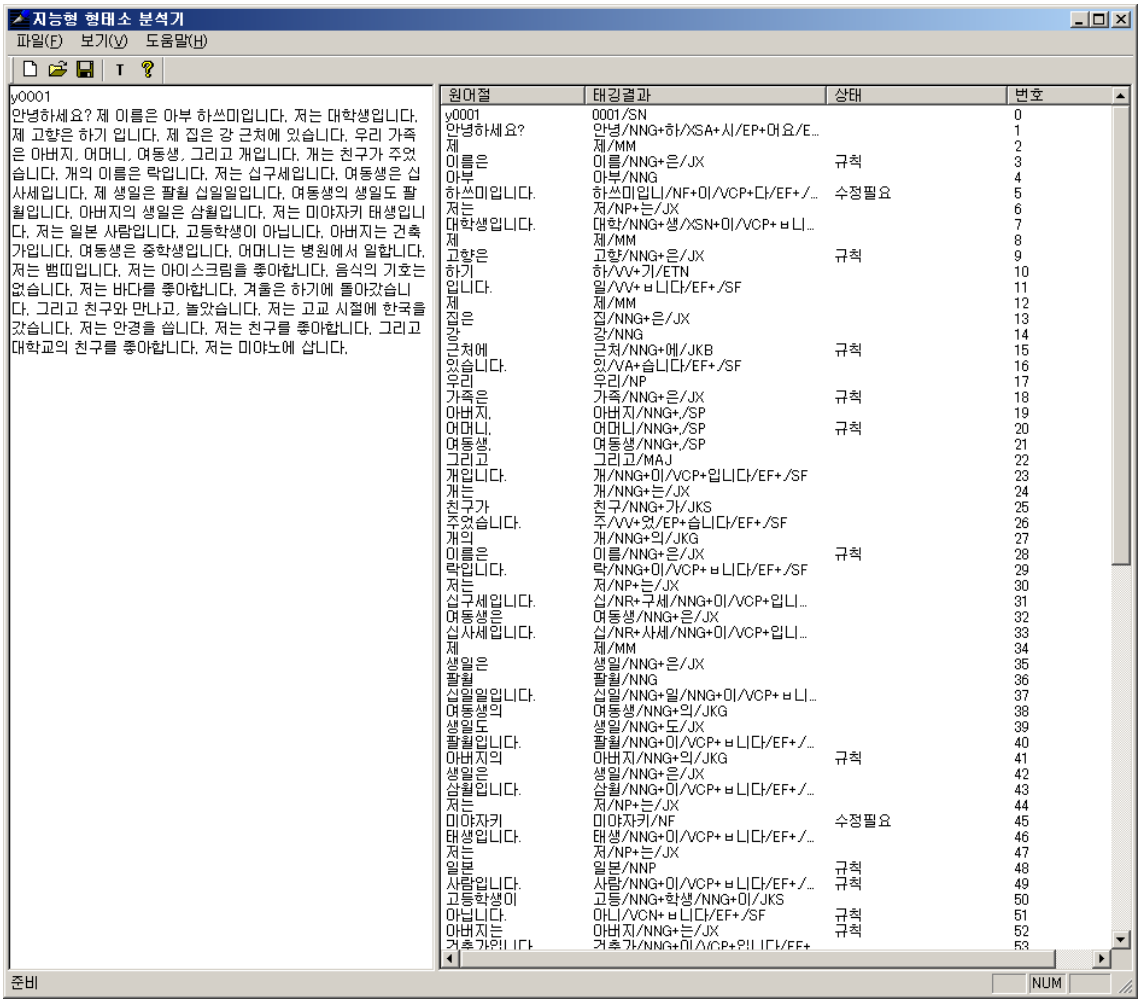
오용 주석의 태그화에 있어서는 전자화한 학습자의 작문을 엑셀화 한 후, 본연구 대표자와 구슈대학의 조미경 교수가 첨삭 정보를 직접 추가 기입하였다. 상기의 오용은 각각 태그화해, 후술 하는 검색 시스템에서도 검색할 수 있게 되어 있다. 또, 형태소 정보가 보관 유지되고 있어 어느 특정의 품사에 관한 오용예등도 수집할 수 있게 되어 있다. 예를 들면, 격조사에 관한 전오용예를 수집한다고 하는 태스크도 가능하다.

현재, 오용의 인정에 관해서는 연구 대표자와 조미경 교수의 교육자로서의 직관에 의지해 가고 있지만, 장래적으로는 제삼자에 의한 판정의 체크 작업등도 생각하고 있다. 또, 향후 선행 학습자 말뭉치와의 데이터 공유를 목표로 해, 첨삭 정보에 대해서는 여러가지 경우에 대응할 수 있도록, 현재 검토중의 오용 타입과 같이, 설정 가능한 오용 타입의 키워드를 수시 추가해 나갈 예정이다.

3-4. 형태소 분석과 수정작업

형태소 분석은 이재호 박사가 실시하였다. 형태소 해석에 대해서는, 세종계획(http://www.sejong.or.kr)이 공개하고 있는 「지능형 형태소 분석기(http://www.sejong.or.kr/dist_frame.php)」를 사용했다.

세종계획「지능형 형태소 분석기」



그 결과, 다음의 문제점이 밝혀졌다.

①형태소 해석기 그 자체의 기능이 그만큼 높지 않다
②학습자 작문에는 비규범적 표현이 많이 혼재하고 있다

현재 일본에서 넓게 이용되고 있는
Chasen(http://chasen.naist.jp/hiki/ChaSen/)이나
MeCab(http://mecab.sourceforge.net/)라고 하는 형태소 해석기가 평균 95% 이상의 해석 정도를 자랑하는데 대해, 한국어의 형태소 해석기는 신문 데이터도 80% 전후의 해석률 정도인 점을 고려하면 결코 고정밀도라고는 할 수 없다.

또한 ②의 문제로서 학습자 데이터에는 많은 오용예가 포함되어 있는 등, 입력 데이터로서 대량의 노이즈를 포함하고 있다. 그 때문에, 해석 에러가 매우 많다. 특히 명사에 관한 오해석이 많아, 경계 인정에서는 70%, 품사 인정에서는 60% 전후의 해석률 정도였다. 따라서 본연구에서는 형태소 해석기로 처리를 실시한 뒤, 데이터중에 에러가 있을 경우에는 다시 한번 사람 손으로 수정을 하여, 보다 신뢰성의 높은 데이터를 작성하였다. 그리고, KC Corpus 검색 시스템에 데이터를 투입해, 정상적으로 검색할 수 있을지를 확인했다. 본 시스템의 품사 체계는 전술한 지능형 형태소 분석기의 품사체계를 이용하고 있다. 구체적으로는 아래와 같다.

대분류	소분류	세분류
체언	명사 NN	일반명사 NNG 고유명사 NNP 의존명사 NNB
체언	대명사 NP	
체언	수사 NR	
용언	동사 VV	
용언	형용사 VA	
용언	보조용언 VX	
용언	지정사 VC	긍정지정사 VCP 부정지정사 VCN
수식언	관형사 MM	
수식언	부사 MA	일반부사 MAG 접속부사 MAJ
독립언	감탄사 IC	
관계언	격조사 JK	주격조사 JKS 보격조사 JKC 관형격조사 JKG 목적격조사 JKO 부사격조사 JKB 호격조사 JKV

		인용격조사 JKQ
관계언	보조사 JX	
관계언	접속조사 JC	
의존형태	어미 E	선어말어미 EP 종결어미 EF 연결어미 EC 명사형전성어미 ETN 관형형전성어미 ETM
의존형태	접두사 XP	체언접두사 XPN
의존형태	접미사 XS	명사파생접미사 XSA 동사파생접미사 XSV 형용사파생접미사 XSA (부사파생접미사 XSB)
의존형태	어기 XR	
기호	마침표,물음표,느낌표	SF
기호	쉼표,가운뎃점,콜론,빗금	SP
기호	따옴표,괄호표,줄표	SS
기호	줄임표	SE
기호	붙임표(물결,숨김,빠짐)	SO
기호	외국어	SL
기호	한자	SH
기호	기타기호(논리수학기호,화폐기호 등)	SW
기호	명사추정범주	NF
기호	용언추정범주	NV
기호	숫자	SN
기호	분석불능범주	NA

3-5. 검색 시스템 개발

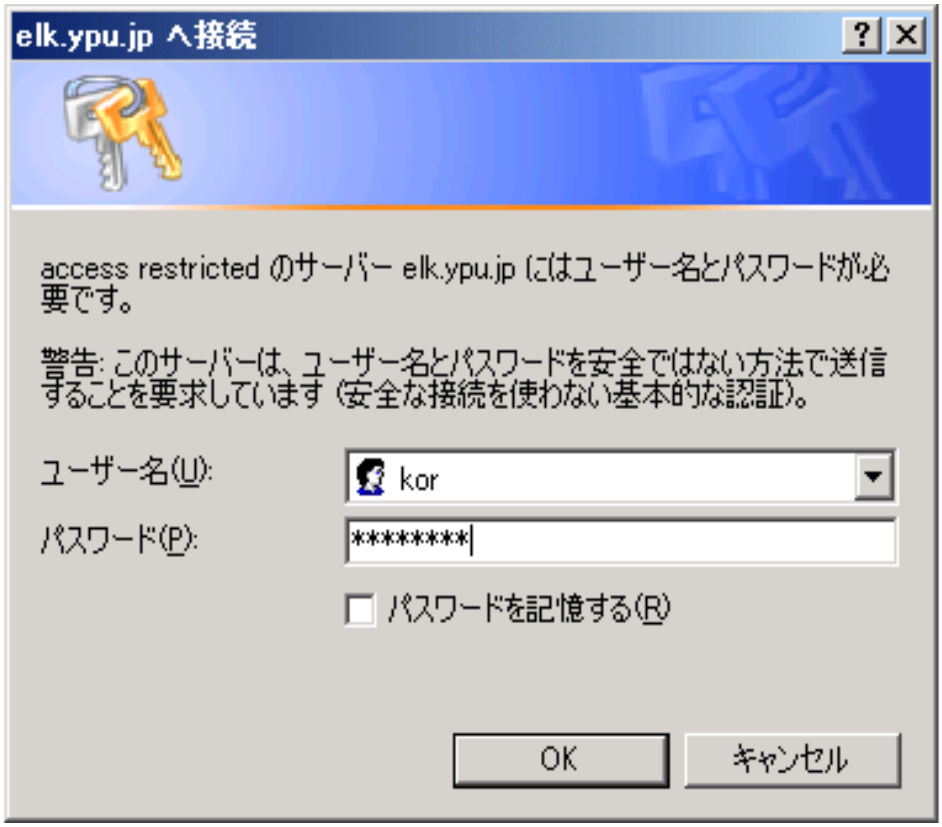
오류 주석의 효율적인 이용과 학습자 오류 연구를 위해서는 전용의 검색 시스템이 반드시 필요하다. 검색 시스템 개발에 있어서는 전문 지식이 없이도 손 쉽게 사용할 수 있고 자작 데이터에도 간단하게 응용할 수 있는 검색기 개발을 목표로 했다. 검색 시스템의 제작은 이재호 박사와 아사오요시히코 씨가 담당했다.

검색 시스템 환경으로서는 2종류를 준비했다. VBA(Visual Basic for Applications)를 베이스로 구축한 엑셀 환경에서의 이용과 웹 베이스의 이용 환경이다. 엑셀의 환경에서는, 일본어 코파스의 검색 시스템인 E-KWIC 차마메용 (아사오·이 2008)을 이용하여, 엑셀의 매크로를 실행하는 것만으로, KWIC (Keyword in Context) 검색을 할 수 있는 시스템을 개발했다. 웹 베이스의 이용 환경에서는, 베이직 인증 수속을 실시하는 것만으로, KWIC 검색과 데이터의 보존을 할 수 있는 시스템을 구축했다.

Microsoft Excel 를 이용하는 메리트로서 인문계의 연구자도 일상적으로 사용하고 있는 경우가 많아, 사용에 큰 어려움이 없다는 것을 들 수 있다. 또, 검색 결과를 그대로 Excel 의 워크시트에 보존할 수 있으므로, 검색 결과를 한층 더 가공·분석하는 일도 용이하다. 이 검색 시스템에서는 지정된 검색어에 대해서, Keyword in Context (KWIC) 형식으로 학습자의 용례를 표시할 수 있다.

본검색 시스템은 단순하게 검색어를 지정한 검색외, 품사 또는 오용의 종별에 의한 추출에 대응하고 있으며, 검색어를 지정하지 않고 특정의 품사의 말을 모두 뽑아내는 사용법도 가능하다. 또, 학습자 레벨에 의한 추출도 가능하다. 오류 검색 방법은 다음과 같다.

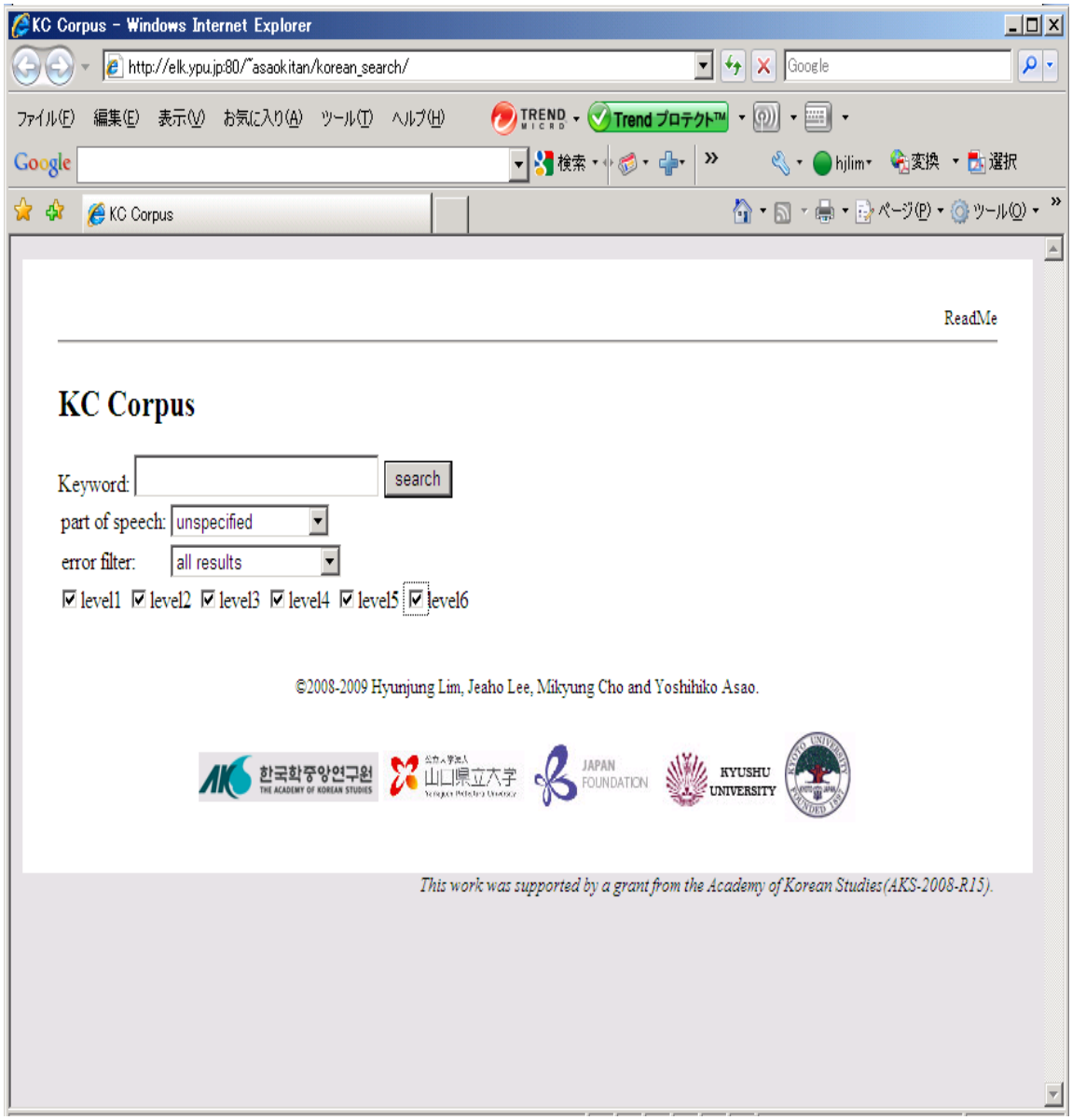
(1)검색 시스템이 들어있는 서버에 악세스(ID와 패스워드는 사전 인증이 필요)



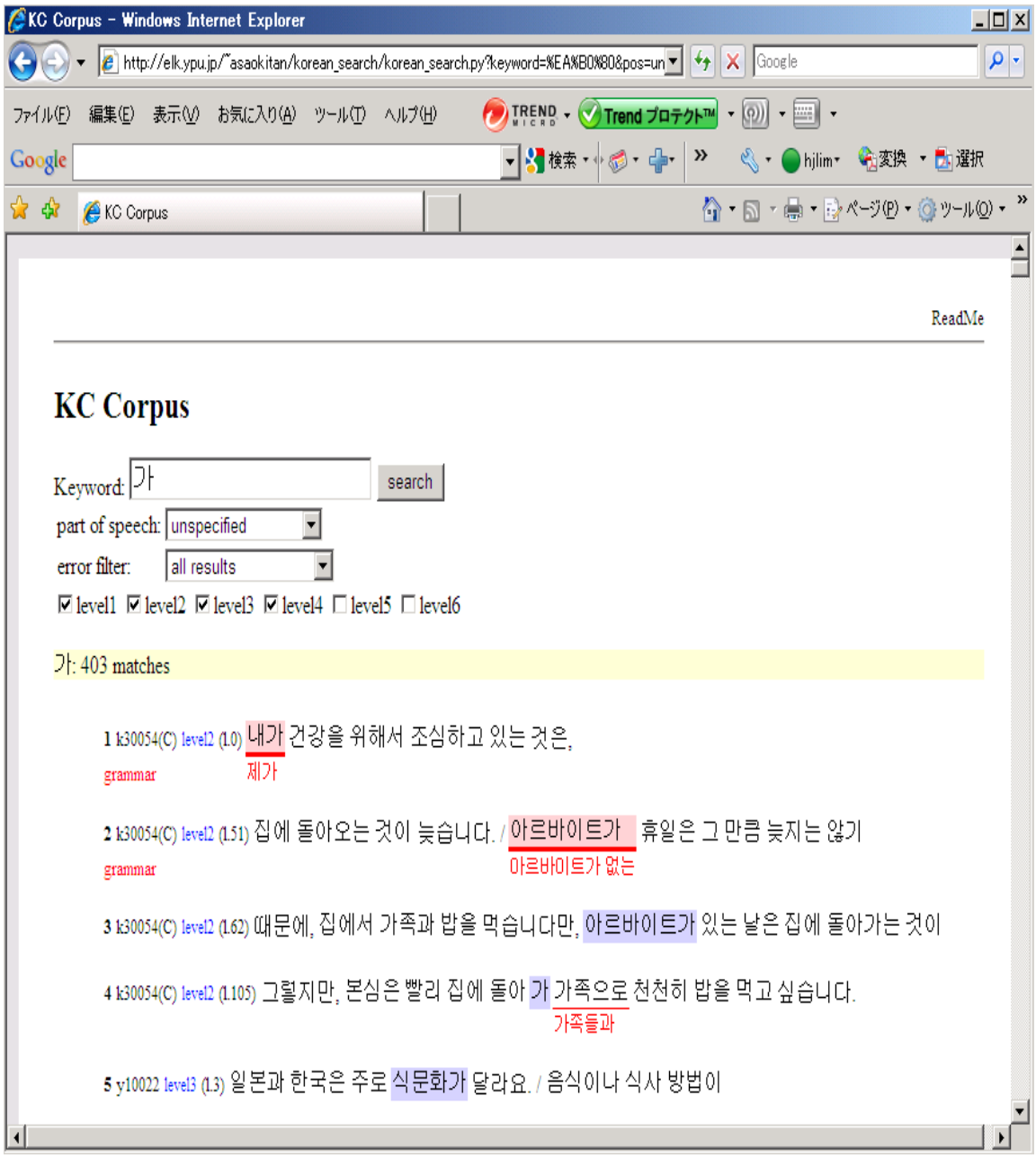
(2)검색 창

학습자 오용 말뭉치의 오류를 검색하려고 할 때는, 검색하고자 하는 조건에 대한 체크박스를 체크한 후에 검색 버튼을 눌러주면 검색조건에 맞는 말뭉치와 개수가 출력된다, 이때 찾아보려고 하는 오류의 조건은 사용자가 임의로 선택할 수 있다. 아래의 그림은 검색의 조건 입력부와 그 화면이다.

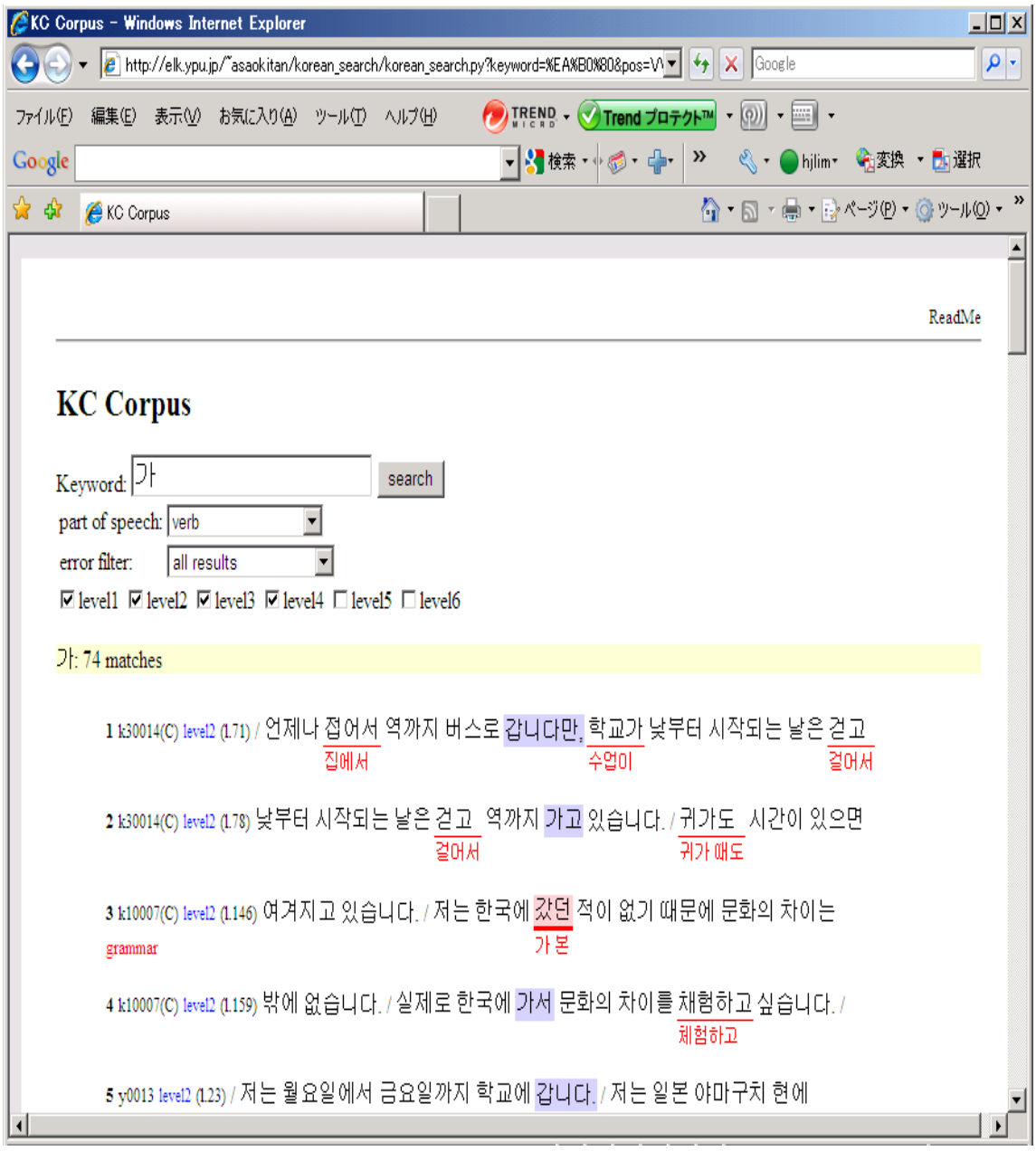
●단순한 문자열로 검색한 화면



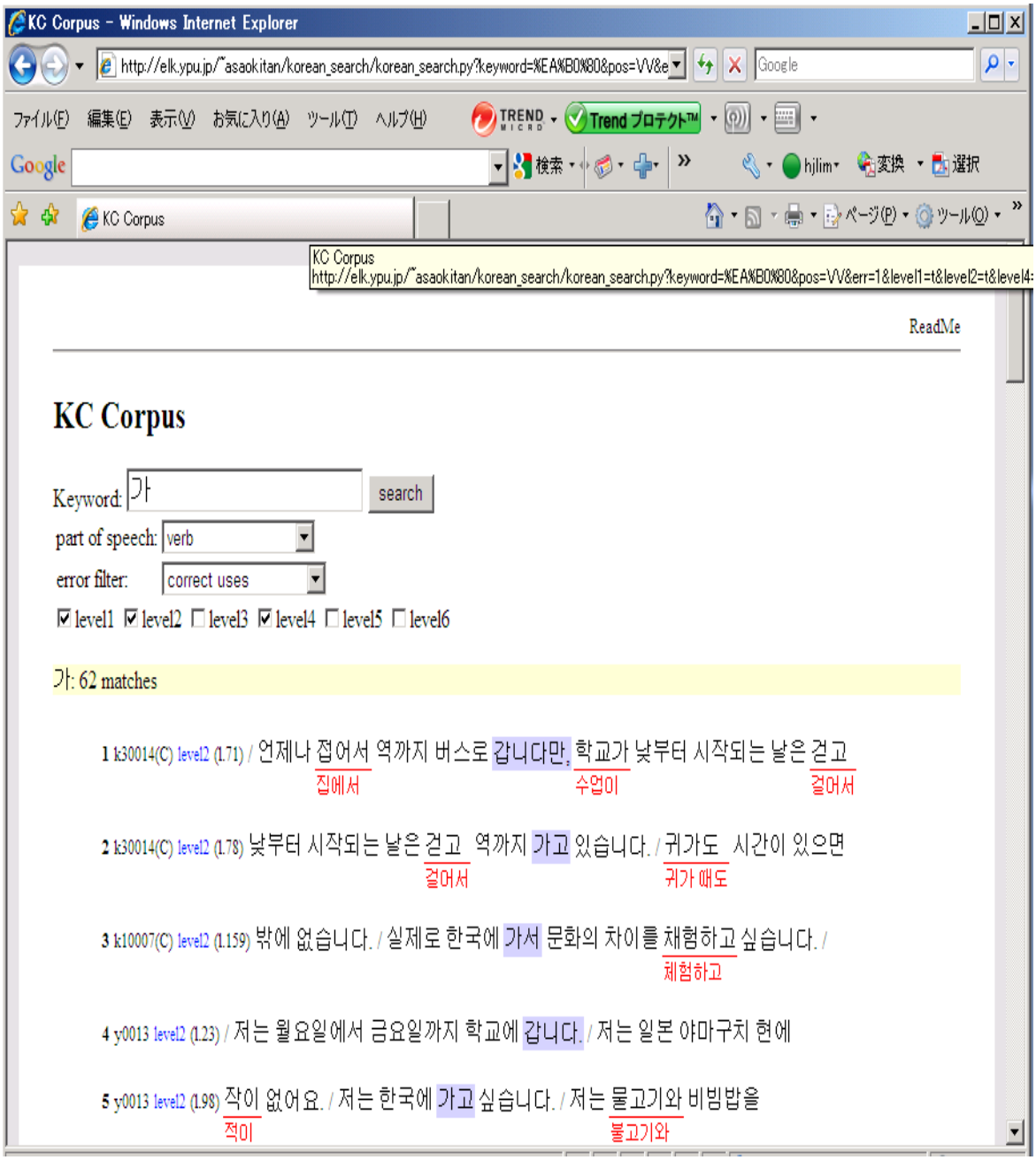
●[가]를 포함하는 전부의 문자열을 Kwic 형식으로 검색한 화면
(레벨 1-4:총 403 건)



●동사로서의 [가(다)]를 포함하는 전부의 문자열을 Kwic 형식으로 검색한 화면 (레벨 1-4:총 74 건)



●[가(다)]를 포함하는 문장 중에서 정문만을 Kwic 형식으로 검색한 화면 (레벨 1-4:총 62 건)



●[가(다)]를 포함하는 문장 중에서 학습자 오용만을 Kwic 형식으로 검색한 화면 (레벨 1-4:총 7 건)

아직 부족한 점도 많지만 오류 주석 시스템의 개발로 보다 쉬운 방법으로 오류에 대한 주석을 할 수 있게 되었으며, 학습자 오용의 패턴을 레벨별로 살펴볼 수 있게 되었다.

3-6. 배포

본연구는 한국학 중앙 연구원의 연구 조성에 의해서 기획·진행되었으며, 한국어 학습자의 데이터의 공유 자원화를 목적으로 하고 있기 때문에, 완전 프리로 공개할 예정이다. 이것에 대해서는, 데이터의 제공자 (한국어 학습자)에게는 데이터 수집의 단계에서부터, 본연구의 취지를 설명한 후, 공개에 관한 동의를 받고 있다. 그 때문에, 저작권에 관한 문제는 완전하게 클리어 되어 있다. 데이터의 제공에 있어서는, 기본 데이터와 웹상에서 KWIC 검색을 할 수 있는 시스템을 매개해 제공할 예정이다.

이상, 한국어 학습자 말뭉치 구축에 관한 최종 보고로서 본말뭉치 전체 설계 및 자료 수집, 그리고 검색 시스템에 대해 소개했다. 말뭉치를 이용한 습득 연구의 유효성이나 교육상의 활용의 중요성은 많은 선행 연구에 의해서 지적되어 왔다(예를 들면, 서(외), 2002; 石川 2008 등). 그러나, 한국어 관계의 학습자 말뭉치는 그 수가 적고, 한층 더 이용에 제한을 마련하고 있는 것이 대부분이다. 또, 일본에서 한국어 학습자가 산출한 데이터에 형태소 정보와 같은 언어 정보가 부여된 말뭉치나 그것을 간단하게 검색할 수 있는 시스템을 제공하고 있는 학습자 말뭉치는 전무라고 해도 과언이 아니다. 앞으로, 학습자의 오용 데이터를 이용한 말뭉치 규모를 한층 더 확충해 나가는 것을 통해서, 교사의 수업 개선이나 교수법의 지원은 물론, 학습자 사전의 작성, 커리큘럼의 개발, 한국어 학습자를 위한 교과서의 개발에 이용할 수 있다. 또한, 학습자의 언어 능력 테스트 개발의 기초 데이터로서도 활용할 수 있어 한국어 교육의 활성화에 기여할 수 있다고 생각한다.

향후 계획

향후, 다른 연구자와의 연계를 보다 깊게 하면서 한국어 학습자 말뭉치의 점진적인 확충과 실용성을 모색해 가고자 한다. 한국어학습자 작문 코퍼스(corpus)의 규모의 확충과 KC Corpus 의 검색 툴의 개량과 추가 기능의 검토 안은 다음과 같다.

①학회 발표등을 통해 타기관의 한국어교육 관계자들에 대해 KC Corpus 에 대한 이해와 협력을 요청한다.
②기존의 데이터의 오용 태그를 더욱 세분화하여 세밀한 검색이 가능하도록 한다.
③학습자의 오용에 관한 추적 분석을 가능하게 하여, 각 레벨의 학습자에게 적합한 학습 계획을 제시할 수 있도록 한다.
④검색 결과를 보존할 수 있는 옵션을 첨가하여, 데이터의 가공·분석을 할 수 있도록 한다.
⑤검색 결과를 일본어 원문 작문과 대비해서 볼 수 있는 기능을 검토하여, 대조언어학적인 관점에서 오용의 원인을 분석할 수 있도록 한다.

[참고 문헌]

[1] 고석주· 김미옥· 김재열· 서상규· 정희정· 한송화(2004), 한국어 학습자 말뭉치와 오류분석, 한국문화사.
[2] 서상규· 유현경· 남윤진(2002), 한국어 학습자 말뭉치와 한국어 교육, 한국어 교육 13-1, 국제한국어교육학회. 127-157.
[3] 淺尾仁彦· 李在鎬 (2008), 日本語学習者コーパス検索ツールの開発, 言語科学会第10回年次大会(静岡県立大学)大会論文集, p.182
[4] 石川慎一郎(2008) 『英語コーパスと言語教育』大修館書店
[5] 林炫情· 李在鎬· 曺美庚· 浅尾仁彦(2008), 韓国語学習者コーパス構築：韓国語学習者作文コーパスにおける検索ツールの開発, 信学技報(2008-40), 電子情報通信学会, 21-26.
[6] 李 在鎬· 淺尾 仁彦· 濱野 寛子· 佐野 香織· 井佐原 均(2008)「タグ付き日本語学習者コーパスの開発」, 2008 年度自然言語処理学会全国大会(東京大学)『大会論文集』, pp.658-661.

[7] 佐渡島沙織(2001), 作文コーパス研究における対訳の有効性：対訳が添削に及ぼす影響, 日本語教育のためのアジア諸言語の対訳作文データの収集とコーパスの構築, 国立国語研究所.

4. 서명(Signature)

Project Director 임현정