영어어휘교육용 전자사전용 외래어 변이형 데이터베이스의 구축

천승미
(서울디지털대학교)

Cheon, Seung-mi. 2010. Development of the Contents of an Educational Tool, EDUT. *The Linguistic Association of Korea Journal.* 18(1). 105-122. This paper aims to develop an educational tool for English as a second language learners in Korea, particularly with respect to vocabulary. This paper has examined the value of the ELK(Electronic Dictionary of English Loanwords in Korean) database as an educational tool for English language learners, particularly its capacity to motivate students who are already familiar with many English loanwords in Korean.

Key Words: English loanwords, vocabulary, ELK, database, corpus, EDUT, UNITEX, FSGraph, CALL, educational tool, dictionary

1. 서론

실제 코퍼스에서 사용되는 외래어를 살펴보기 위해, 무작위로 추출된 2009년 3월 25일 날짜의 아시아경제신문 기사[1]를 살펴보면, 310개의 단어 중 외국어와 외래어가 88개로 총 단어의 약 30%를 차지하고 있음을 볼 수 있다.

> 휴대폰은 이제 현대인에게 빼놓을 수 없는 필수품이 됐다. 때문인지 한해에도 수십가지의 휴대폰이 출시돼 소비자들의 선택을 기다린다. 이 휴대폰들의 가격이나 성능 또한 천차만별이다. 풀터치에 모션 센서까지 적용한 휴대폰이 있는 반면 오로지 통화 기능에만 충실한 휴대폰도 있다.... 풀터치의 WVGA(480×800)화면에 위젯기능, 위성DMB, 500만화소 카메라, DivX 동영상 플레이어, 풀브라우징 인터넷에다 문서업무까지 가능하며

1) http://www.asiae.co.kr/uhtml/read.jsp?idxno=2009032514455975065

> 모션 센서 기능이 있어 가로 세로 화면이 자동 전환되고 GPS까지 탑재하고 있다... 500만 화소 카메라는 물론 터치와 무빙 센서가 통합된 무빙 터치 기능까지 겸비해 69만 9600원이 책정됐다. 홀드 버튼이 상단 커버를 올리는 방식으로 바뀐 프레스토폰은 위젯 기능과 함께 터치방식 단축키로 음악을 편하게 들을수 있고 지상파DMB 기능을 갖추고 있다... LG전자 관계자는 "터치 기능에다 강화유리, 카본 파이버 등을 소재로 사용했기 때문에 가격대가 높아질 수밖에 없다. 풀터치폰이라고 무조건 비싼 것이 아니라 소재와 기능에 따라 달라지는 것"이라고 이유를 설명했다. (아시아경제신문, 고재완기자)

이러한 다양한 외래어의 한글음차표기가 사전검색시나, 한국어자동처리, 정보검색의 자동화에 있어서 문제점으로 나타나는데 고유명사나, 특히 전문용어의 경우, 적절한 한국어 번역을 발견할 수 없는 경우 해당 영어를 한국어로 음차표기해서 사용하거나, 정확한 의미전달을 위해서 원어 그대로 사용하기도 한다. 그래서 이미 들어와 자리 잡은 영어중심의 외래어와 곧 외래어로 자리 잡을 듯한 영어의 사용 가운데, 서로 통하지 않는 표현들이 많아서 각각의 외래어에 해당하는 올바른 영어표현을 제시해 주는 것이 필요하다. 더구나 원어와 외래어 사이의 간격이 점점 넓어지고 있어서 영어 철자와 발음에 세심한 주의를 기울여하는 어휘들이 증가하고 있고(정국, 1984), 영어의 의미가 크게 변해서 사용되고 있거나, 영어에 없는 한국식 또는 일본식 영어들이 많아서 영어 어휘학습에 장애가 되고 있다. 이런 이유에서 원어와 한글외래어표기에 대한 효과적인 처리의 필요성이 증대되고 있다. 즉, 외래어 중 고유명사, 전문용어와 더불어 여러 가지 음차가능한 표기와 함께 원어가 제시되어져 있는 외래어 사전의 구축이 시급하다.

2. 연구의 목적 및 필요성

외래어와 외국어의 차이는 소속하는 언어체계와의 사이에 나타나는 동화의 정도가 사회적 승인의 유무로써 결정되는데, 그 차이는 사용자의 외국어에 대한 의식의 유무에 의해 결정되는 것으로 다분히 주관적인 성격을 띤다고 할 수 있다. 외국과의 교류가 증대되면서 새로운 지시대상이나 개념이 필요하게 되고 이러한 것들에 대한 언어적 명명이 자국어가 가진 어휘만으로는 충분하지 못하여 새로운 어휘가 필요함에 따라 방대한 양의 외래어 또는 외국어의 수용이 불가피하다.

외국어가 도입되는 경로를 살펴보면, 크게 3가지로 나눌 수 있는데, 우선 같은 알파벳 체계를 가진 언어에서는 원어표기 그대로 사용되는 경우가 있는가하면, 언어에 따라서는 새로운 용어가 나타나도 그대로 차용해서 쓰지 않고, 모국어를 이용한 새용어를 만들어 사용하기도 하고, 자국어에 있는 단어로 바꿀 수 없는 경우 외국어를 자국어로 음차표기하여 사용하

는 경우가 있는데, 한국어의 경우에는 외국어의 원어나 번역된 표현보다는 한글로 전사되어 쓰여지는 비중이 큰 만큼 이 문제는 중요하게 다루어져야 한다. 이러한 외래어의 한글음차표기 상황을 살펴보면, 영어발음 자체가 일관성이 없을 뿐더러, 한국식 발음에 의한 한국식 영어에 의존하여 음차를 하는 경우도 있고, 영어에 익숙한 사람은 그렇지 않은 사람보다 원어의 발음에 가깝게 음차표기하려는 경향이 있다(정국, 1986, 1988). 즉, 외래어 음차표기시 한국식 문법과 한국식 발음표기, 한국인들의 사고방식과 한국말을 그대로 전이시키려는 데다가, 음운체계와 음절조직이 한국어와 크게 다른 외국어의 속성상 사람마다 표기를 다르게 하기 쉽다. 실례로, 표 1에서 보이듯이, 정규표현식에 매칭되는 문자열을 찾아서 해당 라인만을 출력하는 hgrep 프로그램[2]을 이용하여 무작위로 추출된 문서에서 사용된 외래어를 분석해본 결과, 하나의 외래어 단어가 다양한 음차표기로 사용되는 경우가 다양하게 나타났다.

표 1. 외래어 단어의 다양한 음차표기[3]

원어	한글음차표기	사용빈도
gauze	거즈	66%
	가제	34%
napkin	냅킨	75%
	내프킨	25%
clarinet	클라리넷	46%
	클래리넷	54%
dollar	달러	89%
	달라	11%
dance	댄스	91%
	땐스	9%
double	더블	78%
	떠블	22%
debut	데뷔	59%
	데뷰	41%
badge	배지	73%
	뱃지	11%
	빳지	16%

그 밖에, "linger"에 해당되는 음차표기로 "링거(21%)," "링게르(14%)," "링겔(51%)" 등, 세 가지 이상의 음차표기가 동시에 사용이 되고 있고, "net"도 "네트(47%)," "넷

2) http://acme.com/software/hgrep/

3) 추출된 대상문서는 21세기 세종계획(http://www.sejong.or.kr, 국립국어원 http://www.korean.go.kr)에서 무료로 배포하고 있는 말뭉치 자료중 기구축 말뭉치 750만절을 기초로 분석된 자료이다.

(29%)," "넷트(18%)" 등 여러 가지 음차표기가 공존함을 볼 수 있다. 이러한 다양한 외래어의 한글음차표기가 사전검색시나, 한국어자동처리, 정보검색의 자동화에 있어서 문제점으로 나타나는데 결국 같은 문서 내에서 영어 및 다양한 외래어 표기의 혼용은 정보검색에서 심각한 단어불일치문제를 야기하며 정보검색의 성능을 크게 저하시키는 역할을 하게 된다. 즉 같은 개념을 표현하는 용어이지만 질의와 문서에 사용된 용어가 다를 경우 단순한 패턴매칭방법으로는 검색이 되지 않게 되는 것이다.

이렇게 한 단어의 외래어가 여러 가지 음차표기형으로 사용되고 있고, 학자들도 주관적이고 직관에 의한 기준으로 인하여 옳은 외래어 표기어형에 대한 판단은 어려운 실정에서, 외래어 표기 논란에 관한 해결 방안으로 1991년 정부와 언론이 외래어 심의 공동 위원회를 만들고 국립국어연구원에서 외래어 표기를 심의하여 표준 외래어 표기법을 만들어 각 언어마다의 표기 세칙을 둔 상황이지만, 표기법 자체가 충분히 자세하지 않아서 모든 외래어를 적는 데는 여전히 어려움이 있고, 보통사람이 이 표기법을 정확히 이해하고 실천하는 것은 사실상 어려운 일이다(정국, 2002, 2003). 또한 표준외래어표기법에 예외사항이 많아 오히려 이 표기법 자체가 다양한 외래어표기를 만들어 내는 원인이 되고 있다.

이에 본 연구는 한국어 어휘체계에서 사용되고 있는 영어 외래어에 대한 한국어 음차표기 유형들에 대해서, 이에 대응되는 영어 원어 표기들을 찾아내고, 각 원어 표기에 대응 가능한 모든 전사표기 변이형의 목록을 구축하여 이 데이터베이스로 활용한 영어 어휘 교육용 프로그램을 개발하는 것을 목적으로 한다.

3. 영어외래어를 이용한 영어어휘교육

스키마이론에 의하면 지식은 독립적인 상태로 각각 존재하는 것이 아니라 일련의 관계로 구조화되어 있어서, 학습자가 새로운 정보를 이미 알고 있는 정보에 연관지어 이해하려 하고, 이 작용이 잘 이루어질수록 학습효과가 뛰어나다고 한다. 영어 학습량의 증가와 더불어 요즘은 영어 어휘의 상당수가 외래어로 사용되고 있는 이유로 자연스레 외래어의 사용빈도가 높아지고 있으므로 학습자들은 이들 외래어에 대한 분명한 의미와 원어는 모를 수 있지만 여러 매체를 통해서 눈과 귀에 익숙한 것들이 많다. 영어어휘교육시 그 어휘의 여러 가지 의미와 사용에 대해 언급한다면, 학습자에게 익숙한 해당 어휘의 인지는 한층 쉬울 것이고 외래어로 사용되는 어휘의 의미에 대해서 학습자가 높은 인지를 나타내므로, 이것은 영어어휘학습에 긍정적인 전이요인이 될 수 있다.

이는 외국어 어휘교육에 있어서 외래어의 선택이 대단히 중요한 의미를 지님을 보여주는 실례이다. 어휘 교육시 영어 외래어와 문자를 접목시켜 경험하게 함으로써 정규 영어수업에 대해 긍정적인 태도를 갖고 주변에 있는 영어 문자에 대한 친근감과 식별 능력을 높일 수 있

는 것이다. 그러나 잘못 수용되어 사용되는 외래어는 그 어휘의 언어를 학습하는데 오히려 커다란 간섭 요인이 된다. 사실 언어학습자가 언어를 이해하고 쓰는데서 생기는 어려움은 어휘지식이 정확하지 않은데서 생기는 것이다. 이러한 간섭 현상은 학습량이 증가하고 성취도가 높아져도 쉽게 제거되지 않는다.

대다수의 외래어가 실제 영어 발음과는 상당한 차이가 있고(정국, 1988a, 1988b) 외래어로 사용되고 있는 단어의 뜻도 원래의 뜻과는 상반된 경우가 많아 별도의 지도가 필요하며 외래어를 통해 영어 학습을 할 때에는 파닉스[4]를 통해 원 발음을 충분히 익히는 것이 중요하다. 즉, 외래어의 의미는 물론이거니와 문법 발음상 학습자에게 혼동의 가능성을 내포한 어휘에 대해서는 그 어휘가 제시될 때마다 반복적이고 철저한 지도가 필요한 것이다. 또한 영어문법이 한국어문법에 동화된 한국식 영어발생을 최소화하고 바른 영어 사용을 위해서라도, 올바른 영어식 표현을 교육현장에서 제시하여야 하며, 잘못 형성된 외래어와 한국식 영어를 가급적 최소화 할 수 있도록 외래어의 정확한 표현과 관련하여 목록이 작성되어져야 한다. 실례로, 호주 Monash Univ.의 한국어교육과에서 한국어 어휘교육을 할 때 학습자들에게 이미 익숙한 차용어를 이용하여 한국어 어휘교육을 하고 있음을 확인할 수 있다[5].

Crothers와 Suppes(1967)도 차용어를 이용한 어휘학습에 관한 연구에서, 학습자들이 백여개 정도의 단어짝을 완전히 습득하는데 7회 반복으로 충분했으며, 국제적 차용어는 암기에 절대적으로 용이하여 학습자들의 어휘량을 가장 쉽게 늘일 수 있는 효과도 있다고 하였다. 여기서 어휘습득에 있어서 7회 반복으로 어휘의 완전한 습득이 이루어졌다고 하였는데 일반적으로 새로운 어휘나 표현의 완전한 습득을 위해서 최소 50번의 반복이 이루어져야 한다는 학자들의 주장을 미루어보면 경이적인 숫자임에 틀림이 없다. 이는 외국어 교육시 학습자의 모국어에 이미 도입된 외래어를 사용하는 것이 효율적임을 보여주는 중요한 예라고 할 수 있다. Carter(1998)도 외국어 단어와 모국어 단어 간의 형태적 전이의 기회가 많으면 많을수록 기억의 기회가 증가한다고 주장하며, "텔레비전"이나 "라디오" 같은 국제적인 차용어인 경우에는 기억이 더 용이하다고 하였다.

4. 외래어 사전의 구축

4.1 외래어데이터베이스의 구축

기존의 연구 논문들(이재성 & 최기선(1997), 이재성 (1999), 오종훈 (2000)등)에서는 이

4) 파닉스(phonics)는 영어알파벳이 가지는 소리를 익혀 영어의 자음, 모음, 연속자음, 자음이중자 등의 소리를 먼저 익혀 영어단어를 보고 발음하는 발음중심의 어학교수법을 뜻한다.

5) http://www.arts.monash.edu.au/korean/klec/index.php

미 구축된 말뭉치를 중심으로 hgrep 프로그램과 같은 다양한 검색프로그램을 이용하여 빠른 검색으로 이루어진 데이터베이스를 사용하고 있으나, 본 연구에서는 사용된 데이터베이스는 총 4개의 사전에서 수작업으로 추출된 외래어가 사용이 되었다.

먼저 외래어 일반명사 어휘 데이터인 "한국어 단순명사 사전(Nam, 1994)"에서 외래어 어휘 부분만을 추출하고, 이를 국립국어연구원에서 사전과 코퍼스를 중심으로 추출한 외래어 어휘 데이터와 통합하였다. 그리고 외래어 고유명사의 어휘 데이터로는 국립국어연구원에서 추출한 외래어 고유명사 데이터와 "한국어 고유명사 전자사전(Nam, 2003)"에서 추출한 데이터가 사용되었다 (Cheon, 2005). 이 때, 표제어에서 중복된 데이터를 가려내고 외래어의 표기 특성상 나타나는 표제어 자체의 변이형이 많아서 이를 가려내는 작업이 필요하였는데, 데이터의 모든 변이형이 동일한 중요성을 지니나, 사전으로서의 형태를 갖추기 위해 데이터의 여러 가지 음차 표기 변이형 중에서 사전에 등재되어 표제어(색인)로 사용될 수 있는, 현재 표준어를 사용하고 있는 집단 14명을 대상으로 조사하여 가장 일반적이라고 판단되어지는 표기 형태를 대표형으로 설정하였다.

다음 작업으로 각각의 데이터에 해당되는 원어를 찾아서 넣어주는 작업을 하였는데, 이는 외래어 어휘의 데이터 분석에서 가장 힘든 작업이었다. 외래어 연구시 가장 어려운 점의 하나가 외국어에서 들어오지 않은 "외래어"를 가려내는 작업이다. 또한, 앞서 확인하였듯이, 실제 텍스트를 살펴보면 기존의 자료 및 사전에서 발견된 외래어들과는 다른 형태의 음차표기들이 발견된다. 기존의 사전에서의 표제어 목록은 편찬자의 직관이나 개인적인 판단에 의하여 선정된 것이 대부분이라고 할 수 있어서, 누락된 외래어 데이터의 수집을 위해서 코퍼스를 이용한 데이터의 수집을 하고, 코퍼스를 바탕으로 한 빈도에 입각한 표제어의 선정과 기존사전의 분석 작업이 동시에 이루어져야만 올바른 어휘목록을 만들어낼 수가 있다.

즉, 사전편찬을 위한 코퍼스의 수집에서 반드시 고려해야 할 것은 코퍼스의 양과 질에 관련된 것이다. 본 연구와 같은 기초적 모색과 실험단계에서는 모형 코퍼스의 크기가 작은 규모로 제한될 수밖에 없는 만큼 이에 적합한 영역 및 자료의 선택이 불가피하다. 이를 고려하여 본 연구에서는 우선 최신의 용어가 포함될 수 있도록 최신의 경제(증권), IT 분야의 신문을 1차적 구현을 위한 샘플 채취 대상으로 설정하고, 본격적인 코퍼스 구축작업을 시작하기 전에, 외래어가 나타나는 비중을 살펴보기 위해서 표본작업을 실시하였다. "조선일보" 2003년 06월 26일자 신문기사를 표본으로 분석하였더니 외래어가 가장 많이 나타나는 기사는 스포츠 면이고 비교적 적게 나타나는 분야가 정치면이었다.

다음 작업으로 "조선일보"와 "중앙일보"의 다른 기사들에 비해 외래어로 된 전문용어가 비교적 많이 나오는 경제면만 살펴보기로 하고, 2003년 6월 26일자 신문부터 2003년 8월 4일자 신문까지 40일치의 조선일보와 중앙일보의 각 경제면 기사 44개를 추출하여 수작업으로 분석하였다. 총 10,740개의 어절 중, 이에서 나온 외래어 데이터가 포함된 어절의 수는 844개로, 외래어의 비중이 약 8%에 해당함을 알 수 있다. 그리고 "중앙일보" 경제면에서만

7월 30일과 31일에 나온 기사 각각 94개, 92개의 기사를 추출하여 살펴보았더니, 7월 30일자 94개 기사에서는 외래어 데이터가 1,412개가 나왔다. 7월 30일 기사의 총 낱말 수는 17,779개이고 그중 외래어데이터수가 1,412개이므로 약 8%가 외래어인 셈이다. 7월 31일자 92개의 기사인 경우에는 총 16,922개의 낱말 중 외래어는 1,408개이고 외래어의 비중은 약 9%로 그다지 차이가 나지 않았다.

마지막으로 2007년도의 연합뉴스들을 인터넷에서 검색하여 코퍼스를 구성하였다. 이때, 경제면에 해당하는 부분에서 주식과 IT 관련 712개의 기사를 선정하여 외래어 어휘에 관한 분석을 한 결과, 12,239개의 외래어 어휘데이터가 추출되었다.

코퍼스에서 추출한 데이터 중 앞서 구축한 외래어 데이터베이스에 등재되어 있지 않은 어휘수는 3,211개로, 코퍼스에서 구축한 데이터베이스의 약 73%에 해당하는 것으로, 상당수가 기존 사전 자료에서는 누락되어 있는 것을 발견할 수 있었다. 이러한 사실은 데이터베이스 구축시 코퍼스 사용의 중요성을 일깨워 준다.

4.2 외래어 변이형 데이터베이스의 구축

앞에서 영어 원어가 한국어 텍스트에 도입되어 사용될 때 많은 경우 한국어 전사 표기 형태로 사용되며, 이때 단일 원어에 대하여 여러 가지의 음차 표기 형태가 실현되는 현상을 확인할 수 있었다. 즉, 이상에서 구축된 대표형 기반 데이터베이스는 모든 가능한 변이형들을 포함하는 형태로 확장되어야 진정한 의미의 언어 자원으로서 기능할 수 있게 된다. 이를 위하여 가능한 모든 음차 변이들을 자동으로 가정하여 생성하는 방식(이재성 & 최기선(1997), 이재성(1999), 정길순(1998) 외)과 실제로 발견될 수 있는 변이형들의 목록을 개별적으로 구축하는 방식이 가정될 수 있다. 그러나 전자와 같은 방식으로 추정하기에는 실제 음차 표기들이 이처럼 체계적이지 않고, 전혀 불필요한 형태들을 포함하게 됨으로써 언어 자원으로서의 신뢰성을 저하시키는 면이 나타나므로 본 연구에서는 각 대표형에 대하여 개별적인 변이형들을 고려해야 한다는 입장을 취하였다. 이러한 관점은 근본적으로 언어 현상을 바라보는 부분문법의 입장과 상응한다.

부분문법 이론은 자연언어 텍스트에 대한 자동처리를 위해 어휘문법이론과 함께 고안된 문법 모델이다(남지순, 2005). 어휘문법이론이 통사-어휘적으로 자유로운 단문 구조에 대한 체계적인 연구라고 한다면, 부분문법이론은 부분적으로 굳어진 관용 표현을 함유하고 있는 일정 시퀀스나 의미적으로 동의관계를 가진 일정 어구들, 또는 특정 분야에 관련된 언어 표현들 및 그 외 부분적으로 통사-어휘-의미적으로 제약 관계를 보이는 모든 언어 형태들에 대한 부분적인 문법을 구성하는 것이 부분문법의 기능이라 할 수 있다[6]. 부분문법으로 구현된 언어 정보들은 텍스트에 대한 자동 처리시 분석이나 생성을 위한 문법적 정보로 사용될 수

6) http://maincc.hufs.ac.kr/~namjs

있으며, 분석시 발생하는 중의성을 제어하는 데에도 매우 효율적으로 기능할 수 있다. 부분문법은 비순환그래프의 형태로 실현될 수 있으며, 이는 텍스트 처리시 유한 트랜스듀서나 유한 오토마타의 형태로 바로 전환되어 시스템에서 작동할 수 있다. 프랑스 LADL/IGM 연구소에서 개발한 INTEX 프로그램과 UNITEX 프로그램[7]은 바로 이와 같은 작업을 용이하게 하는 그래픽에디터를 내장하고 있다.

ELK(Electronic Dictionary of English Loanwords in Korean)라고 이름붙인 본인의 외래어 전자사전은 앞서 획득된 외래어 대표형 목록에, 가능한 모든 변이형들을 기술하여 하나의 원어표현에 대한 모든 한국어 음차 표기들을 연결시켜주는 정보를 내장한 사전으로, 바로 이와 같은 그래픽에디터를 이용하여 다양한 전사 변이형을 가지는 외래어 데이터들을 하나로 그래프로 표현할 수 있었다.

이처럼 외래어 어휘 중 전사변이를 허용하는 형태들은 UNITEX 프로그램[8]의 FSTs 그래프를 이용하여 기술함으로써, 리스트 방식으로 일일이 열거하는 어려움과 혼동을 피할 수 있었으며, 음차표기 가능성들을 조합적으로 결합해 봄으로써 보다 논리적으로 접근할 수 있는 장점을 가질 수 있었다. 이와 같이 구축된 그래프 전체를 통합하여 나온 변이형들 목록의 전체가 궁극적으로 외래어전자사전인 ELK(Electronic Dictionary of English Loanwords in Korean)의 데이터베이스를 구성하게 된다.

4.3 FST 그래프에디터를 이용한 외래어전자사전의 구축

기존의 연구들과는 달리 본 연구에서는 영어 외래어 어휘의 음차규칙을 새롭게 설정하여 그에 따른 한 가지 또는 예외의 표기법을 정하는 것이 아니라, 영어 외래어 어휘들이 변이형, 즉 자국어로 음차 표기시 나타나는 여러 가지 형태의 다양한 변이형들을 일일이, 가능한 표기는 모두 변이형으로 넣어준 뒤, 여러 가지 표기들을 하나의 그래프와 같은 형식으로 묶어서 색인해 두었다가 사용하는 것이다. 이렇게 구축된 데이터베이스를 어휘교육 프로그램의 엔진으로 활용하고, 이를 이용하여 해당 외래어 어휘의 정확한 원어를 찾아주는 것에 중점을

---

7) UNITEX 프로그램은 프랑스 파리 제7대학(University of Paris 7)의 LADL연구소에서 개발된 INTEX 프로그램에 상응하는 S/W로서, 마른느-라-발레 대학(University of Marne-la-Vallee)의 IGM연구소에서 개발되었다. UNITEX 프로그램은 기본적으로 자동 처리를 위한 전자사전(electronic lexicon)을 구축하는 사전 에디터의 기능을 가지고 있다. 부분문법을 직접 구성하기 위한 그래픽 에디터도 내장하고 있으며, 이와 같이 사용자가 구축한 사전들은 프로그램에 내장되어 있는 DELA 전자사전과 함께 사용되어 텍스트에 대한 직접적인 분석을 가능하게 한다. 텍스트에 대한 형태소분석 및 구문분석, 그리고 중의성 제어 결과 등이 제공되며, 사용자가 요구하는 일정 패턴에 대한 용례추출기(concordance program)의 기능도 가지고 있다. UNITEX는 기본적으로 다국어 처리 프로그램으로서, 사용자가 지정하는 언어 체계에 따라 사전 및 문법 정보를 구성하는 언어 정보 구축기의 기능과 함께, 축적된 언어 정보에 따라 텍스트를 실제 처리하여 그 결과를 제공하는 언어 텍스트 분석기의 기능을 동시에 가진다.

8) http://www-igm.univ-mlv.fr/~unitex/

두었다.

즉, 영어 외래어 어휘를 규범적인 단어의 형태(canonical form)로 만들어서 활용하는 것이 아니라, 다양하게 나타날 수 있는 데이터를 충분하게 넣어서 영어의 원어(origin word)에 해당하는 단어로 연결시켜주는 것이다.

아래 그림 1은 UNITEX 프로그램의 FST 그래프를 이용하여 "초콜렛(chocolate)"의 모든 음차표기를 하나의 그래프로 나타낸 예이다. 이 그래프는 리스트로의 변환이 가능한데, 표 2의 형태로 총 1,681가지 음차표기형의 데이터로 태깅된다.

각각의 한글 음차 표기 데이터는 "초코렛트, 초콜렛. NFS (chocolate) CON; NC" 즉, "음차변이형, 대표형, NFS:명사(Noun)+외래어(Foreign)+단순명사(Simple Noun), (올바른 영어표현), 구체명사(Concrete noun), 자음으로 끝나는 명사(Noun ended with consonant)"의 정보를 내장하게 된다.

그림 1. FST 그래프 (초콜렛 (chocolate))

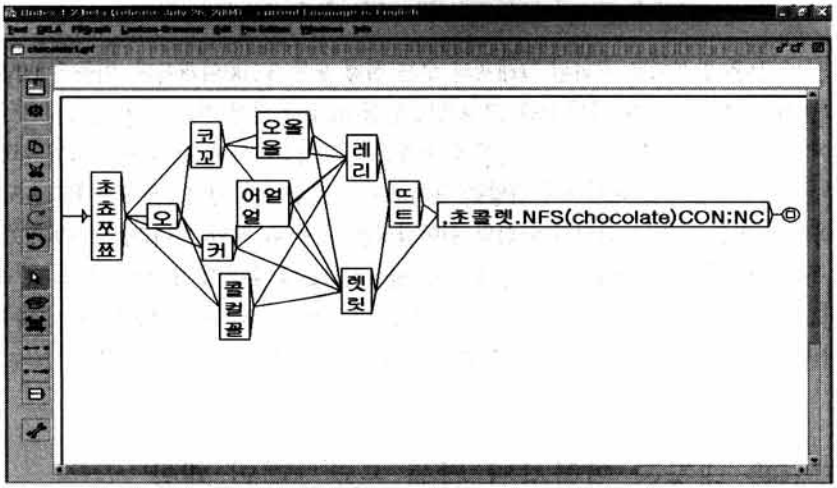


표 2. FST 그래프(초콜렛(chocolate))의 태깅 예

TAG
초콜릿,초콜렛.NFS(chocolate)CON;NC
초콜릿트,초콜렛.NFS(chocolate)CON;NC
초콜릿뜨,초콜렛.NFS(chocolate)CON;NC
초콜렛,초콜렛.NFS(chocolate)CON;NC
초콜렛트,초콜렛.NFS(chocolate)CON;NC
초콜렛뜨,초콜렛.NFS(chocolate)CON;NC

초콜리트,초콜렛.NFS(chocolate)CON;NC
초콜리뜨,초콜렛.NFS(chocolate)CON;NC
초콜레트,초콜렛.NFS(chocolate)CON;NC
초콜레뜨,초콜렛.NFS(chocolate)CON;NC
초오꼬올릿,초콜렛.NFS(chocolate)CON;NC
초오꼬올렛,초콜렛.NFS(chocolate)CON;NC
초오꼬올레뜨,초콜렛.NFS(chocolate)CON;NC
초오꼬오올릿,초콜렛.NFS(chocolate)CON;NC
초오꼬오올렛,초콜렛.NFS(chocolate)CON;NC
초오꼬오올리뜨,초콜렛.NFS(chocolate)CON;NC
초오코올릿,초콜렛.NFS(chocolate)CON;NC
초오코올릿트,초콜렛.NFS(chocolate)CON;NC
...

이상과 같은 방식으로 구현된 그래프의 수는 현재 모두 3,100여개이며, 이들은 UNITEX 프로그램에서 자동으로 한 개의 FST 그래프로 통합(merge)되어 여러 응용 분야에서 효율적인 언어 자원으로 활용될 수 있다. 이렇게 구축된 외래어전자사전 ELK는 외래어 컨텐츠의 핵심부분인 외래어 어휘 데이터의 다양한 음차표기를 포함하고 있어, 영어 어휘학습시 한국어로 음차 표기되어 있는 외래어 어휘의 원어검색을 쉽게 할 뿐만 아니라, 기계번역기나 인터넷, 문서 검색과 같은 정보검색의 효율성을 향상시키는데 활용할 수 있을 것이다. 구축된 외래어전자사전은 모든 변이형을 인식할 수 있게 고안되어 있어 실제 텍스트[9]에서 출현하는 외래어의 90% 이상을 올바르게 인식하고 처리할 수 있음을 확인할 수 있었다.

5. ELK를 내장한 영한사전 검색 프로그램의 개발

본 연구의 궁극적인 목표인 영어어휘교육에 사용할 수 있는 교육용 프로그램을 개발하기 위해서 앞서 구축한 외래어 전자사전인 ELK(Electronic Dictionary of English Loanwords in Korean)를 내장하여 교육용 틀인 EDUT[10](Database Look-up & Linking On-line Dictionary)을 개발하였다. 해당 프로그램은 포털사이트 야후(Yahoo)의 인터넷 영한전자사전과 연계되도록 개발되어, 검색한 단어의 결과물이 야후 영한사전의 검색

9) 대상문서는 21세기 세종계획(http://www.sejong.or.kr, 국립국어원http://www.korean.go.kr)에서 무료로 배포하고 있는 말뭉치 자료와 인터넷 연합뉴스에서 무작위로 추출하여 사용하였다.
10) 본 어휘교육용 프로그램 개발에 프랑스인 프로그래머 Ivan Berlocher의 도움을 받았다.

결과물과 일치한다.

이 툴은 본 연구에서 구축된 데이터베이스의 성능을 확인하는 데에도 의미가 있을 뿐 아니라, 나아가 이와 같은 언어 자원을 이용한 교육용 툴의 개발의 한 예를 보여주는 데에도 의미가 있다고 생각된다.

이 프로그램(EDUT)[11]은 사용자가 한글로 어떠한 형태의 외래어 전사 표기를 입력하여도 이에 대응되는 영어 원어를 인식하여 이를 표제어로 하는 영한 사전의 항목을 제시해 주는 기능을 갖도록 고안되었다.

프로그래밍 언어 C++로 구현된 본 어휘 교육용 프로그램 EDUT의 개발 과정을 소개하면 다음 표 3과 같다.

표 3. EDUT 알고리즘

알고리즘	설명
1. Procedure READ_INPUT_STRING	사용자가 스크린 화면에 찾고자 하는 영어외래어의 전사된 표현을 입력한다.
2. Procedure FIND_EXPRESSION	외래어전자사전에 내장되어 있는 외래어데이터베이스의 변이형들 중 일치하는 데이터베이스를 검색한다.
3. Procedure FIND_WORD	검색한 데이터베이스의 영어원어를 검색하여 가져온다.
4. Procedure TYPE_WORD	인터넷으로 연결되어 있는 영한사전(야후)에 영어원어를 검색한 영어원어를 입력한다.
5. Procedure FIND_TERM	외래어전자사전의 스크린 화면에 사용자가 검색한 외래어 단어의 철자, 발음, 의미, 관련된 표현들까지 나타내어준다.

그림 2는 구축된 외래어전자사전 EDUT의 메인 화면을 캡쳐한 것으로, (1)이라고 표시된 부분에 사용자가 검색하고자 하는 외래어의 한글전사표기를 입력하고, (2)라고 표시된 부

11) EDUT 프로그램 설명

- <Key-Value> 쌍으로 구성된 리스트를 해쉬맵(hashmap) 데이터구조로 저장한다. 해쉬맵 데이터구조는 C++나 Java와 같은 프로그래밍 언어에서 다음과 같은 연산을 가능하게 하는 기본 구조이기 때문이다.
- 각 키(key)에 따라 값(Value)을 획득하는 것이 가능(n개의 엔트리가 있는 사전이라면 복잡도는 C=O(log N)이 된다.
- 또한 <키-값>의 쌍을 추가하는 것이 가능(이때의 복잡도도 위와 같이 C=O(log N)이 된다.)하다.

분을 클릭하면, (3)에서 검색한 결과를 표시하여 준다.

그림 2. EDUT 프로그램의 메인화면

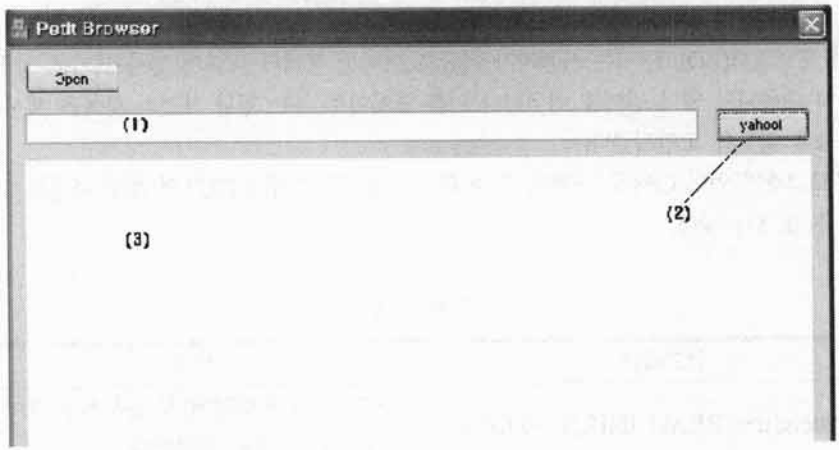


그림 3, 4, 5는 개발된 EDUT 프로그램에서 "chocolate"의 서로 다른 음차표기형인 "쪼꼬렛", "초꼬렛", "초코렛뜨"를 각각 키워드로 입력한 경우, 결과물로 제시된 영한사전 항목으로써, 모두 올바른 영어 원어를 인식하여 동일한 결과를 제시하고 있음을 볼 수 있다.

그림 3. 한글음차표기 "쪼꼬렛" 입력시의 검색 결과 화면

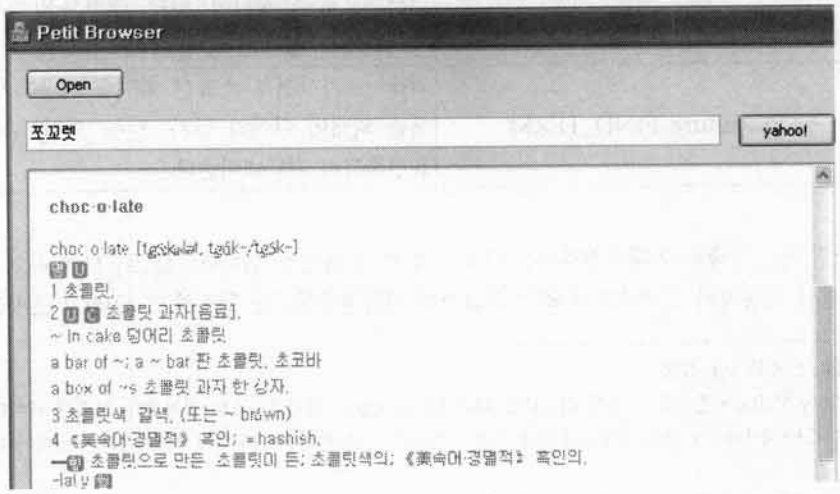


그림 4. 한글음차표기 "초꼬렛" 입력시의 검색 결과 화면

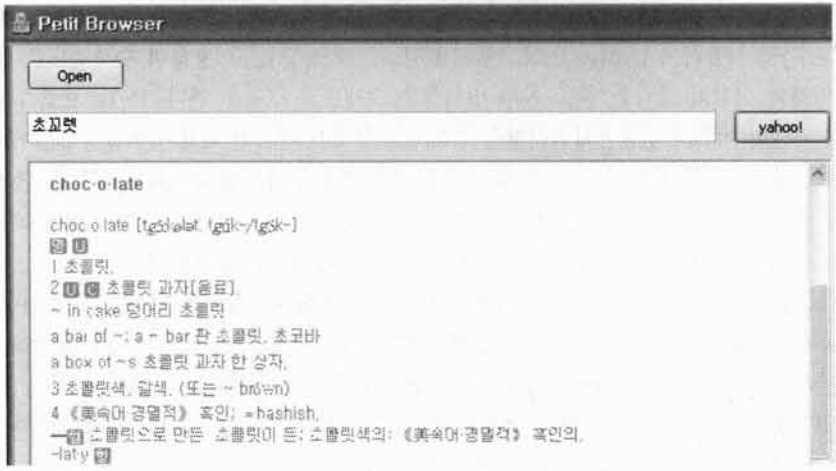
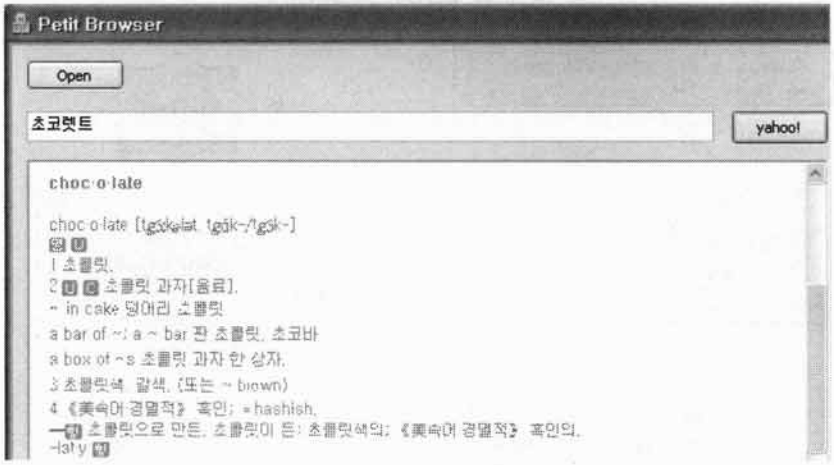


그림 5. 한글음차표기 "초꼬렛트" 입력시의 검색 결과 화면



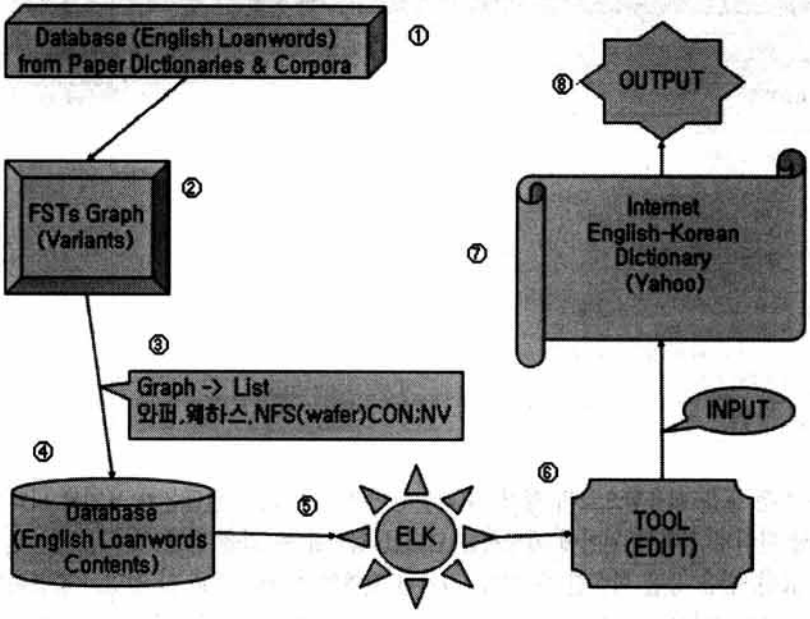
이와 같은 틀을 이용함으로써 영어 외래어 단어가 여러 가지 한글음차 표기로 나타나 있어도, 해당 단어의 정확한 원어에 해당하는 단어를 검색할 수 있을 뿐만 아니라, 검색된 단어의 의미, 표현, 발음 등을 학습할 수 있다. 검색된 결과의 정보나 내용 및 설명이 충분하지 않은 경우는 검색되어진 정확한 영어철자를 가지고 다른 종류의 사전에서 검색할 수 있어 더 많은 정보를 획득할 수 있다.

실례로, 신문에서 무작위로 발췌한 기사에서 나온 외래어들의 검색을 해 본 결과, EDUT에서 90% 이상의 성공적인 검색이 이루어짐을 확인할 수 있었다.

본 연구에서 구현된 ELK와 이를 이용한 영한사전 검색 프로그램 EDUT은 그림 6과 같이 간단히 기술될 수 있다. 먼저 여러 가지 유형의 사전과 코퍼스에서 추출된 외래어 데이터베이스(①)를 사용하여 Unitex 프로그램의 FSTs 그래프(②)를 이용하여 수작업으로 일일이 해당 외래어 어휘의 다양한 한글 음차 변이형을 그래프로 구축을 한다. FSTs 프로그램으로 각각의 외래어 어휘의 한글음차변이형을 그래프로 구축시 각각의 외래어에 해당하는 영어 원어(③)를 찾아 입력함으로써 영어 어휘 교육용 프로그램에 사용될 수 있는 가장 중요한 자원(④)이 구축되어 진다. 이 데이터베이스(⑤)를 활용할 수 있는 간단한 전자사전(⑥)의 검색 형태 도구를 만들어 앞서 구축한 데이터베이스 ⑤를 콘텐츠로 활용한다.

이 도구는 학습자가 인터넷에서 쉽고 간단하게 사용할 수 있도록 인터넷 포털사이트 야후(www.yahoo.co.kr)에서 무료로 제공하고 있는 영한사전과 연계가 되도록 개발을 하여, 검색한 외래어가 야후 영한사전(⑦)에서 검색한 결과물과 일치한다. 즉, 해당 외래어 어휘가 야후 인터넷 사전에서 해당 어휘를 검색할 수 있도록 하는 도구를 가지고 해당 외래어 어휘의 정확한 영어 원어를 검색(⑧)할 수 있도록 한다.

그림 6. EDUT 프로그램



6. 향후 연구방향

본 연구에서 가장 중요한 자원이라고 할 수 있는 영어 원어 정보가 들어가 있는 외래어 데이터베이스의 수는 5천개 이상이며, 영어 외래어의 특성상 나타나는 다양한 한글음차표기 형태까지 고려한다면, 본 연구에서 개발한 도구가 실제적으로 검색할 수 있는 어휘수는 20만 개 이상이다. 실제로 신문과 문어 및 구어 코퍼스[12]에서 나타나는 2가지 이상의 표기 형태를 가진 외래어를 중심으로 본 연구에서 개발된 전자사전에서 검색을 한 결과, 98% 이상의 성공률을 나타내고 있다.

그러나 ELK와 EDUT 프로그램이 실제 학생들에 의해 시범 사용되어지지 않은 상태라서 직접 사용해본 학생들을 대상으로한 설문조사라던가 학생들의 평가 정보가 요구되어진다. 또한 지속적인 외래어 데이터베이스의 업데이트가 이루어져야 하는데, 이와 병행하여 자동으로 전사표기를 추정할 수 있는 자동프로그램(이재성 & 최기선, 1997)을 사용하면 보다 효율적으로 현단계의 한계를 보완하는 효과가 있을 것이다.

7. 결론

본 연구에서는 영어 외래어 변이형 데이터베이스 ELK 외래어 전자사전의 구축 과정과 이를 이용한 영어 어휘 교육용 프로그램인 EDUT 프로그램의 개발과 활용에 대해서 간단히 살펴보았다. 최근 영어와 외래어 사이의 철자와 발음에 해당하는 간격이 점점 넓어짐에 따라 나타나는 영어 외래어에서 나타나는 다양한 전사 표기 형태가 야기하는 여러 가지 문제에 대해서 본격적으로 해당 데이터베이스를 구축하여 그 해결 방안을 모색하고자 하였다.

이와 같이 다양한 표기 방식으로 나타나는 외래어에 대한 체계적인 데이터베이스의 구축은 여러 응용분야에서 매우 요긴하게 사용될 수 있다. 가령 자동처리 시스템에서 가장 처리하기 힘든 문제 중의 하나가 바로 미등록어의 문제이다. 이러한 미등록어의 주된 원천이 영어 외래어인 만큼, 해당 프로그램이 이러한 외래어를 정확히 인식하는 것만으로도 문제를 크게 완화시킬 수 있을 것이다. 실제로 21세기 세종기획에서 무료로 배포하고 있는 코퍼스에서도 확인할 수 있듯이, 한국어로 작성된 수많은 문서에서 많은 고유명사 명칭과 전문용어가 외래어로 나타나기 때문에 고유명사 추출이나 전문용어 추출에도 도움을 줄 수 있다.

또한 연구에서 개발된 영어어휘 교육용 프로그램에서 본 것과 같이, 영어 외래어의 정확한 색인이나 검색이 가능해지기 때문에 다양한 정보검색 시스템에도 유용하게 사용될 수 있다.

---

12) 대상문서는 21세기 세종계획(http://www.sejong.or.kr, 국립국어원 http://www.korean.go.kr)에서 무료로 배포하고 있는 말뭉치 자료와 인터넷 연합뉴스에서 무작위로 추출하여 사용하였다.

최근 여러 형태의 기사[13]에서도 나타나듯이 영어 학습자들이 가장 많이, 효율적으로 사용하고 있는 영어 학습 프로그램은 인터넷 포털 사이트에서 유·무료로 제공되고 있는 자동번역 시스템이라고 할 수 있다. 이러한 자동번역 시스템에서도 한국어 분석시 다양한 음차표기 형태들을 올바르게 대응 원어에 연결시켜 줄 수 있기 때문에 번역의 질을 향상시키는 데에도 많은 도움이 될 것으로 보인다.

궁극적으로 본 연구 결과는 교육용 목적을 위해서도 중요한 자원이 될 수 있는데, 가령 일반적으로 영어단어를 검색하기 위해서는 원하는 영어단어의 정확한 철자를 입력해야 하지만, 영어단어의 철자 대신에 발음만을 기억하는 경우에 정확한 철자를 모르는 성인들을 위한 영어사전과 최근 초등학생들을 위한 조기영어교육용 영어사전으로도 이용될 수 있을 것이기 때문이다.

앞서 본 것처럼 잘못 사용되는 영어 외래어가 초기의 영어 학습자들의 영어 학습에 부정적인 영향을 준다는 점에서 무엇보다도 외래어 표기가 보다 효율적으로 영어 원음의 소리를 가장 잘 나타낼 수 있도록 구성되어야 함을 지적할 수 있다. 하지만 앞서 지적한 바와 같이 영어와 한국어는 서로 다른 음운체계를 지니고 있으므로 두 언어 간의 정확한 대응은 기대하기 어렵다(정국, 1986). 현 외래어 표기법만 살펴보아도 '가제(gauze)'와 같이 실제 영어의 원음과는 거리가 먼 표기가 대다수이며 동시에 언중들이 사용하지 않는 형태를 표준형으로 표기하고 있는 것들이 많다. 또한, 국립국어원에서 제정한 외래어의 표준 표기 원칙이 있기는 하지만, 말이란 누가 제정하는 것이 아니라 사람이 쓰는 것이기 때문에 개개인의 발음에 따라서 각각의 어휘에 관한 음차표기가 다양하게 나타날 수밖에 없는 것이다. 그래서 본 연구의 외래어 데이터베이스 ELK는 외래어 데이터의 대표형뿐만이 아니라, 다양한 가능한 음차표기 형태들을 모두 고려하여 포함함으로써 온라인, 오프라인에서의 여러 응용 분야에서 사용되어 질 수 있게 하였다.

현재 본 연구에서 영어 어휘 교육용 프로그램의 가장 중요한 자원인 외래어 변이형 데이터베이스 ELK는 아직 초기 단계에 불과하고 보충되어야 할 작업이 더 많이 남아있는 게 사실이다. 그러나 이에 대한 지속적인 보완과 확장을 통해 현 연구의 결과가 더 많은 방면에서 다양하게 활용되어 질 수 있기를 기대해본다.

13) http://news.nate.com/view/20090617n01267

참고문헌

남지순. (2005). 프랑스어 언어 자원 구축을 위한 부분문법(grammaire locale) 방법론의 소개. *한국프랑스학논집. 49*, 67-94.

이재성 & 최기선. (1997). *정보 검색을 위한 외래어 자동표기 모델.* 한국과학기술원.

이재성. (1999). 다국어 정보검색을 위한 영한 음차 표기 및 복원 모델. *한국과학기술원 전산학과 박사학위 논문*

오종훈. (2000) 전문분야 사전과 코퍼스 및 외래어 인식에 기반한 전문용어 추출. *한국과학기술원.*

정국. (1984). 외국어 발음의 다섯가지 문제(Five Problems in Korean speakers' Pronunciation of English). *언어와 언어학, 10.*

정국. (1986). 외국어발음 : 인식과 재생의 심리과정(The Recognition and Reproduction of Foreign Sounds; A Generative Approach). *언어와 언어학, 12.*

정국. (1988a). 외국어투의 우리 말과 글. *국어생활, 12.*

정국. (1988b). 음운론의 제 이론과 외국어 발음교육(Phonological Theories and the Teaching of Foreign Language Pronunciation). *영어 영문학, 34*(2)

정국. (2002). 외국어 외래어 한글표기의 문제점과 기본원칙. *한국외국어대학교논문집, 34.*

정국. (2003). 외래어 표기법과 외국어 발음(Transcription of loanwords and pronunciation of foreign languages). *외국어교육연구논문집, 17.*

정길순. (1998). 정보검색을 위한 외래어 자동추출 및 영어단어로의 자동음역. *충남대학교 대학원 학위논문집.*

Carter, R. (1998) *Vocabulary: Applied linguistic perspectives* 2nd ed.. London: Routledge.

Cheon, S-M. (2005). Constructing ELC Database. *언어학, 13*(2), 96-119.

Crothers, E., & Suppes, P. (1967). *Experiments in second language learning*. New York, NY: Academic Press.

Nam J. S. (1994). *The Dictionary of Simple Nouns in Korean.* Dictionnaire Des Noms Simples Du Coreen. LADL, Universite Paris 7 - CNRS.

Nam J. S. (2003). *The Dictionary of Proper Nouns in Korean.* Technical Report 03-03. DICORA.

천승미
121-040 서울시 마포구 도화동 560
서울디지털대학교 영어학부
전화: (02) 2128-3021
이메일: smcheon@sdu.ac.kr

Received: 1 April, 2009
Revised: 16 February, 2010
Accepted: 14 March, 2010