외국어로서의 한국어 음성 코퍼스 구축과 이를 통한 외국인의 한국어 음성·음운체계 습득 양상 연구[1)]

이석재* 김정아** 장재웅***
* 연세대학교 영어영문학과
** 연세대학교 언어정보연구원
*** 안양대학교 중국어과

Speech Corpus for Korean as a Foreign Language and the Aspects of the Foreign Learners' Acquisition of the Phonetic and Phonological Systems in the Korean Language

Seok-Chae Rhee*, Jeongah Kim**, Chae-Woong Chang***
* Dept. of English Language and Literature, Yonsei University
** Institute of Language and Information Studies, Yonsei University
***Dept. of Chinese Language, Anyang University

scrhee@yonsei.ac.kr, jonakim_99@yahoo.co.kr, chang-jw@hanmail.net

Abstract

This study aims to establish a speech corpus for Korean as a foreign language (L2 Korean Speech Corpus, L2KSC) and to examine the aspects of the foreign learners acquisition of the phonetic and phonological systems in the Korean Language. In the first year of this project, L2KSC will be established through the process of reading list organizing, recording, and slicing, and the second year includes an in-depth study of the aspects of foreign learners Korean acquisition and a contrastive analysis of phonetic and phonological systems. The expectation is that this project will provide significant bases for a variety of fields such as Korean education, academic research, and technological development of phonetic information.

Ⅰ. 서론

발화자 자신의 모국어에 대한 음성코퍼스의 연구와 구축 작업은 국내외에서 이미 상당한 수준으로 이루어진 데 반해, 외국어로서의 한국어 음성코퍼스 설계 및 구축 작업과 이를 통한 한국어 음성·음운 체계 습득 과정의 연구는 매우 미비한 실정이며, 외국인을 대상으로 한 한국어 발음 교육도 교수자의 개별적 소규모 자료와 경험을 바탕으로 이루어 진 것이 대부분이다.

이에 본 연구에서는 지금까지 국내외에서 구축되지 않은 외국인 한국어 학습자의 한국어 발화 목록을 설계, 실제 디지털형식으로 녹음 가공한 대용량 음성코퍼스의 구축, 즉 연구의 바탕이 되는 기초 자료 칭출을 그 1차 목적으로 한다. 또한 L2 한국어의 음성·음운 체계가 외국인에게 습득되는 양상을 조사하여 언어간 간섭에 따른 인간의 말소리 습득 특징과 그 양상에 관한 음성·음운론적 지식 생산을 본 연구의 2차 목적으로 삼는다. (본 연구에서 구축되는 「외국어로서의 한국어 음성코퍼스」는 「L2KSC」 (=L2 Korean Speech Corpus)라 간략히 줄여 부르기로 한다.)

Ⅱ. 연구방법

본 연구는 다음과 같은 질문에 대한 답을 구하기 위한 작업이라 할 수 있다. 첫째, 한국어 학습자 외국인

1) 본 논문은 한국학술진흥재단 기초학문육성 인문사회일반 연구(2004-8-0033) 지원으로 수행되었습니다.

의 한국어 발화에서 한국어 음성 및 음운 체계(음운규칙/음운제약)의 어떠한 면의 습득에서 어떠한 특징을 보이며, 둘째, 외국인의 모국어 (약 40개 언어) 특징에 따른 한국어 음성·음운 체계의 습득 정도는 어떠한 양상을 보이는지, 그리고 한국어 학습 수준과 한국어 음성·음운 체계의 습득 정도는 어떠한 관련성에 대한 연구를 하게 될 것이다. 더불어 한국어 학습자 외국인의 한국어 발화에서 한국어의 분절음과 초분절적 운율 요인의 습득 양상에 어떠한 차이를 보이며, 또한 이것이 외국인의 한국어 학습량과 어떠한 상관관계를 보이는지 다루게 될 것이다. 교실 외 언어 환경이 외국어로서의 한국어 습득에 끼치는 영향에 대한 연구도 수반될 것이다. 물론 외국어로서의 한국어 음성코퍼스 L2KSC의 설계와 녹음 과정 진행에 대한 논의 및, 완성된 L2KSC가 사용자의 편의를 위해서는 어떠한 구조로 구성되어야 하며 주석 방법의 효용성을 높이기 위하여 어떤 방법을 취해야하는지, 외국어로서의 한국어 교육에 어떻게 응용 및 활용될 수 있는 방안에 대한 논의도 이루어 질 것이다.

1. 1차년도 진행

음성 코퍼스 구축의 과정은 다음과 같은 단계에 따라 진행될 것이다. 즉 음성 자료 수집 전 단계로서, 외국인들이 발화하게 될 발화 목록을 준비하는 단계인 L2KSC 설계의 단계를 거쳐, 녹음과 L2KSC 구축의 단계에 이르게 될 것이다. L2KSC 구축 단계에서는 녹음이 완료된 것을 슬라이싱(slicing)하여 파일화하고, 체계적인 디렉터리 구조를 설정하며, 여기에 모든 발화자 외국인의 부가 정보(모국어, L2 한국어 학습수준, L2 한국어 학습시간, 연령, 출신국가와 특정 지역, 교육정도)를 부여한다.

1.1. L2KSC 설계

구축될 L2KSC는 다음과 같은 특징을 갖는다.

ㅇ 정형 음성코퍼스: 발성할 내용을 미리 준비
ㅇ 독립어 어휘음성과 연속음성 (문장 및 이야기) 복합 음성코퍼스
ㅇ 낭독체와 대화체의 복합 음성코퍼스
ㅇ 다중 언어 음성코퍼스: 각 외국인 화자가 자신의 모국어를 발화한 부분을 포함.
ㅇ 다중 병렬 음성코퍼스: 한국인의 한국어 발화도 포함하여 같은 내용에 대한 외국인 발화와의 비교 분석을 수월하게 함.
ㅇ 분석용 음성코퍼스: PBW와 PBS가 되도록 함

본 연구에서 이루어질 코퍼스가 한국어 발음 교육용 음성 DB로서 활용되기 위해서는, 먼저 한국어 학습자들이 주로 어떤 발음 오류를 보이는지를 살펴보고, 이들을 교정할 수 있는 자료를 선정해야 한다. 기존의 연구에서 이루어진 자음과 모음, 음운 변화의 발음 뿐 아니라 억양의 오류 유형도 함께 살펴보고자 하며, 이를 고려하여 발화 목록을 작성하였다.

발화 목록 작성시 고려되었던 외국인 학습자들의 대표적인 발음 오류 사항들은 다음과 같다. 대표적으로 일본어권 학습자, 중국어권 학습자, 영어권 학습자의 발음 오류를 유형별로 제시하면 다음과 같다.[1]

ㅇ 영어권 외국인 발화자의 발화 애로:
- 저해음(obstruents)의 발화에서 음절 내 각 위치에 따른 평음, 경음, 격음의 구분,
- 음절 말 저해음 파열 양상;
- 공명음의 탈겹자음화(degemination)현상;
- 단모음의 원순음화(labialization) 현상;
- 이중모음에서 원순음 현상;
- 성문마찰음의 탈락현상;
- 영어 고유 강세와 리듬 패턴의 전이 현상

ㅇ 중국어권 외국인 발화자의 발화 애로:
- 저해음(obstruents)의 발화에서 음절 내 각 위치에 따른 평음, 경음, 격음의 구분;
- 종성의 탈락 현상,
- 설측음의 탄설음화, 탈겹자음화, 탄설음의 설측음화;
- [h]음의 연구개 마찰음화;
- 이중모음의 복모음화, 원순모음과 비원순모음과의 구분, 이중모음의 단모음화;
- 의문사가 있는 의문문의 문미 억양 체계, 문장 내에서 강세구 끝의 억양 체계, 한자어의 중국어 성조체계로의 발화

ㅇ 일본어권 외국인 발화자의 발화 애로:
- 저해음의 발화에서 음절 내 각 위치에 따른 평음, 경음, 격음의 구분;
- 종성자음(coda)의 발화;
- 종성에서 비음의 위치성 구분;
- 일본어의 비명시 모음 첨가에 의해 음절 덧붙임 현상;
- 설측음과 탄설음의 구분,
- 단모음 및 이중모음에서 원순모음/비원순모음의 구분;
- 문미 문중에서 일본어의 피치 악센트 전이 현상

위의 사항들을 고려하면서, 본 연구는 과제 수행을

통해 새로이 구축되는 대용량 L2KSC 실제 결과에 대한 음성실험 H/W, S/W를 이용한 분석을 통해 이를 과학적, 계량적으로 객관화하는 동시에, 외국인의 한국어 음운규칙/제약의 습득 순서를 밝히고 이를 외국인 언어권별로 학습량에 따라 조사하고, 또한 아직 발견되지 않은 L2 한국어 발화상의 음성·음운의 특징을 새로이 밝힐 것이다.

상기에 제시한 것과 같은 방법으로 녹음 대상 한국어 학습자 외국인의 모국어 (약 40개 언어 예정: Bangladesh, Bulgarian, Congo, Czech, Danish, Finnish, French, German, Cantonese, Hungarian, Indonesian, Italian Japanese 등) 대비 한국어 소리 특징을 관찰 고려하여, 이를 발화 목록 작성에 고려한다.

발화 목록은 한국어의 음성·음운적 특징과 각 발화자 모국어의 특징 및 외국인 한국어 학습자의 한국어 오류 등을 고려하여 작성되었으며, 그 구성은 다음과 같다.

- ▸ 발화목록 SET#1: [발화자 모국어 어휘: 자·모음]✇
- ▸ 발화목록 SET#2: [무의미 어휘]✇
- ▸ 발화목록 SET#3: [공통 어휘]✇
- ▸ 발화목록 SET#4: [공통 대화 문장]✇
- ▸ 발화목록 SET#5: [공통 이야기]✇
- ▸ 발화목록 SET#6: [자유 발화]✇
- ▸ 발화목록 SET#7: [영어 문장]✇
- ▸ 발화목록 SET#8: [영어 이야기]✇
- ▸ 발화목록 SET#9: [한국어 추가 이야기]✇
- ▸ 발화목록 SET#10: [한국어 추가 대화 문장]✇

본 연구가 구축하려는 음성코퍼스는 다중 언어 병렬 음성코퍼스의 성격을 지니게끔 되는데, 본 연구는 각 외국인 화자가 자신의 모국어를 발화하는 것도 녹음하여 개별 외국어의 소리 특성과 한국어 소리 특징과의 대조를 위한 비교 기준 자료로 삼을 것이다.

1.2. 발화 외국인의 구성

녹음에 참여할 외국인 발화자는 모국어별, 한국어 학습 정도별로 구분한다. 한국어를 학습 중인 외국인은 연세대학교 외국어학당에서 섭외한다. 발화자 녹음 협력 기관의 2003년 겨울 학기 기준 국적별 분포(아래)를 보면 일본어권과 중국어권, 영어권 발화자가 대다수를 차지하나, 그 외 언어를 모국어로 하는 발화자도 각기 소수이기는 하지만 매우 다양하게 존재하므로 이들의 한국어 발화도 L2KSC에 포함되도록 한다.

▸ Australian 9; Bangladesh 3; Brazilian 9; British 7; Bulgarian 1; Canadian 9; Chinese 125; Columbian 1; Congo 1; Czech 1; Danish 3; Finnish 1; French 6; German 8; Cantonese 9; Hungarian 1; Indonesian 6; Indian 6; Israeli 1; Italian 4; Japanese 221; Kazakhstan 1; Kyrgyzstan 1; Laos 1; Lithuanian 1; Mexican 2; Mongolian 12; Morocco 1; Myanmar 1; Nepali 1; Dutch 3; Nigerian 2; Philippians 3; Polish 1, Russian 33; Slovak 2; South African 1, Swedish 3, Swiss 3; Taiwanese 20; Thai 3; Turkish 2; Ukrainian 2; American 35; Uzbekistan 14; Vietnamese 11, Korean-American 20

국내의 한국어 학습자 예정 발화자 구성 및 인원수는 영어권, 일본어권, 중국어권 외국인화자 각 40명과 기타 40개 언어의 화자 150명, 그리고 한국인 화자 50명으로 총 300명~320명을 예상 인원으로 삼고 있다. 현재 영어, 일본어, 중국어권 화자의 녹음은 이미 완료되었으며, 몽골, 러시아, 우크라이나, 베트남 등 기타 언어의 화자들을 포함하여 총 204명의 녹음이 이루어진 상태이며, 한국인 화자 50명의 녹음도 완료되었다. 기타 언어의 화자를 가능한 다양하게 섭외하여 예상 인원의 녹음을 완료할 계획으로 있다.

또한 영어권, 일본어권, 중국어권은 외국 각 현지에서의 한국어 교육을 받은 외국인 학습자와 국내에서 학습 받은 외국인 학습자의 비교 연구를 통해 교실 외적 학습 환경의 중요도를 파악하기 위하여 외국 각 현지에서 비슷한 학습량을 지닌 발화자를 섭외하여 녹음을 실시하고자 한다. 이들의 예정 발화자는 영어권, 일본어권, 중국어권 각 20명으로 기대하고 있다.

1.3. 녹음진행

녹음은 방음실에서 실시하는 것을 원칙으로 하며 아래의 기기를 사용한다. (샘플링 주파수: 48kHz, 양자화 비트수: 16bit, 녹음 장비 및 장소는 음성코퍼스 부가정보에 반드시 포함한다)

- ○ 디지털 녹음기: TASCAM DA-20MKII, TASCAM DA-P1
- ○ 오디오 믹서: Behringer MXB1002
- ○ 헤드셋 마이크: Sennheiser HMD25-1, Shure SM10A

녹음 후 처리 작업은 DAT에 녹음된 것을 16kHz, 16bit로 PC파일(raw형식)화 하고, 이를 슬라이싱(slicing) 툴을 사용하여 어휘는 각 어휘별로, 문장은 문장별로, 이야기 읽기는 이야기 읽기별로, 대화체는 대화 세트별로 파일화 한다. 이때 화자의 고유 번호를 발화자

외국인의 모국어, 남/여, 연령, 한국어 능력별로 구분할 수 있는 체제로 한다. 또한 제작될 음성코퍼스의 디렉터리 구조나 파일 이름 체계는 가능한 표준화된 체제를 따르도록 하며, 국내 및 국외 녹음을 파일 상에서 구분을 지울 수 있도록 한다.

[표1] 음성코퍼스 파일 이름 체계

분류	모국어	성별	발화자	SET	해당 파일
	영어EN 일본어JP 중국어CN 러시아어RU ...	남1 여2	일련번호 1~	s1~s10	01~ (001~)

예를 들어, 모국어가 일본어인 남성 발화자가 한국어 학습 최고급반의 2번 발화자라면, 이 발화자의 모든 음성 파일은 JP132로 시작하면서 각 음성 파일에 부합하는 번호가 매겨지게 될 것이다. 다음 예에서 s#으로 표현된 것은 발화자 모국어 자모음(s1), 무의미어휘(s2), 공통어휘(s3), 대화(s4) 구분에 따른 표시이다 (JP132s1_01, JP132s1_02, JP132s1_03, JP132s1_04, JP132s1_05, JP132s1_06, JP132s1_07, JP132s1_08...)

목표량은 정확한 실제 발화자 수와 완성될 발화목록에 따라 차이가 나겠지만, 현재로 1인당 약550개 음성파일 (단독어, 문장, 이야기, 대화체 전체포함)을 예정하고 있으므로 550개 음성파일 x 290명 화자 = 총159,500개의 음성 파일이 될 것이며, 용량은 raw포맷으로 약 15GB정도가 될 것으로 예상된다.

2. 2차년도 진행

2차 연구년도의 진행은 1차 연구년도에 구축된 음성자료를 바탕으로 한국어 대 타 언어(발화자들의 각 모국어)의 다중 언어 대조 언어학적 연구가 된다. 대조비교 분석(Contrastive Analysis)에서 주장한 대로 외국어 학습 시 나타난 오류의 상당수는 모국어(L1)의 음운 체계가 목표 언어(L2)의 음운 체계에 작용한 간섭(interference)에 기인한 것이므로, 광의의 음운 체계에 대한 대조 분석을 바탕으로 L2 한국어 학습에 있어 외국인에 의한 L2 한국어 소리 체계 습득 양상을 집중 연구할 것이다.

외국인의 L2 한국어 발화에서 나타나는 음성·음운의 과학적 대조 연구를 통한 소리 체계 학습 과정 연구를 위해서는 1차 연구년에 구축된 코퍼스의 음성 파일을 대상으로 일정 기준안에 따른 정밀한 소리 정보의 부여 작업이 우선적으로 필요하다. 이렇듯 컴퓨터로 재생되는 디지털화된 음성 파일을 대상으로 소리를 직접 들으며 동시에 파형(waveform)이나 스펙트로그램(spectrogram) 등을 보면서 발화된 각 소리의 특징 정보를 컴퓨터상에서 사전에 마련된 기준에 따라 정보를 부여하는 작업을 레이블링이라고 칭하며, 비교 언어학적 연구는 1차년도에 이루어진 슬라이싱과 더불어 레이블링 작업이 이루어진 파일들을 대상 자료로서 활용하여 이루어질 것이다. (레이블링은 [2], [3]의 기준안을 참조하여 진행될 것이다.) 한국인의 한국어 발화 대비 각 외국인의 모국어별, 학습 수준별, 학습 환경별 변이에 따른 음성(phonetic) 특징에 관한 계량적, 통계적 조사 연구는 이와 같은 자료에 토대하여 이루어질 것이다.

상기의 다중 언어 음성 및 음운 체계 습득의 대조언어학적 연구 결과는 외국인 한국어 학습자의 한국어 음성·음운 체계 습득의 교육 수월성을 높이기 위하여 어떻게 응용될 수 있는가의 문제로 연결된다. 이를 위하여 본 연구는 외국인의 한국어 발음 학습 교재 개발의 구체적 시안을 마련하고자 한다. 교육에의 부가적 2차 응용으로 본 연구는 외국인 한국어 학습자 오류 사전, 오류 용례 검색 시스템 등의 기초 자료로 활용될 수 있는 방안을 연구한다.

Ⅲ. 결론

본 연구에서 얻게 될 외국어로서의 한국어 음성코퍼스와 음성·음운 체계 습득 과정에 대한 연구 결과는 다음과 같은 여러 가지 분야에서 활용될 수 있을 것이다. 먼저 교육적 방면에서는 모국어의 특성이 고려된 L2 음성코퍼스를 통해 한국어 음성·음운 체계의 특징을 객관화 시킬 수 있으며, 연구 결과물은 음성, 언어 및 각종 사전 데이터베이스 등 양질의 한국어 교육 자료의 응용으로 확대시킴으로써 한국어 교육을 한 단계 높은 수준으로 향상시킬 수 있다. 둘째, 학문적으로 볼 때 가시적이면서도 신뢰할 수 있는 음성데이터 자료에 토대하여 음성·음운 분야의 비교·대조 연구를 할 수 있으므로, 음성·음운 시스템에 관한 언어 연구의 위상을 높이고 새로운 지식을 창조할 수 있다. 셋째, 기술적 방면에서는 음성정보기술개발을 체계적으로 지원할 수 있다는 점을 꼽을 수 있을 것이다 최근 컴퓨터와 인간간의 대화 수단으로 음성을 활용하는 기술인 음성정보기술의 연구가 활발하고 이에 필요한 잘 정비된 음성코퍼스는 필수적인 요소이다. 여기에 본 연구에 의해 구축되는 음성코퍼스 L2KSC를 통한 언어학적 지식 축적은 언어 연구의 활용성 확대에 일조할 것이다. 마지막으로 사회적인 방면에서는 표준적인 외국

어로서의 한국어 음성코퍼스를 구축함으로써 그 동안 영어권에 편중되었던 한국어 교육에서 벗어나 각 나라별로 맞춤교육이 될 수 있으므로 한국어를 보다 널리 보급시킬 수 있을 것이다.

참고문헌

[1] 정명숙, "한국어 발음 교육을 위한 음성 DB 구축 방안", *말소리* 47, pp.51-72, 2003.

[2] 이숙향, 신지영, 김봉완, 이용주, "음성 코퍼스 구축을 위한 SiTEC 분절음·운율 레이블링 기준의 검토 및 제안", *말소리* 46, pp. 127-143.

[3] Sun-Ah Jun, "K-ToBI (Korean ToBI) labelling conventions: Version 3", *UCLA Working Papers in Phonetics* 99, pp.149-173, 2000.