한국학연구 제62집 | 245~265쪽 | 2021.8.

한국어 학습자 음성 자료의 전사 방식에 대한 연구*

-발음 교육용 음성 자료 구축을 위하여-

하신영**

〈차 례〉

1. 서론
2. 음성 자료의 가공 방식 : 철자법 전사와 음소 전사
3. 음소 전사 자료의 활용
4. 결론

[국문초록]

본고는 한국어 학습자의 발음 교육 연구를 위해 학습자 구어 말뭉치를 구축하여 가공하는 과정에서 어떠한 전사 방식을 사용할 것인가에 대한 논의이다. 한국어의 음성 자료를 대상으로 한 말뭉치는 사용 목적에 따라 철자법 전사, 음소 전사, 음성 전사 등 다양한 방식으로 전사할 수 있다. 그러나 지금까지 구축된 말뭉치 자료들은 대개 철자법 전사를 적용하여 원시 말뭉치를 구성했기 때문에 발음과 관련된 특징을 관찰하는 데에 불편함이 많았다. 대규모 학습자 말뭉치도 기본적으로 철자법 전사가 이루어졌고 음소 전사 방식이 부분적으로 적용되었으나 그런 방식으로 전사된 자료는 학습자의 발음을 관찰하기에 적절하지 않다는 것을 확인했다. 본고는 적어도 한국어 발음 교육의 분야에서 사용될 전사 자료는 음소 전사 방식으로 처리되어야 함을 주장하며 1차적인 기준을 제안하였다. 덧붙여 음소 전사된 학습자 음성 자료의 예를 간단히 제시하였다.

[주제어] 학습자 음성 자료, 철자법 전사, 음소 전사, 발음 교육

---

* 이 논문은 인천대학교 2020년도 자체연구비 지원에 의하여 연구되었음
2021년 국어학회 여름 학술 대회 발표문을 수정 보완한 논문임.
** 인천대학교 국어국문학과 조교수

1. 서론

이 글은 한국어 학습자의 발음 교육 연구를 위해 학습자 구어 말뭉치를 구축하여 가공하는 과정에서 어떠한 전사 방식을 사용할 것인가에 대한 논의이다. 한국어의 음성 자료를 대상으로 한 말뭉치는 사용 목적에 따라 철자법 전사, 음소 전사, 음성 전사 등 다양한 방식으로 전사할 수 있다. 그러나 지금까지 구축된 말뭉치 자료들은 대개 철자법 전사를 적용하여 원시 말뭉치를 구성했기 때문에 발음과 관련된 특징을 관찰하는 데에 불편함이 많았다. 발음 교육을 전제로 한 음성 자료의 구축을 위해서는 그 목적에 맞는 방식으로 전사된 자료가 필요하다. 본고는 철자법적 전사를 기본으로 이루어진 기존의 전사 자료로는 발음 교육을 위한 자료로 쓰이기에 적절하지 않음을 보이고 음소 전사 방식의 자료가 필요함을 확인하는 기초 연구라고 할 수 있다. 이를 위해 2장에서는 기존의 한국어 구어 자료와 한국어 학습자 구어 자료의 전사 방식을 면밀히 살피고 부족한 부분을 검토하여 발음 교육을 위한 음소 전사 방식을 제안한다. 3장에서는 음소 전사 방식으로 전사된 자료의 예를 통해 이 방식이 학습자의 발음을 관찰하고 오류를 수정하는 데에 어떻게 활용될 수 있는지 살펴보도록 하겠다.

2. 음성 자료의 가공 방식 : 철자법 전사와 음소 전사

문자로 이루어진 자료가 아니라 음성으로 이루어진 자료는 연구 목적에 따라 두 가지로 나뉘어 다루어져 왔다. 하나는 구어 자료에 대한 언어학적 기술을 목적으로 대규모로 구축되었던 것으로서 텍스트 형태의 자료가 기본이 되고 품사, 형태, 통사, 의미, 담화 등의 언어학적 주석이 이루어졌다. 현재 일반적으로 구어 말뭉치라고 일컫는 자료가 이에 해당한다.[1] 또 하나는

1) 이때의 구어란 음성으로 이루어진 자료라는 매체적인 의미보다는 문어와 대비되는 의미로 쓰이고 있는 것으로 보인다. '구어 형태-문어 형태', '구어 문법-문어 문법'처럼 문어와 대비될 수 있는

실험 음성학이나 음성 공학 분야에서 음성 인식, 음성 합성 등을 목적으로 구축된 자료들이다. 음성 파일 형태의 자료가 기본이 되고 발화 신호가 전사 자료와 매핑되고 여러 음성 정보가 부가되는 형태이다(김형정, 2002). 이 분야에서는 매우 정밀한 전사가 이루어지기도 하나 기계 처리에 중점을 두기 때문에 언어학적 연구에 사용되기 어려운 점이 있다.

목적에 따라 자료가 모아지면 그 다음으로 목적에 따라 가공이 이루어진다. 구어 자료에 대한 언어학적 기술이라는 전자의 목적을 이루기 위한 말뭉치 자료를 가공하기 위해 대개 철자법적 전사가 이루어진다.[2] 이렇게 구축된 텍스트 형태의 자료는 형태소 이상의 여러 언어 단위를 관찰하는 데에 적절하기 때문이다. 그러나 음성 자료의 소리 특성-음소, 음운 규칙, 억양 등의 특징을 관찰하려면 철자법적 전사는 적당하지 않다. 얼마나 정밀한가의

특징을 관찰하기 위해서는 구어와 문어를 아우르는 처리/주석 방식이 필요하다. 그러나 대응되는 문어가 존재할 수 없는 음성(발음)의 특징을 관찰하고자 할 때에는 발음의 모습을 가장 잘 드러낼 수 있는 방식의 처리가 필요하다. 그렇게 전사된 자료를 통해서 구어의 음운론적 특징을 관찰해 내고 나아가 장소원(1986 : 193)에서 구분한 구어체 구어와 문어체 구어의 음운론적 특징을 구분해 낼 수도 있을 것이다.

2) 전사방식은 몇 가지로 수준으로 구분된다. 구어 말뭉치 구축의 표준안 마련과 관련된 여러 논의가 진행된 *Eagles Standards and Resources for Spoken Language Systems*(1997)에서는 아래와 같이 최대 9단계의 전사 수준을 구분한다. 사용 목적에 따라 이중 하나, 또는 여러 단계의 전사가 이루어질 수 있다.

1. Recording script
2. Orthographic transcription
3. Morphological, syntactic , semantic, and pragmatic representations
4. Citation-phonemic representation
5. Broad phonetic transcription or phonotypic transcription
6. Narrow phonetic transcription
7. Acoustic-phonetic transcription
8. Physical transcription
9. Prosodic transcription

국내 언어학 연구 관련 말뭉치 전사 층위는 김형정(2002 : 120)에서 다음과 같이 정리하였는데, 이 중 철자법 수준과 한글 음소 수준의 전사 방식이 주로 사용된다.

전사 층위	형태적인 면	음운적인 면	사용 기호
① 철자법 수준	형태소	음소	한글 자모
② 한글 음소 수준	변이형태	음소	한글 자모
③ 음성 수준		변이음	IPA 기호
④ 운율 수준			특수 기호

차이는 있을지라도 관찰하려는 소리의 특징을 보여주는 방식으로 음소 전사 혹은 음성 전사가 이루어져야 한다. 언어 자료를 통해 관찰하려는 언어학적 단위가 무엇인가에 따라 자료를 처리하는 전사 방식은 달라질 수밖에 없다.

1) 한국어 구어 말뭉치 전사 방식

(1) 철자법 전사

대표적인 구어 말뭉치 자료인 21세기 세종 계획 현대 국어 구어 말뭉치의 전사 방식은 기본적으로는 철자법적 방식을 따르되, 발음을 표준어형으로 치환하기 어려운 경우 발음대로 표기하거나 태그에 넣어 보충하는 등, 음소 전사 방식을 부분적으로 택하고 있다.[3] 음성 자료에 드러나는 음성적 특징을 완전히 배제하여 전사할 수는 없으니 어느 정도 반영하여 보여 주고자 타협점을 찾은 것으로 보인다.

(1) ㄱ. 그케 스킨 발르구
ㄴ. 오븐에 넣을〈phon〉느을〈/phon〉 수만 있게끔
ㄷ. 표현을 확실하게 해 버리는〈phon〉뻐리는〈/phon〉 게
ㄹ. 비극의 역사를
ㄹ′. 비극의〈phon〉비극의〈/phon〉 역사를

이 전사 기준에 따르면 발음 표기는 두 부류로 나뉜다. 발음 나는 대로 표기하는 경우와 철자법대로 표기하는 경우이다. (1ㄱ)의 '그케'와 '발르구'는 표준어형이 아님에도 쉽게 원래 형태로 이해될 수 있는 형태라고 판단하

3) 구축 지침 25쪽에는 비표준 발음에 대해서 다음과 같이 전사하도록 하고 있다.

> 또한 발음대로 전사한 것 가운데, 철자법에 맞는 것을 함께 적어 주지 않으면 내용 이해에 어려움이 있는 것이나, 자주 출현하지 않고 일시적으로 나타나는 비표준 발음은 〈phon〉 〈/phon〉로 표시하였다. 〈phon〉 앞에 나온 말을 〈phon〉과 〈/phon〉 사이의 말로 발음한 것이므로, 〈phon〉과 〈/phon〉 사이의 말이 실제 발음된 것이다.

여 발음 그대로 전사하였다. 이에 반해 (1ㄴ)과 (1ㄷ)은 표준어형으로 전사한 후 실제 발음형을 첨기하였다. 이 두 부류를 나누는 명확한 기준은 알 수 없으나 기본적으로 단어의 형태는 표준어형으로 표기하고 그 발음이 표준 발음과 다른 현실 발음을 반영하고 있을 때에는 발음형을 첨기하며, 표준어형과 매우 동떨어진 발음의 경우에는 발음대로 전사하는 방식을 취한다. 이는 표준어와 표준 발음을 기준으로 철자법적으로 전사하는 방식이며 형태소의 음성 형태인 '발음'보다는 문자 형태인 '철자'를 중시하는 방식이다.[4] 철자법적 전사 방식이 가지는 '발음'에 대한 한계일 수밖에 없다.

이렇게 철자 형태를 중심으로 구축된 자료로는 발음을 다루는 음운론적 연구를 진행하기에 번거로울 뿐만 아니라 다른 분야의 연구를 진행하는 데에도 불편함이 있다. 발음 형태는 단순히 음운론적 연구에만 관련된 것이 아니기 때문이다. 대표적인 사회언어학적 연구인 Labov(1972)의 말투 변이 관찰 실험도 음소 'r'의 발음을 대상으로 이루어졌고 Trudgill(1974)에서도 언어변종을 관찰하기 위해 언어가 표기되는가 발화되는가에 따라 구분하여 분석한다. 표준 발음도 발음의 여러 변이형 중 하나(이진호, 2012 : 30)이기에 철자 형태가 드러내는 표준 발음이 연구의 비교 대상이 될 수는 있으나 언어의 변화에 관심을 가지는 연구에서 사용하기에도 철자 전사 자료는 매우 불친절한 형식의 자료이다.

이같은 입장에서 (1)에 제시된 발음 형태들은 다양한 관점에서 연구를 진행할 수 있다. 표준어형 '넣-'이 '눟-'으로 발음되고 '-고'가 '-구'로 발음되는 것은 음변화의 입장에서 말뭉치 자료를 관찰할 수도 있고 방언형과 표준어형의 관계로 관찰할 수도 있다. '바르다'와 '발르다'의 관계는 활용형에서의 재구조화를 보여주는 자료로 관찰할 수도 있고 사용역에 따른 어휘 선

4) 데이비드 싱글턴(2008) 6장에서 어휘와 음운론, 어휘와 정서법이 상호작용을 한다는 여러 증거를 확인할 수 있다. 어휘 단위에는 음운 단위와 표기 단위가 저절로 결정되며 단어의 선택이 음운론적 형태와 표기 형태를 똑같이 결정한다는 설명은 형태소의 '형태'가 무엇인가에 대해 생각하게 한다. 본고는 형태소의 1차적인 기본 형태가 음운론적인 형태인 '발음'이라고 생각하나 문자를 사용하는 사회에서 철자법적 형태인 '표기 형태'도 형태소의 중요한 형태임은 틀림없다. 두 형태간의 중요성 차이보다는 이 둘을 한 형태소의 다면적인 형태, 혹은 이형태로 구분하여 인식하는 것이 언어 현상을 보다 깊이 이해하는 데에 도움이 될 것이라고 생각한다.

택으로 볼 수도 있다. 하지만 표준어를 기준으로 한 철자법적 전사가 이루어진 자료에서 (1)과 같이 부가적으로 덧붙여진 형식으로 발음 형태가 제시되면 구어 자료에서 드러난 발음의 특징을 단순한 오류나 변이형으로 치부하게 되기 쉽다.

(1ㄹ)은 약간 다른 기준으로 설명된다. '비극의'라는 전사자료는 조사 '의'가 단모음 [에]로 발음되었다는 것을 표시한다. 지침에 따르면 조사 '의'를 [에]가 아닌 [의]로 발음하는 화자는 매우 드물기 때문에 전사는 '의'로 하되 [의]로 발음되었을 경우에만 (1ㄹ′)처럼 발음 정보를 넣으라고 되어 있다. 대부분의 경우 [에]로 발음이 됨에도 '의'로 전사하기로 한 것은 매우 철자법에 따르는 전사 방식이며 구어의 발음 형태를 관찰하기 쉽도록 고안된 방식은 아니다.

가장 최근에 이루어진 구어 말뭉치 자료로는 음성 자료와 함께 구축된 '일상대화 말뭉치'가 있다. 음성 인식 등의 연구에 사용될 수 있도록 전사 자료를 음성 자료와 함께 구축하였다. 전사 지침에 따르면 일상 대화 말뭉치는 철자 전사와 발음 전사의 두 층위로 전사가 이루어졌다.[5] 그러나 이때의 발음 전사 부분도 음성 자료의 소리 특징을 그대로 보여주는 부분이라기보다는 형태소의 구어적 이형태를 보이는 것으로 보인다.[6] (2)의 예와 같이 전사를

---

5) 이경일 외(2020 : 71) 일상 대화 말뭉치 전사지침의 기본 원칙은 다음과 같다. (밑줄 및 번호는 필자) ①번 전사 지침만 보면 발음 전사와 철자 전사가 병행되므로 발음 전사는 철자 전사와 구분되는 별개의 정보를 전달하는 것 같지만 ②번 지침에서 알 수 있듯이 단지 맞춤법 표기에 따른 발음, 즉 표준 발음과 차이가 있는 경우에만 발음을 반영하고 그렇지 않은 부분은 철자 전사와 동일한 방식으로 이루어짐을 알 수 있다.

① 발화는 발화된 그대로 전사하는 발음 전사와 한글 맞춤법 및 표준어 규정에 따른 철자 전사를 병행하여 전사하는 것을 기본 원칙으로 한다.
② 발음 전사는 구어의 발음 특성이나 개인적인 발음 특성, 지역적인 특성 등의 이유로 표준 발음에서 벗어난 형식으로 발화하거나 표준 발음이 여러 개인 경우 등 한글 맞춤법 표기에 따른 발음과 차이가 있는 경우에 적용하여 발음 나는 대로 적는다.
③ 철자 전사는 한글 맞춤법 및 표준어 규정에 따라 적는 것으로, 발화 내용은 기본적으로 한글 맞춤법 및 표준어 규정에 따라 전사하며 띄어쓰기도 한글 맞춤법에 따른다.

6) 구어적 이형태라는 용어는 논란의 여지가 있다. '조금'과 '쪼금', '쪼끔'을 두고 보았을 때 의미의 차이가 없다면 자유변이형으로 볼 수 있다. 그러나 철자 형태로는 '조금'으로 고정되고 음운 형태로는 '쪼금'이 사용된다고 하면 사용역에 따라 구분되는 분포를 보이는 이형태의 일종으로 생각할

진행하면 동등한 무게를 표현하는 병렬 코퍼스인 것처럼 보이나 사실은 세종 구어 말뭉치에서 특정 부분에 발음 정보를 표시하던 것을 전체 문장으로 풀어 놓은 것에 불과하다.

(2) ㄱ. 철자 전사 : 자 상담소에는 어떤 걸 기대하고 왔을까?
발음 전사 : 자 상담소에는 어떤 걸 기대하고 왔으까?
ㄴ. 철자 전사 : 조금이라도
발음 전사 : 쪼금이래도

기본적으로 철자법적 전사 방식, 그 중에서도 표준어형을 기준으로 한 철자법 전사 방식은 '소리'라는 구어의 특징을 드러내기 위해 고안된 방식이 아니기 때문에 발음의 특징을 드러내는 데에 한계가 있을 수밖에 없다.

(2) 음소 전사

기존의 대규모 전사 자료 중에서 음소 전사로 이루어진 자료로는 지역어 조사 자료가 있다. '지역어 조사 사업'(2004-2013)의 구술 발화 부분은 (3)과 같이 음소 전사로 이루어졌다.

(3) 〈지역어 조사의 구술자료 전사 예〉(2013년 강원 삼척 지역어 전사 자료)

그럼 그걸 또 어뜨게 하느냐 하며는 이러케 벼치 나며는
{그럼 그걸 또 어떻게 하느냐 하며는 이렇게 볕이 나면}
그 식쩌네 그러케 인자 방아다
{그 식전에 그렇게 인제 방아에다}

수 있지 않을까 한다. 물론 한두 음소 차원의 교체를 보이는 수준에서 나아가 '에게'(문어형)와 '한테'(구어형), '와/과'과 '이랑', '하고' 등의 관계에까지 설명하려하면 말투전환(style-shifting)의 음운론적 차원으로 볼 수도 있을 것이다. 이에 대해서는 앞으로 많은 연구가 필요하다고 생각한다.

음소 전사된 자료를 통해 지역어 어휘의 발음 형태를 잘 관찰할 수 있고 어떤 음운 규칙이 적용되는지 혹은 되지 않는지 한눈에 확인할 수 있다. 그러나 철자법 전사와 반대로 어간과 어미의 구분이 쉽지 않고 의미 해석에 시간이 더 걸리는 불편함이 있다. 음소 전사로 이루어진 자료 해독의 불편함을 줄이기 위해 표준어 대역이 부가되었다. 이런 불편함이 있음에도 음소 전사 방식이 선택된 것은 국내 방언학 연구가 음운론적인 연구를 중심으로 이루어진 것과 관련이 있을 것이다.

음소 전사된 자료를 통해 자료의 음운론적인 특징을 잘 관찰할 수 있다는 장점이 있지만 문법이나 담화 연구에 불편함이 있다는 평가가 이루어지기도 한다.[7] 이러한 평가는 관찰하려는 언어 층위에 따라 자료를 처리하는 전사 방식이 달라질 수밖에 없다는 것을 반증한다. 이 같은 상황을 반영하여 '지역어 보완 조사' 정승철 외(2020 : 90)에서는 결과물의 활용 가치를 높이기 위해 음소 전사로만 이루어졌던 기존의 구술 발화 자료에 형태음소 전사 부분을 덧붙여 다음과 같이 보완하였다.

(4) 번저나가는 쪽뿌터 발러 드러가더라구요. 〈음소 전사〉
번저 나가는 쪽부터 발러 들어가더라구요. 〈형태음소 전사〉
{번저 나가는 쪽부터 발라 들어가더라고요.} 〈표준어 대역〉

(4)의 예에서 말하는 '형태음소 전사' 방식은 기존의 철자법적 전사 방식의 온전한 형태라고 말할 수 있다. 음운 규칙으로 예측될 수 있는 부분은 표기하지 않되 형태소의 발음형(기저형)은 그대로 표기에 드러내는 방식으로서 표준어형이라는 기준이 배제된 철자법적 전사 방식의 전형을 적용한 것이다.

철자 전사와 음소 전사 방식의 실례를 살펴보면서 말뭉치의 사용 목적에 따라 전사 방식이 달라져야 함을 알 수 있었다. 한 가지 전사 방식에 여러 층위의 정보를 부가하려다 보니 철자법적 전사 방식에 음소 전사 방식을 혼

7) 대화를 통해 할 수 있는 언어학적 작업이 주로 문법이나 담화 연구로 제한된다면, 음운 전사보다는 오히려 형태 전사가 더 효율적이라는 생각 때문이다(이기갑, 2010 : 75).

합한 (1)과 같은 전사 자료가 만들어졌다. 그러나 한 가지 자료에 한 가지 전사 방식만을 적용해야 한다는 생각에서 벗어나면 (4)와 같이 여러 방식의 전사 자료를 구축할 수 있다.[8)]

2) 한국어 학습자 구어 말뭉치 전사 방식

(1) 철자법 전사

그렇다면 한국어 학습자 말뭉치 자료에서는 어떠한 전사 방식을 택하고 있을까? 학습자의 구어는 학습자 모어 음운 체계의 부정적 전이 때문에 목표어의 일반적인 특징과는 다른 모습을 보이며, 목표어에 도달하지 못한 채 부정확하고 미숙한 체계를 가지므로 일반적인 구어 전사보다 훨씬 더 복잡한 문제에 직면할 수밖에 없다(강현화 외, 2016). 그러나 그간의 연구에서는 대규모 말뭉치 자료의 처리에 있어서 경제성과 문어 자료와의 호환성, 전사자 훈련의 어려움 등을 고려하여 구어 자료도 철자법 전사로 처리하는 방향으로 이루어져 왔다(안의정 1998, 김형정 2002, 강현화・한송화 2016).

대표적인 학습자 말뭉치인 한국어 학습자 말뭉치도 철자법 전사로 구축되었다.[9)] 철자법 전사가 기본이 되기는 하지만 발음에 대한 정보를 담기 위해 음소 전사 방식을 결합하는 형식을 취했다. 학습자의 발음 양상은 한국어 모어 화자의 발음 양상보다 훨씬 복잡하고 다양한 모습을 보이기 때문에 발음

---

8) 경주 지역어 텍스트 자료인 최명옥 외(2007)에서는 음성 전사 자료까지 더하여 아래와 같이 4단계로 제시하였다.

ㄱ. piɾo ɜpt'ammɾida
ㄴ. 피로 업땀마리다
ㄷ. 필오 없단 말이다
ㄹ. 필요 없단 말이다

(ㄱ)은 음성 전사, (ㄴ)은 음소 전사, (ㄷ)은 형태음소 전사, (ㄹ)은 표준어 대역이므로 4단 병렬 말뭉치라고 할 수 있다.

9) '한국어 학습자 말뭉치'는 국립국어원에서 2015년부터 2020년까지 이루어진 사업의 결과이다. 2020년까지 75개국 45개 언어권의 학습자 자료로 구축된 약 4,400,000어절에 이르는 원시 말뭉치 중 구어 말뭉치가 약 1,100,000만 어절이다.

과 관련된 전사 방식에 대한 부가 설명이 매우 길게 달려 있다. 그 중 몇 가지를 살펴보면 다음과 같다.

(5) 〈한국어 학습자 말뭉치 구어 전사 지침〉[10]

ㄱ. 발화 내용은 기본적으로 철자법 수준의 전사를 한다. 다만, 구어의 발음특성, 외국인 학습자의 발음 특성이나 오류, 지역적인 특성 등에 의해 철자법대로 소리 나지 않는 발음(표준 발음이 아닌 경우)에 대해서는 발음 나는 대로 적는다. 이 경우 학습자의 발화는 괄호 안에 밝혀 주되 규범 표기는 괄호 밖에 전사하여 준다.

ㄴ. 국가라는 거하고(거하구) 직결되는 : : 과정이죠.

ㄷ. 친구와(칭구와) 강남에(간남에) 갔습니다(갔슨니다)

ㄹ. 먹는(먹는)〈note〉자음동화 안 됨〈/note〉

ㅁ. 앞에(앞에)〈note〉연음 안 됨〈/not〉

ㅂ. 가구(가구)〈note〉'가'의 ㄱ을 유성음으로 발음〈/note〉

(5ㄱ)의 서술에서 알 수 있듯이 철자법 전사를 하되 발음의 특징은 모두 괄호 속에 넣어 표시하도록 하고 있다. 이때 말하는 철자법이란 철저히 맞춤법에 따른 것으로서 구어의 발음 특성을 드러내는 (5ㄴ)의 '거하구' 같은 경우에도 '거하고'로 전사하도록 한다. 이는 앞서 살핀 세종 구어 말뭉치에서 '바르고'의 구어형인 '발르구'를 그대로 전사하도록 허용한 것보다 훨씬 엄격한 기준이다. 이렇게 전사된 자료에서 괄호 부분을 빼 버리면 발음의 특징은 전혀 반영되지 않은 표준어형 철자법 전사 자료가 될 것이다.

(5ㄷ)은 개별 음소가 표기법이나 표준 발음법에 맞지 않게 발음된 경우를 다루고 있다. 학습자가 종성 비음 'ㄴ,ㅁ,ㅇ'을 구분하는 데에 어려움이 있음을 확인할 수 있는 자료이나 '갔슨니다'와 같은 표기형은 철자법 전사(갔습니

10) 한송화 외(2020 : 67-78)에서 발췌.

다)도 아니고 음소 전사(가쓴니다)도 아닌 어중간한 표기형으로 나타나고 있다. 이는 문제가 되는 '습'만 '슨'으로 바꾸어 적은 것으로 보인다. 또한 일반적인 한국어 화자가 적용하는 음운 규칙이 적용되지 않은 경우는 철자법적 전사로 나타낼 수 없기 때문에 (5ㄹ,ㅁ)의 예와 같이 그때마다 주석을 붙여 주는 수밖에 없는 것이다. 이 주석 자료에는 (5ㅂ)처럼 음성적인 특징까지도 표기하도록 되어 있다.

(2) 음소 전사

최근에 와서는 자료의 활용 방안에 따라 전사 층위를 나눌 필요성에 대한 논의가 이루어지기 시작했다. 안의정(2018)에서는 한국어 구어 전사 말뭉치가 다양한 분야에서 활용되기 위해서는 자료의 활용 방안에 따라 발화 내용과 발화자 층위만 구분하는 간략 전사와 모든 요소를 전사하는 정밀한 전사로 나누어 병행하는 방식에 대해 언급했다. 이승연(2019)에서는 한국어 학습자의 문법 습득 연구를 위한 구어 말뭉치를 구축함에 있어서 철저한 철자법 전사 방식을 택했는데, 문법 습득 연구와 직접 관련이 없는 부분은 전사에 반영하지 않고 발음이나 형태에서 오류가 발생한 경우에는 오류를 교정하여 연구 목적에 맞는 전사 자료를 구축하였다.

학습자를 대상으로 한 어휘 교육, 문법 교육, 담화 교육을 연구하려는 자료로서 구축된 구어 말뭉치라면 철자 전사로 구축되는 것이 적절하다. 그러나 발음 교육을 위한 자료로서 사용되려면 학습자의 발음을 오롯이 반영하여 보여 줄 수 있는 음소 전사 방식이 필요하다. 한국어의 음소 전사 방식은 앞서 살핀 지역어 조사 사업의 음소 전사 방식을 참고 할 수 있다. 그러나 학습자의 음성 자료를 다룰 때에는 한국어 모어 화자의 음성 자료를 다룰 때에는 고려하지 않았던 몇 가지 문제를 해결해야 한다.

먼저 변이음에 대한 음성 전사를 할 것인가의 문제이다.

(6) ㄱ. [kagu] 가구
ㄴ. [gagu] 가구

ㄷ. [gaku] 가쿠

(6ㄱ)이 일반적인 한국인의 발음이라고 할 때 음소 전사를 하면 '가구'로 표기된다. 외국인 학습자가 한국어 음소 /ㄱ/의 변이음에 대한 지식이 부족하여 (6ㄴ)과 같이 [gagu]로 발화한 경우에는 어떻게 전사할 것인가. 유성음으로 발음된 초성의 /ㄱ/은 외국인 악센트로서 어색하게 들릴 수는 있으나 음소 /ㄱ/으로 받아들이기에는 충분하다.[11] 따라서 음소 전사의 입장에서는 '가구'로 전사하는 것이 맞다. (6ㄷ)의 예도 /ㄱ/의 변이음에 대한 지식이 부족한 경우라 볼 수 있다. 그러나 이때에는 어중 초성의 /ㄱ/이 한국어 화자들에게 /ㅋ/으로 인식되므로 음소 전사 결과가 '가쿠'로 달라진다.

두 번째는 한국어의 특정한 음소로 표기하기 힘든 경우이다.

(7) ㄱ. [Φul] 풀* ('ㅍ' 발음이 명확하지 않음)
ㄴ. [koʔ] 곧* ('ㄷ' 발음이 명확하지 않음)

(7ㄱ)과 같은 경우는 마찰성의 정도에 따라 '풀'로 전사될 수도 있고 '훌'로 전사될 수도 있을 것이다. 약간 어색한 발음이더라도 특정한 음소로 인식될 수 있으면 음소 전사를 하면 된다. 그러나 한국어의 음소로서 구분하기 어려운 경우라면 전사자가 듣기에 가장 가까운 음소로 전사하고 그 발음이 명확하지 않음을 표시(*)하기로 한다. 그 발음이 정확히 어떤 음성인지 듣고 정밀 전사하는 것은 필요하지 않다. 한국인이 음소로 인식하기 힘든 발음은 의사소통을 방해하기 때문에 분명한 발음 교육이 필요한 지점이라는 것을 확인하는 것으로 충분하다고 본다.

세 번째는 연음의 표기 방식에 대한 부분이다.

---

11) 모어 화자와 같은 수준의 발음 정확성을 요구하는 것이 아니라 의사소통에서의 이해명료성(intelligibility)에 초점을 두는 발음 교육에서는 외국인 악센트 정도로 인식되는 소리는 수정해야할 교육 대상이 되지 않는다. 다만 언어 간 간섭의 정도를 관찰하려는 언어학적인 연구나 한국어 모어 화자와 동일한 수준의 정확성을 목표로 하는 학습자를 위한 교육 자료로 활용하려는 경우를 위해 이러한 발음의 특징을 표시해 둘 필요는 있다.

(8) ㄱ. 집이
ㄴ. 지비

앞서 살핀 세종 구어 말뭉치나 한국어 학습자 구어 말뭉치의 철자법적 전사 방식에 따르자면 (8ㄱ)에 전사된 '집이'는 [지비]라는 발음을 표시하며 (8ㄴ)의 '지비'도 [지비]라는 발음을 표시하게 된다. 음소 전사 방식에서는 (8ㄴ)의 '지비'만이 [지비]라는 발음을 표시하게 된다. 학습자 구어 자료에서는 '집이'라는 표기도 함께 사용되는데 이때에는 연음되지 않고 종성에 발음된 'ㅂ'을 나타내게 된다.

네 번째는 형태음소적 표기나 철자법 표기와 대응시키기 어려운 경우가 있다는 점이다.

(9) ㄱ. 친구는 나는 생가캄니다
ㄴ. 친구는 나는 생각합니다

전사된 자료만 보았을 때에는 (9ㄱ)은 발음상으로는 오류가 없는 발화로 평가될 수 있다. 개별 음소의 발음이 정확하고 유기음화라는 음운 규칙도 적용되었기 때문이다. 그러나 실제로 (9ㄱ)을 발화한 학습자가 의도한 발화가 '나를 생각합니다'이었으나 한국어의 /ㄹ/ 음소의 발음에 어려움이 있어서 (9ㄱ)과 같이 발음되었다면 발음의 오류라고 판단해야 할 것이다. 연구자나 교사가 이 자료의 전체적인 발음 특징을 분석한 후 오류 여부를 판단하고 교육할 수 있다. 그러나 문제는 이 문장만으로는 화자의 의도가 목적격 조사 '를'이었는지 보조사 '는'이었는지 확인할 수 있는 방법이 없다는 점이다. 이런 경우에 (9ㄴ)과 같이 철자법적 전사를 하게 되면 화자의 문법적 능력과는 다르게 조사의 사용에 오류가 있는 것으로 판단될 수밖에 없다. 음소 전사 자료를 철자법 전사 자료와 대응 시킬 때 고민이 필요한 부분이다.

3. 음소 전사 자료의 활용

음소 전사 방식의 활용에도 약간의 어려움은 있지만 학습자의 발음 교육을 위한 음성 자료를 처리하는 데에는 음소 전사 방식이 가장 적절하다고 판단된다. 이 같은 인식에 기반하여 필자가 참여하여 현재 진행하고 있는 다국어 학습자 말뭉치(INU-MULC)개발 사업의 한국어 부문에서는 구어 자료 구축에 1차적으로 음소 전사 방식을 적용해 보기로 하였다.[12] 학습자 코퍼스 사업의 구어 전사 자료를 음소 전사 방식으로 처리한 첫 번째 시도라고 할 수 있다.

음소 전사 방식은 기본적으로 지역어 조사의 전사 지침을 따르고 기호의 사용 방식은 한국어 학습자 말뭉치의 표기 지침을 참고하였고 학습자 언어의 특징을 반영하여 수정하였다.

- 음소 표기 원칙

① '위'가 상향 이중모음 [wi]로 실현되는 경우와 단모음으로 실현되는 경우 모두 '위'로 전사하되 단모음으로 실현되면 ( ) 안에 그 사실을 밝혀준다.[13]

② '외'가 단모음으로 실현되면 '외'로 전사하고, 이중모음 [we]로 실현되면 '웨'로 전사한다.

③ '에(e)'와 '애(ɛ)'가 변별적 기능을 갖는 경우는 'ㅔ'와 'ㅐ'를 구별하여 표기하고, 이 두 소리가 변별적 기능을 갖지 못하는 경우에는 'ㅔ'로 통일시켜 전사한다. '궤'나 '괘'의 경우도 마찬가지이다.

---

12) INU-MULC(Incheon National University Multi-Language Learner Corpus)는 외국어 학습자 언어 빅데이터를 구축하기 위해 인천대학교에서 2019년부터 진행되고 있는 사업으로 첫해에는 한국인 대학생 학습자의 영어 에세이 자료와 독백 자료가 수집되었고, 2020년에는 한국인 중국어 학습자와 외국인 한국어 학습자의 자료로 범위가 확대되었다. 2021년에도 계속 진행되고 있는 사업으로 지금까지 영어 43만 단어, 중국어 15만 단어, 한국어 7만3천 어절이 구축되었다. 한국어 학습자 자료는 토픽 3급~6급 학습자 48명을 대상으로 수집되었으며 그중 구어 자료는 475분 분량의 음성 파일과 3만 어절 이상의 전사 자료로 구성된다.

13) 익명의 심사자께서 지적하신 것처럼 '위'의 경우 실은 [ɥi]가 자연스러운 한국어 발음일 것인데 어떻게 발음하든 한국어 모어 화자는 /ㅟ/라고 인식할 수밖에 없을 것이다. 학습자의 발음을 정확하게 적는 것만큼이나 한국어 모어 화자가 어떻게 받아 들이느냐도 전사 방식을 정할 때에 고려되어야 한다는 의견에 적극 동의한다.

④ 숫자는 발음나는 대로 전사한다.

- 띄어쓰기 : 음소 전사를 하되 띄어쓰기는 맞춤법에 관계없이 발음에 따라 한다.

- 기호 사용

?	-상승억양	
.	-하강억양	
X ...	-알아들을 수 없는 소리	예) 그때 XX에 가서, 친구랑 ... 만나써
~	-머뭇거림 소리	예) 어~, 음~
: :	-표현적 장음	예) 엔 : : 나레
○○○	-개인 정보가 드러나는 경우	예) 저는 ○○○○○○○임니다

- 전사자 코멘트 : 해당 어절 뒤 괄호( ) 안에 넣는다.

예) 식푸믈(ㅍ 발음이 f처럼 들림)

- 전체적인 발음의 특징 : 언급할 필요가 있을 경우 자료 맨 아래에 덧붙인다.

이 전사 방식을 활용하여 구축한 자료의 일부를 가지고 음소 전사된 구어 자료를 통해 학습자의 발음이 어떻게 드러나고 있는지 관찰해 보기로 한다. 해독을 돕기 위해 아래에 표준어 대역을 병기했다.

(10) 베트남어 화자, 토픽 4급

ㄱ. 처뻔체에는 어~처번체에는 민칩모차 의 의상 소게에드리 께씀니다.
{첫 번째에는 어~첫 번째에는 밀짚모자의 의상 소개해 드리겠습니다.}

ㄴ. 어 잠시만녀. 혹씨 지금 피티 피피티는 화면 본수이쓰세요?

{어 잠시만요. 혹시 지금 피티 피피티는 화면 볼 수 있으세요? }

ㄷ. 게속하 께슴니다. 엄~

{계속하겠습니다. 엄~}

ㄹ. 네 베트나메는 춤믄 엄~민칩모자는

{네 베트남에는 춤은 엄~밀짚모자는}

ㅁ. 엄~종 종신는 담믄 춤므 춤미며?

{엄~정 정신을 담은 춤으 춤이며}

ㅂ. 베트남저셩 여성든이 아름따으믄 특칭 춤임니다.

{베트남 여성 여성들이 아름다움을 특징 춤입니다}

ㅅ. 엄~베트남 베트남 오슨 예기하면 사람든니 바로 아오자이는 생가함니다

{엄~베트남 베트남 옷을 얘기하면 사람들이 바로 아오자이를 생각합니다}

이 자료에서 가장 먼저 눈에 띠는 것은 자음 발음의 오류가 많다는 것이다. (10ㄱ)에서는 '처뻔체(첫번째)', '민칩모자(밀짚모자)' 등의 발음에서 /ㅊ/과 /ㅈ/, /ㅉ/의 발음에 어려움이 있음을 확인 할 수 있다. '민칩모자(밀짚모자)', '화면본수(화면 볼 수)', '여성든이(여성들이)' 등의 발음에서 종성 /ㄹ/은 제대로 발음되지 않는다는 것을 알 수 있다. 연음을 적용하는 데에도 오류가 관찰된다. (10ㅁ)의 '종신는', '담믄', '춤미며'에서처럼 종성 자음을 중복하기도 하며 (10ㅂ)의 '여성든이', '춤임니다'에서처럼 아예 연음을 적용하지 않기도 한다.

이런 발음 오류의 특징을 관찰하면 일차적으로 해당 학습자의 발음 교육에 적용할 수 있다. 그리고 이런 자료들이 모여 체계적으로 분석되면 베트남 학습자들의 한국어 발음 오류 경향을 관찰할 수 있게 된다.

(11) 몽골어 화자, 토픽 3급

ㄱ. 이거는 바로 저기 으~마냐게 전제니 나타네며는 전제니 나타나며는

{이거는 바로 저기 만약에 전쟁이 나타나면은 전쟁이 나타나면은}

ㄴ. 아~초 가뜽거로 머 엔나레는 초 가뜽거로 이러께 전젠 헤써쓰며는
{아~총 같은 거로 뭐 옛날에는 총 같은 거로 이렇게 전쟁했었으면은}
ㄷ. 요즘믄 으~여러분드리 헨무기라는 말 드러보셔쓸꺼에여
{요즘은 으~여러분들이 핵무기라는 말 들어보셨을 거예요}
ㄹ. 이게 저기 지금 부카네서 지금 엄첨 마니 만들고 인는 헨무기
{이게 저기 지금 북한에서 지금 엄청 많이 만들고 있는 핵무기}
ㅁ. 마냐게 그러 그걸로 전제 :: 늘 하게뒈며는 으~사람 사람드리 다 주글꺼에여.
{만약에 그러 그걸로 전쟁을 하게 되면은 으~사람 사람들이 다 죽을 거예요}

예(11)에서 관찰되는 발화는 유창성은 매우 뛰어나나 특징적인 발음의 오류가 관찰되는 경우이다. 종성의 /ㅇ/ 발음에 오류가 있는데, (11ㄴ)의 '초(총)'에서처럼 아예 발음이 되지 않거나 (11ㄱ)의 '전제니(전쟁이)'에서처럼 'ㄴ'으로 발음되고 있다. (11ㄹ)의 '엄첨 마니'의 경우를 보면 '엄청'의 종성 비음이 뒤따르는 '마니'의 영향으로 'ㅁ'으로까지 발화되고 있다. (11ㄴ)의 '가뜽거로(같은거로)'에서는 'ㅇ'이 드러나는데 이때에는 'ㅇ'을 정확히 발음할 수 있어서 변별적으로 발음했다기보다는 후행하는 'ㄱ'의 영향으로 드러났을 것이다. 단순히 종성 /ㅇ/의 발음에 오류가 있다는 사실뿐만 아니라 조음 위치에 따라 발음이 달라지고 있다는 점을 알게 되면 이 학습자의 발음 교육 방향을 잡는 데에 도움이 될 것이다.

4. 결론

한국어의 음성 자료를 대상으로 한 말뭉치를 구축하고 사용하기 쉽도록 가공할 때에는 사용 목적에 따라 여러 가지 방식으로 전사할 수 있다. 지금까지 구축된 말뭉치 자료들은 대개 철자법 전사를 적용하여 원시 말뭉치를 구성했기 때문에 발음과 관련된 특징을 관찰하는 데에 불편함이 많았다. 적어도 한국어 발음 교육의 분야에서 사용될 전사 자료는 음소 전사 방식으로

처리되어야 하며 같은 자료라도 다양한 교육, 연구 분야에서 사용될 목적에 따라 적절한 방식으로 전사되고 주석 작업이 이루어져야 한다. 한 가지 전사 방식을 중심에 두고 덧깁기 하는 것보다는 음소-철자 병렬 말뭉치로 구성하는 것이 효과적일 것이라고 생각한다.

음소 전사 방식으로 구축된 한국어 학습자 말뭉치가 활용될 수 있는 방향을 정리하며 글을 맺고자 한다.

첫째, 한국어 교육 연구 분야에서는 전체 교육 방향을 제시하기 위한 기초 자료가 된다. 학습자 말뭉치에서 교육 단계별로 나타나는 학습자 발음 특징이나 오류의 유형을 확인하여 한국어 발음 교육 전반의 교육 방향을 제시할 수 있다. 또한 언어권별로 나타나는 오류의 경향성을 확인하여 학습자의 모어에 따라 차별화된 교육 내용을 구성할 수도 있다.

둘째, 언어 간 간섭 현상을 관찰할 수 있는 자료가 된다. 학습자의 모어가 지닌 음운론적 특징이 한국어 학습 과정에서 어떻게 전이되어 나는지, 한국어 능력이 향상되어 가면서 그 특징들이 어떤 방식으로 어떻게 변해 가는지를 관찰할 수 있는 좋은 자료가 될 수 있다.

덧붙여 교육 현장에서 교사가 모아서 전사한 자료는 학습자의 발음 진단 자료로 쓰일 수 있고, 개별 학생들의 발달 과정을 관찰하고 오류를 수정해 줄 수 있는 기초 자료로 사용될 수 있다.

▌참고문헌

강현화・한송화, 「한국어 학습자 구어 자료 전사에서의 쟁점과 구어 전사의 실제」, 『한국사전학』 27, 2016, 75~108쪽.

김형정, 「구어 전사 말뭉치의 표기 방법」, 『한국어 구어 연구(1) – 구어 전사 말뭉치와 그 활용 – (서상규・구현정 공편)』, 한국문화사, 2002, 115~175쪽.

데이비드 싱글턴, 『언어의 중심 어휘』(배주채 옮김), 삼경문화사, 2008.

박동근, 「음운 주석 말뭉치 구축」, 『한국어 구어 연구(1) – 구어 전사 말뭉치와 그 활용 – (서상규・구현정 공편)』, 한국문화사, 2002, 199~222쪽.

안의정, 「한국어 입말뭉치 전사 방법 연구」, 연세대학교 대학원 석사학위논문, 1998.

_____, 「구어 전사 말뭉치 구축에 관한 현황과 쟁점」, 『언어와 문화』 14(3), 2018, 81~101쪽.

이기갑, 「구술 발화의 전사와 분석」, 『배달말』 47, 2010, 67~91쪽.

이승연, 「한국어 학습자 문법 습득 연구를 위한 구어 말뭉치 구축과 활용」, 『한국어학』 83, 2019, 69~95쪽.

이익섭, 『사회언어학』, 민음사. 1994.

이진호, 『한국어의 표준발음과 현실발음』, 아카넷, 2012.

장소원, 「문법 연구와 문어체」, 『한국학보』 12(2), 1986, 191~212쪽.

전영옥, 「구어 원시 말뭉치 구축 방법」, 『한국어 구어 연구(1) – 구어 전사 말뭉치와 그 활용 – (서상규・구현정 공편)』, 한국문화사, 2002, 75~112쪽.

최명옥・김주석, 『경주지역어 텍스트 – 1』, 역락, 2007.

Gibbon D., Moore R., Winski R. (eds.), Eagles Standards and Resources for Spoken Language Systems, Walter de Gruyter Publishers, Berlin & New York, 1997.

김봉국 외, 2013년 강원 지역어 전사 보고서, 국립국어원, 2013.

이경일 외, 2020년 일상 대화 말뭉치 구축 최종 보고서, 국립국어원, 2020.

정승철 외, 2020년 지역어 보완 조사 결과 보고서, 국립국어원, 2020.

한송화 외, 2019-20년 한국어 학습자 말뭉치 연구 및 구축 사업 결과 보고서, 국립국어원, 2020.

21세기 세종 계획 말뭉치 구축 지침
https://ithub.korean.go.kr/user/total/referenceView.do?boardSeq=5&articleSeq=103&boardGb=T&isInsUpd=&boardType=CORPUS

What is the proper transcription method of Korean learners' speech material for pronunciation education?

Ha, Shin Young*

This paper is a discussion on which transcription method to use in the process of constructing the learners' spoken corpus for the study of pronunciation education for Korean language. The corpus of spoken data can be transcribed in various ways, such as spelling transcription, phoneme transcription, and phonetic transcription, depending on the purpose of use. However, since most of the corpus materials constructed so far consisted of primitive corpora by applying spelling transcription, it was inconvenient to observe the characteristics related to pronunciation. Even a large corpus of learners was basically transcribed in spelling and partially transcribed by phoneme, but it was confirmed that data transcribed in such a manner was not suitable for observing learners' pronunciation. This paper asserts that at least transcription materials to be used in the field of Korean pronunciation education should be processed in a phonemme transcription method. And the primary criteria for phoneme transcription were presented. In addition, an example of phoneme-transcribed learner spoken data is briefly presented.

* Incheon National University

Key words : learners' spoken data, spelling transcription, phoneme transcription, pronunciation education

논문투고일 : 2021년 7월 06일 ‖ 심사완료일 : 2021년 8월 06일 ‖ 게재확정일 : 2021년 8월 11일