말소리와 음성과학 제1권 제3호 (2009)
pp. 103~108

발음 사전에 기반한 영·한 음차 표기 사전의 구축

Building English-to-Korean Transliteration Dictionary Based on Pronouncing Dictionary

이 도 길[1)]
Lee, Do-Gil

ABSTRACT

This paper proposes a method for building a transliteration dictionary, which is based on pronouncing information extracted from two kinds of existing dictionaries. Also, it proposes a method for transforming the pronouncing information into Korean translitered words. To express the pronouncing information, we define Phoman code system. In order to avoid phonetic estimation process of English words which is the most important problem, the proposed method uses the pronouncing information extracted from the existing dictionaries. Therefore, unlike previous approaches, the proposed method does not need any incomplete phonetic estimation process so that it can produce accurate transliteration results. The proposed method has been fully implemented.

Keywords: transliteration, pronouncing dictionary

1. 서 론

음차 표기(transliteration)는 한 언어로 쓰인 단어를 다른 언어로 표기하는 것이다. 예를 들면, “television”이라는 영어 단어를 “텔레비전”과 같이 한글로 표기하는 것을 의미하며, 그 반대의 과정은 음차 복원이라고 한다. 한국어 문서에는 한글뿐만 아니라 영어, 한자 등과 같은 외국어 문자 표기와 이에 대한 한글 음차 표기가 혼용되고 있다[1]. 이러한 언어 환경 하에서는 동일한 단어를 나타내는 다양한 표현 즉, 이형태가 있기 때문에 단어 불일치(mismatch) 문제가 발생하게 된다. 단어 불일치 문제는 정보 검색 등에서 동일한 개념의 단어에 대한 다양한 표현으로 인해 사용자가 원하는 문서의 검색을 어렵게 만드는 하나의 원인이 된다. 단어 불일치 문제를 해소하기 위해서는 자동 음차 표기와 음차 복원에 대한 연구가 필요하다.

본 논문은 외국어 중에 가장 빈번하게 쓰이는 영어 단어의 한글 음차 표기 문제에 대해 논한다. 음차 표기는 원단어의 발음에 근거한다. 음차 표기를 하는 방식으로는 “입말 표기”와 “눈말 표기”가 있다[2]. 입말 표기는 영어 단어를 발음에 근거하여 한글로 표기하는 방식이고, 눈말 표기는 영어 단어의 철자로부터 발음을 추측하여 한글로 표기하는 방식이다.

음차 표기로 인한 문제를 해결하는 가장 단순하고 효과적인 방법은 영·한 음차 표기 쌍이 저장된 음차 표기 사전을 두어 사전에 등록된 영어 단어를 찾고 대응하는 한국어 표기를 알아내는 방법이다. 그러나 음차 표기 쌍을 수집하는 일에는 많은 노력이 필요하고, 사전에 등록되지 않은 고유명사나 신조어로 인해 미등록어 문제가 발생한다. 따라서 전적으로 사전에만 의존하는 방법에는 한계가 있다. 이러한 이유로 자동 음차 표기에 대한 연구가 진행되어 왔다[1,3-6].

자동 음차 표기를 위한 기존 연구는 “직접 방식”과 “피봇 방식”으로 나눌 수 있다[3]. 직접 방식은 영어 단어로부터 한글 음차 표기를 직접 생성하는 방식이고, 피봇 방식은 영어 단어로부터 발음을 생성하고, 생성된 발음을 이용해 한글 음차 표기로 변환하는 방식이다. 직접 방식은 눈말 표기에 적합하고 피봇 방식은 입말 표기에 적합하다.

자동 음차 표기를 위한 기존 연구들은 대부분 통계적인 접근 방법을 사용한다. 여기에는 확률 모델을 이용한 방법[3], 최대 엔트로피 모델을 이용한 방법[1], 음운패턴이라고 명명한 가변 길이의 영어 발음열과 한국어 표기열의 쌍을 이용한 방법[4], 결정 트리(decision tree)를 이용한 방법[5] 등이 있는데, 이 연구들은 모두 직접 방식에 기반하고 있다. 피봇 방식에 기반한 방법으로는 [6]의 연구가 있다. [6]의 음차 표기 과정은 크게 “발

1) 고려대학교 motdg@korea.ac.kr
접수일자: 2009년 8월 1일
수정일자: 2009년 9월 16일
게재결정: 2009년 9월 17일

음 생성 단계"와 "음차 표기 생성 단계"가 있는데, 발음 생성은 발음 사전을 검색하거나, 발음 추정을 통해 해당 단어의 발음을 결정한다. 발음 추정과 음차 표기 생성 단계는 메모리 기반 학습과 결정 트리를 이용한다.

자동 음차 표기에 있어서 가장 어려운 문제는 영어 단어로부터 정확한 발음을 추정하는 것이다. 발음을 생성하는 과정에서 발생하는 오류는 음차 표기 단계에 전파된다. 명시적으로 중간 단계인 발음 표현을 생성하지 않는 직접 방식도 발음을 추정하는 과정이 포함되어 있다고 볼 수 있다. 따라서 영어 단어에 대한 정확한 발음을 알 수 있다면 음차 표기에 매우 효과적일 것이다.

본 논문은 기구축된 사전으로부터 발음 정보를 추출하여 발음 사전을 구축하고 이로부터 한글 음차 표기로 변환함으로써 음차 표기 사전을 구축하는 방법을 제안한다. 기존 연구에서는 자동으로 발음을 생성함으로써 발음 추정 과정에서 오류가 발생하나, 제안하는 방법은 발음 사전으로부터 정확한 발음 정보를 얻을 수 있으므로 불완전한 발음 추정 과정이 필요하지 않으므로 기존 방법보다 정확한 음차 표기가 가능하다.

본 논문은 다음과 같이 구성된다. 2장에서는 기분석 사전과 발음 사전의 구축에 대해서 설명하고, 3장에서는 발음 사전으로부터 음차 표기 사전을 구축하기 위한 변환 과정을 설명한다. 4장에서는 결과 및 고찰, 5장에서는 결론 및 향후 연구에 대해 논한다.

2. 사전 구축

2.1 기분석 사전

본 논문은 영·한 음차 표기에 초점을 맞추고 있으나, 음차 표기의 원칙은 "외래어 표기법"[2]을 따른다. 외래어 표기법 제1장 제5항의 "이미 굳어진 외래어는 관용을 존중한다"는 원칙이 있다. 그러나 관용적으로 굳어진 표기는 자동으로 음차 표기한 결과와 상이한 경우가 많다. 예를 들면, 영어 단어 "radio"에 대한 음차 표기는 "레이디오"가 올바르지만 실제로는 "라디오"를, "camera"에 대해서는 "캐머러" 대신 "카메라"를 관용 표기로 인정하고 있다. 물론 관용 표기의 상당수는 앞에서 언급한 눈말 표기에 의한 것이나, 반드시 그런 것만은 아니기 때문에 관용적인 외래어 표기와 같이 예외적인 현상을 용이하게 처리하려면 눈말 표기에 의한 자동 음차 표기보다는 기분석 사전을 구축하는 것이 보다 현실적인 방법이라고 생각된다.

본 논문에서는 실제 언어현상에서 관찰되는 모든 관용 표기를 수집하고자 하는 것이 아니라, 이미 표준으로 자리잡고 언중들에 의해 널리 사용되는 외래어 표기를 음차 표기의 대안으로 삼을 수도 있다는 인식 하에 기분석 사전을 도입하였다.

본 논문에서는 약 1만 8천 개의 영어 단어-한글 음차 표기쌍을 수집하였다.

2.2 발음 사전

발음 사전은 기구축된 사전으로부터 발음 정보를 추출함으로써 구축할 수 있다. 본 논문에서 사용하는 기구축 사전은 약 7만 표제어가 수록된 어학용 영한 사전과 CMU 발음 사전[3]이다. [4]에서는 CMU 발음 사전을 이용하여 발음 사전을 구축한 바 있으나, 아직까지 어학용 사전을 이용한 연구는 없다.

표 1. 영한 사전 발음기호와 포만코드 대조표
Table 1. Mapping table between phonetic symbols of English dictionary and Phoman code symbols

발음기호	포만코드	발음기호	포만코드	발음기호	포만코드	발음기호	포만코드
m̩	M	d	d	i	!	ì	!
É	E	ɔ	c	j	J	ó	O
á	A	g	g	k	K	ò	O
à	A	h	h	l	L	ú	U
é	E	i	i	m	M	ù	U
è	E	j	j	n	N	ɵ	*
í	!	k	k	o	O	è	E
ì	!	m	m	p	P	ʌ̀	^
ó	O	r	r	r	R	ʃ	#
ò	O	s	s	s	S	ə	C
ú	U	t	t	t	T	æ	@
œ	q	u	u	u	U	ð	&
ɔ̃	O	w	w	v	V	á	A
à	A	ʌ́	^	w	W	l	l
ɛ́	E	ɑ	A	ŋ	Q	n	n
ǽ	@	b	B	z	Z	p	p
ə	C	ɔ	O	ɑ́	A	v	v
ɛ	E	d	D	ɑ̀	A	x	X
æ	@	e	E	ɔ̀	O	ð	&
ʌ	^	f	F	é	E	ù	U
ː	:	g	G	è	E	ē	E
ə́	C	h	H	ì	!	ʒ	3
ɔ́	O						

본 논문에서는 영한 사전의 국제 음성 기호로 표시된 발음 정보를 다루기 위해 새로운 코드 체계를 정의하였으며, 이 코드 체계를 "포만코드(Phoman code)"라고 명명하였다. 포만코드는 다음과 같은 정의 원칙을 두고 있다.

- 사전의 발음 정보를 ASCII 코드의 인쇄 가능 영역의 글자로만 표현하며, 가급적 발음 정보와 형태적으로 유사한 글자로 대응시킨다.
- 최대한 원어의 발음을 살리도록 한다. [f]와 같이 한국어에 없는 발음도 가급적 표현한다.
- 묵음에 가까운 발음은 소문자로 표현한다.[4]

2) 문교부 고시 제85-11호 (1986년 1월 7일)

3) http://www.speech.cs.cmu.edu/cgi-bin/cmudict

4) 소문자로 표현된 포만코드의 처리에 대해서는 3.3절에서 자

표 2. CMU 발음기호와 포만코드 대조표
Table 2. Mapping table between CMU phonetic symbols and Phoman code symbols

CMU 발음기호	포만코드	CMU 발음 표현 예	
		단어	발음표현
AA	A	odd	AA D
AE	@	at	AE T
AH	^	hut	HH AH T
AO	O	ought	AO T
AW	AU	cow	K AW
AY	A!	hide	HH AY D
B	B	be	B IY
CH	T#	cheese	CH IY Z
D	D	dee	D IY
DH	&	thee	DH IY
EH	E	Ed	EH D
ER	Cr	hurt	HH ER T
EY	E!	ate	EY T
F	F	fee	F IY
G	G	green	G R IY N
HH	H	he	HH IY
IH	!	it	IH T
IY	!:	eat	IY T
JH	D3	gee	JH IY
K	K	key	K IY
L	L	lee	L IY
M	M	me	M IY
N	N	knee	N IY
NG	Q	ping	P IH NG
OW	OU	oat	OW T
OY	O!	toy	T OY
P	P	pee	P IY
R	R	read	R IY D
S	S	sea	S IY
SH	#	she	SH IY
T	T	tea	T IY
TH	*	theta	TH EY T AH
UH	U	hood	HH UH D
UW	U:	two	T UW
V	V	vee	V IY
W	W	we	W IY
Y	J	yield	Y IY L D
Z	Z	zee	Z IY
ZH	3	seizure	S IY ZH ER

• 소리의 장단은 표시하되 소리의 고저(강세)는 표현하지 않는다. 소리의 고저는 한글 표기에 반영되지 않기 때문이다.

사전에 수록된 각 표제어의 발음 기호를 포만코드로 표현하기 위해서, 영한 사전의 발음 기호와 포만코드의 코드값 간의 대응 관계를 <표1>과 같이 정의하였다.

CMU 발음 사전은 모든 사용 목적에 제약없이 공개된 자료로서, 본 논문에서 사용한 발음 사전의 버전은 0.6으로서 표제어 수는 127,069개이다. CMU 발음 사전의 발음기호가 영한 사전의 그것과 다르므로 포만코드와의 대응 관계를 새로 정의해야 하는데, CMU 발음 사전의 발음 기호와 포만코드 간의 대응 관계는 <표2>에 있다.

3. 한글 음차 표기 변환

2장에서 정의한 포만코드를 이용하여 발음 사전을 구축한 뒤에는 발음 사전에 수록된 표제어를 한글 음차 표기로 변환해야 한다. 이 변환 과정은 외래어 표기법에서 정한대로 국제 음성기호와 한글 대조표에 의해 표기하는 것을 원칙으로 하되, 예외적인 현상은 "3장 표기 세칙"에 따라 각 발음에 대한 규칙을 프로그래밍 코드로 작성한다.

<그림1>은 발음 사전에 수록된 각각의 단어에 대한 포만코드로부터 한글 음차 표기를 생성하는 전체적인 과정을 보여준다. 각 발음 표기에 포함된 포만코드를 순서대로 처리하는데, 모음은 모음 처리부(3.1절)에서, 자음은 자음 처리부(3.2절)에서 각각 처리한다. 모음 처리부와 자음 처리부를 거쳐 생성된 한글 자소열이 불완전한 경우 즉, 초성, 중성, 종성[5]의 순으로 배열되지 않은 경우, 완전한 자소열로 변환하고 이를 한글 음절열로 변환한다.(3.4절)

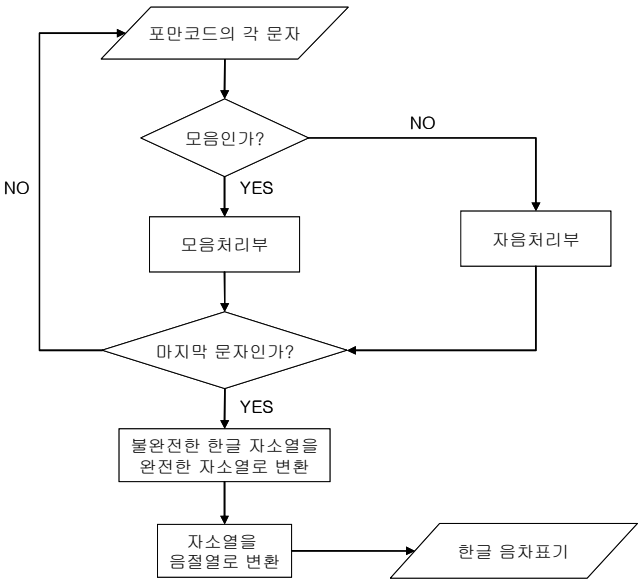

그림 1. 포만코드의 한글 음차 표기 생성 과정
Figure 1. Generating Korean transliteration from Phoman codes

3.1 모음 처리부

포만코드의 모음은 단어의 시작이거나 종성 'ㅇ'이나 모음의 뒤에서는 초성 'ㅇ'을 삽입하고, <표3>에서 정의된 바와 같이 각 모음에 대응하는 한글 자소로 변환한다.

3.2 자음 처리부

포만코드의 자음은 기본적으로 <표4>에서 정의된 바와 같이

세히 다룬다.

5) 종성이 없는 경우에는 채움 코드가 종성을 대신하므로, 모든 한글 음절은 3개의 자소로 이루어진다고 가정한다.

표 3. 포만코드와 한글 자소 대조표 (모음)
Table 3. Mapping table between Phoman code symbols and Korean graphemes (vowel)

종류	포만코드	한글자소
단모음	!, I	ㅣ
	^, C, c	ㅓ
	E	ㅔ
	A	ㅏ
	U	ㅜ
	O	ㅗ
이중모음	@	ㅐ
	q	ㅚ
	OU	ㅗ
	AUC	ㅘ워
	UC	ㅝ
반모음	W!	ㅟ
	W^, WC	ㅝ
	WE	ㅞ
	WA	ㅘ
	WU	ㅜ
	WO, WOU	ㅝ
	W@	ㅙ
	J!	ㅣ
	J^, JC	ㅕ
	JE	ㅖ
	JA	ㅑ
	JU	ㅠ
	JO	ㅛ
	J@	ㅒ

각 자음에 대응하는 한글 자소로 변환하면 되며, 일부를 제외하고 모든 자음은 어말이나 자음 앞에서는 중성 모음 'ㅡ'를 추가한다.

외래어 표기법의 "3장 표기 세칙"에서 언급하고 있는 규정에 따라 자음 처리부에서는 다음과 같은 예외 규칙을 둔다.

- '#'는 어말에서는 '시', 자음 앞에서는 '슈', 모음 앞에서는 '!', '^', 'E', 'A', 'U', 'O', '@'과 결합하여 각각 '시', '셔', '세', '샤', '슈', '쇼', '새'로 표기한다.
- '3'은 어말이나 자음 앞에서는 '지'로 표기한다.
- 어말이나 자음 앞에서 'TS'는 '츠'로, 'T#'은 '치'로 표기한다.
- 어말이나 자음 앞에서 'D3'은 '지'로, 'DZ'는 '즈'로 표기한다.
- 비음 'M'과 'N'은 어말이나 자음 앞에서는 종성으로 표기한다.
- 'L'은 1) 어말이나 자음 앞에서는 종성으로, 2) 어중의 'L'이 모음 앞에 오거나, 모음이 따르지 않는 비음 앞에 올 때는 "ㄹㄹ"로, 3) 비음 뒤에서는 모음 앞에 오더라도 'ㄹ'로 표기한다.
- 짧은 모음 다음의 어말 무성 파열음([p], [t], [k])은 각각 종성 'ㅂ', 종성 'ㅅ', 종성 'ㄱ'으로 표기한다. 무성 파열음을 받침(종성)으로 적거나 다음 음절의 초성으로 적기 위한 구분은 그 앞의 짧은 모음 여부가 판단 기준으로 사용된다. 예를 들어, it[it]는 "잇"으로, beat[bi:t]는 "비트"와 같이 표기한다.

표 4. 포만코드와 한글 자소 대조표 (자음)
Table 4. Mapping table between Phoman code symbols and Korean graphemes (consonant)

종류	포만코드	한글자소
마찰음	S, *, #	ㅅ
	Z, 3	ㅈ
	F	ㅍ
	V	ㅂ
	&	ㄷ
무성파열음	P	ㅍ
	T	ㅌ
	K	ㅋ
파찰음	TS, T#	ㅊ
	D3, DZ	ㅈ
비음	M	ㅁ
	N	ㄴ
	Q	종성 ㅇ
유성파열음	B	ㅂ
	D	ㄷ
	G	ㄱ
유음	L	ㄹ
	R	ㄹ

3.3 예외 처리

포만코드에서 소문자로 표기된 발음은 묵음에 가까운 약한 소리를 의미한다. 발음의 종류에 따라 한글 음차 표기에 반영하는 것이 바람직하기도 하고, 그렇지 않는 것이 바람직하기도 하여 일괄적으로 처리하기 어렵다. 따라서 소문자로 표시된 포만코드를 한글 음차 표기로 변환할 때, 각각의 코드값의 발음을 인정하거나 무시하도록 미리 정의해야 한다. <표5>는 이러한 코드값에 대한 처리 방안을 정의하고 있다.

우리말에 없는 발음 중 하나인 [r]이 초성에 쓰일 때는 'ㄹ'로 변환하면 되지만, 그 외의 경우에는 일관성이 없이, 경우에 따라 'ㄹ'로 표기하거나 묵음으로 두기 때문에 처리하기가 쉽지 않다.

영한 사전은 포만코드로 변환할 때 'R'과 'r'을 구분하여 발음을 표기하고 있으므로, 'R'은 초성 'ㄹ'로 변환하고 'r'은 <표5>에서 정의한 바와 같이 묵음으로 두고 있으나, CMU 발음 사전에서는 'R'과 'r'을 구분하고 있지 않아 음차 표기를 할 때 다음과 같은 문제점이 있다. CMU 발음 기호의 'R'을 모두 포만코드 'R'로 표기하기 때문에 묵음처럼('r') 발음해야 하는 부분에 다음의 예와 같이 '르'를 표기하는 경우가 발생한다.[6]

start :	스타르트*	스타트
party :	파르티*	파티
star :	스타르*	스타

6) '*'는 오류를 의미한다.

표 5. 포만코드에서의 소문자 처리 규칙
Table 5. Rules for lowercase symbols in Phoman code

포만코드	처리 방법	예 / 비고
c	인정	aaron, accidence
d	인정	and, bandbox, grandparent
g	무시	suggest
h	무시	anywhere 단, 첫소리인 경우는 인정 (humiliate, wheelchair)
i	인정	easily
j	인정	endure
k	무시	length
l	인정	palmer
m	무시	ammeter
n	인정	columnist
p	무시	government
r	무시	actioner, arnold
s	무시	disserve, acquaintanceship
t	무시	bankruptcy, nestling
u	무시	biograph
w	인정	quoin, banquo

또한 <표2>에서 정의한 것과 같이 CMU 발음 사전에서는 발음 기호 'ER'을 포만코드 'Cr'로 표기하기 때문에 다음의 예와 같이 음가가 있는 'R'이 무시되는 경우가 발생하기도 한다.

arrival : 어아이벌* 어라이벌
binary : 바이너이* 바이너리

따라서 이러한 오류를 해결하기 위해 다음과 같은 예외 규칙을 둔다.

- 'R'의 경우, 어말이나 자음 앞에서는 무시(묵음처리)한다. 앞 예에서의 "start", "party", "star"가 여기에 해당한다.
- 'r'의 경우, 모음 앞에서는 초성 'ㄹ'로 변환한다. 앞 예에서의 "arrival", "binary"가 여기에 해당한다. 다만, 반모음 [w] 앞에서는 무시한다. (아래 예 참고)

overwash : 오버뤄시* 오버워시

3.4 자소열의 음절 변환

앞에서 설명한 모음처리부와 자음처리부의 출력은 한글 자소열이다. 자소열은 반드시 초성, 중성, 종성의 순서로 연결되어야 하나, 변환 과정에서 이러한 순서에 어긋나는 불완전한 자소문자열이 발견되는 경우는 다음과 같이 처리한다.

- 중성 다음에 초성이 나타나면 종성 채움 코드를 삽입한다.
- 초성 다음에 초성이 나타나면 중성 'ㅡ'와 종성 채움 코드를 삽입한다.
- 종성 다음에 중성이 나타나면 초성 'ㅇ'을 삽입한다.
- 중성 다음에 중성이 나타나면 종성 채움 코드와 초성 'ㅇ'을 삽입한다.
- 초성 다음에 종성이 나타나면 중성 'ㅡ'를 삽입한다.
- 종성 다음에 종성이 나타나면 초성 'ㅇ'과 중성 'ㅡ'를 삽입한다.

초성, 중성, 종성 순으로 이루어진 자소열을 조합형 코드로 변환한 뒤, 이를 완성형 코드로 변환함으로써 완전한 한글 음절열로 변환한다.

표 6. 영·한 음차 표기 결과
Table 6. Results of English-to-Korean transliteration

영어단어	한글음차표기	영어단어	한글음차표기
The	더|디	immediacy	이미디어시
computer	컴퓨터	of	오브|어브
the	더|디	the	더|디
successor	석세서	message	메시지
to	투	transmitted	트랜스미터드
television	텔레비전| 텔러비전	radically	래디컬리
is	이즈	drops	드롭스|드랍스
an	언	audience	오디언스
extended	익스텐디드	attentiveness	어텐티브너스
TV	티브이|티비	as	아스|애스|어즈
medium	미디엄	there	데어
and	앤드|언드	is	이즈
functions	펑크션즈	really	리얼리
essentially	이센셜리	no	노
the	더|디	need	니드
same	세임	to	투
way	웨이	pay	페이
This	디스	attention	어텐션

4. 결과 및 고찰

본 논문에서 구축한 발음 사전은 영한 사전과 CMU 발음 사전을 합쳐서 구성하였다. 영한 사전으로부터 발음 정보를 추출한 단어의 수는 약 7만이고, CMU 발음 사전은 약 13만 개의 단어로 이루어져 있다. 표제어의 수는 CMU 발음 사전이 영한 사전에 비해 매우 많으나, 발음 정보는 영한 사전이 더욱 세밀하고 정확하게 기술되어 있다. 따라서 이 두 사전에 공통으로 수록된 단어는 영한 사전의 발음 정보만을 이용하도록 하였다. 최종적으로 구축된 발음 사전의 크기는 약 17만 단어이다.

음차 표기 사전을 이용하여 영어 문장을 한글 음차 표기한 결과의 예는 <표6>에 있다. 영어 단어가 기분석 사전과 발음 사전에 모두 등록되거나 발음 표기가 둘 이상인 경우에는 하나

의 영어 단어에 둘 이상의 한글 음차 표기가 대응될 수 있다.[7)]

외래어표기법 제3장 제10항의 1에는 “따로 설 수 있는 말의 합성으로 이루어진 복합어는 그것을 구성하고 있는 말이 단독으로 쓰일 때의 표기대로 적는다.”라고 규정하고 있다. 예를 들어, “bookend”는 “book”과 “end”의 합성으로 이루어진 복합어이므로, 외래어표기법에서는 단독으로 쓰이는 “북”과 “엔드”를 붙여 “북엔드”로 표기하도록 규정하고 있다. 그러나 본 논문에서 사용한 사전에는 복합어인지의 여부나 복합어를 이루는 단어들의 경계에 대한 표시가 없기 때문에 정확한 단어의 경계를 알아낼 수 없다. 따라서 현재의 음차 표기 방식으로는 “부켄드”와 같이 결과를 내어준다. 그러나 엄밀히 말하면, 이것은 외래어의 표기와 관련된 문제이지 음차표기와 관련된 문제는 아니다.

5. 결 론

본 논문은 영・한 음차 표기 사전을 구축하기 위하여, 기구축된 두 종류의 사전으로부터 추출한 발음 정보로부터 발음 사전을 구축하고, 발음 사전으로부터 한글 음차 표기로 변환하는 방법을 제안하였다. 발음 정보를 표현하기 위한 포만코드를 정의하고, 포만코드로부터 한글 음차 표기로 변환하는 규칙을 상세히 설계하고 구현하였다.

자동 음차 표기에서 가장 중요한 문제인 발음 추정 문제를 해결하기 위해, 본 논문은 기구축된 사전으로부터 얻은 발음 정보를 음차 표기에 이용하므로 기존의 자동 음차 표기 연구들과는 달리 불완전한 발음 추정 과정이 필요하지 않고 정확한 발음을 알 수 있으므로 정확한 음차 표기가 가능하다.

본 연구에서 구축한 음차 표기 사전은 주로 관용 표기를 수집하고 저장한 기분석 사전과 발음 사전으로부터 변환한 음차 표기 쌍으로 이루어져 있다. 그 자체로 기분석 영한 음차 표기 시스템으로서 활용될 수 있을 뿐만 아니라, 기존 자동 음차 표기 시스템의 전처리 모듈로 사용될 수 있다. 또한 음차 표기를 위한 통계 기반 방식은 영・한 음차 표기 쌍으로 이루어진 많은 양의 학습 데이터가 필요한데, 구축된 음차 표기 사전이 이를 위한 학습 데이터로 사용될 수 있다.

본 연구를 통해 구축된 음차 표기 사전의 크기가 작지는 않지만 사전에 등재되어 있지 않은 단어 즉, 미등록어를 처리하지 못한다는 단점이 있다. 따라서 범용의 영・한 음차 표기 시스템을 구축하기 위해서는 본 연구에서 제안한 음차 표기 사전을 기존의 통계 기반의 음차 표기 모델과 결합하여 사용하는 것이 바람직하다. 향후 연구로는 제안하는 방법이 통계 기반의 음차 표기 모델의 성능 향상에 미치는 영향을 조사하고자 한다.

참 고 문 헌

[1] Kim, T. I. (1999). "English-Korean Transliteration Model Using Maximum Entropy Model for Cross Language Information Retrieval", M.S. thesis, Sogang University.
(김태일, (1999). “최대 엔트로피 모델을 이용한 다국어 정보검색에서의 영-한 음차 표기 모델”, 서강대학교 석사학위논문.)

[2] Lee, H. B. (1979). "On Translitering Loan Words by the Korean Alphabet - A Critical Appraisal of the Revised Koreanization System", *Language Research*, Vol.15, No.1, pp.39-59.
(이현복, (1979). “외래어 표기법 개정 시안의 문제점”, 어학연구, 제15권, 제1호, pp.39-59.)

[3] Lee, J. S. Choi, K. S. (1997). "Automatic Foreign Word Transliteration Model for Information Retrieval, *Proceedings of the Korean Society for Information Management Conference*, pp. 17-24.
(이재성, 최기선, (1997). “정보검색을 위한 외래어 자동표기 모델”, 한국정보관리학회 제4회 학술대회 논문집, pp.17-24.)

[4] Kang, I. H., Kim, G. C. (1999). "English-to-Korean Transliteration using Multiple Unbounded Overlapping Phonemes", *Proceedings of Annual Conference on Human and Language Technology*, pp. 50-54.
(강인호, 김길창, (1999). “복수 음운 정보를 이용한 영・한 음차 표기”, 제11회 한글 및 한국어 정보처리 학술대회 논문집, pp. 50-54.)

[5] Kang, B-J., Choi, K-S. (2000). "Automatic Transliteration and Back-Transliteration by Decision Tree Learning", *Proceedings of the 2nd International Conference on Language Resources and Evaluation*.

[6] Oh, J. H., Choi, K. S. (2005). "An English-to-Korean Transliteration Model based on Grapheme and Phoneme", *Journal of KISS : software and applications,* Vol.32, No.4, pp. 312-326, Apr.
(오종훈, 최기선, (2005). “자소 및 음소 정보를 이용한 영어-한국어 음차표기 모델”, 정보과학회논문지:소프트웨어 및 응용, 제32권 제4호, pp.312-326.)

• **이도길 (Lee, Do-Gil)**
고려대학교 민족문화연구원
서울시 성북구 안암동 5가
Tel: 02-3290-5289 Fax: 02-926-8385
Email: motdg@korea.ac.kr
2008~현재 HK 교수

7) 둘 이상의 음차 표기가 존재하는 경우, ‘|’로 구분한다.