한글-로마자 표기 변환 시스템 구현

김경징 박성현° 최영규 이준환* 이상범
단 국 대 학 교 전자 컴퓨터 공학과, *극 동 대 학 교 정보통신 학부
{jjing75, mzzang°, young, sbrhee}@dankook.ac.kr
*ljh@kdu.ac.kr

Implementation of the Hangul-Roman Conversion System

Kyoung-Jing Kim Sung-Hyun Park°
Young-Kyoo Choi *Jun-Hwan Lee Sang-Burm Rhee

Dept. of Electronics and Computer Engineering, Dankook University
*Dept. of Computer Information, Keukdong University

요 약

본 논문에서는 개정된 국어의 로마자 표기법에 근거한 로마자 표기 변환기를 생성하기 위하여 한글-로마자 표기 변환시스템을 설계하였다.

한글-로마자 표기의 규칙변환을 위하여 로마자 표기법중 표기의 변환에 관련된 항과 그렇지 않은 항으로 분리하여 규칙 변환을 위한 로마자 표기법을 정리하였으며, 로마자 표기법의 근간이 되는 표준 발음법을 페트리넷으로 모델링 후 분석하여 표기-음가 변환표를 생성하고, 표기-음가 변환표에서 로마자 표기법에 해당하지 않는 부분을 제거하여 한글 - 로마자 표기 변환표를 생성하고 이를 바탕으로 한글-로마자 변환 시스템을 구현하였다.

1. 서론

국어의 로마자 표기법[1](이하 로마자 표기법)은 한국어를 외국인이 읽는 것을 전제로 한 표기 규칙으로 국어의 표준 발음법에 따라 적는 것을 원칙으로 한다. 또한 특수 부호를 사용하여 표기하던 것을 로마자 이외의 부호는 사용하지 않는 방향으로 개선되었다[2,3].

로마자 표기법과 함께 발표된 용례사전[3]에 수록된 어휘의 수가 한정되어 한글을 로마자로 변환하는데 많은 어려움이 따랐다. 한글-로마자 표기 변환을 일대일 치환 방법이 아닌 규칙기반으로 처리하기 위해서는 로마자 표기의 바탕이 되는 한글 표준 발음법[4]에 대한 무결한 모델링이 선행되어야 하며, 이를 바탕으로 한 로마자 표기법에 대한 무결한 모델링을 수행하여야 한다.

PetriNet[5]은 그래픽적이고 수학적인 모델링 도구로 여러 가지 시스템에 응용할 수 있다. 정보처리시스템을 기술하고 학습하는데 유망한 도구이다. 그래픽적인 도구로서, PetriNet은 visual-communication을 도입한 플로우차트, 블록다이어그램, 네트웍처럼 사용할 수 있으며 수학적인 도구로서, 상태 방정식, 논리 방정식의 setup이 가능하고, 시스템의 동작을 수학적으로 모델링 할 수 있다[5,6].

페트리넷을 이용한 언어의 모델링을 위하여 모델링 영역의 설정과 모델링 표기를 정의하고, 동적 모델링 방법인 페트리넷 모델의 정적 표현을 위하여 근접행렬로의 변형을 수행하여 한글 표준 발음법과 한글 로마자 표기법을 모델링 할 수 있다[6,7].

본 논문에서는 표준 발음법과 로마자 표기법을 페트리넷으로 모델링하여 규칙변환을 위한 변환테이블을 구성하고 구성된 테이블을 검증하기 위하여 한글-로마자 표기 변환 시스템을 설계 구현한다.

본 논문의 구성은 1장의 서론에 이어 2장에서는 개정된 국어 로마자 표기법의 특징을 알아보고, 페트리넷을 이용한 표준 발음법과 로마자 표기법의 모델링 방법에 대하여 기술한다. 3장에서는 한글-로마자 표기 변환을

위한 변환 테이블 생성하는 방안에 대하여 설명한다. 4장에서는 로마자 표기 변환표를 이용한 한글-로마자 표기 변환기의 구현에 대하여 기술한다. 끝으로 5장에서는 본 연구의 결론과 향후과제에 대하여 논의한다.

2. 로마자 표기법과 페트리넷 모델링

2.1 로마자 표기법

2000년 7월 4일 문화공보부에 의해 개정 공표된 새 「국어의 로마자표기법」의 가장 큰 변화는 로마자를 표기할 때 컴퓨터와 인터넷에서 편리하게 쓸 수 있도록 기존의 매큔라이샤워 표기법을 따를 때 나타나던 반달표(˘)와 어깻점(′)이 없어진 것이다. 주요 변화를 살펴보면 우선 말머리에 오는 우리말 자음 "ㄱ·ㄷ·ㅂ·ㅈ"은 각각 "g·d·b·j"로 표기한다. 또 국어로마자표기의 가장 골칫거리로 지적되던 "어"와 "으"는 논란 끝에 각각 "eo"와 "eu"로 구별한다. 다른 주요 내용을 보면 "ㅋ·ㅌ·ㅍ·ㅊ"은 종전에는 "k′·t′·p′·ch′"로 또한 "ㅓ"와 "ㅡ"를 "ŏ·ŭ"로 했으나 특수부호를 없앤다는 원칙을 따라 "′"부호와 "˘"의 부호도 모두 삭제하였다. 또한 종전에는 "ㅅ"과 모음 "ㅣ"가 어울릴 때는 "sh"로 표기하던 것을 "ㅅ"은 항상 "s"로 적기로 하였다.

<표 1> 개정 전후의 로마자 표기법 비교

모음	ㅏ	ㅓ	ㅗ	ㅜ	ㅡ	ㅣ	ㅐ	ㅔ	ㅚ	ㅑ	ㅕ
개정전	a	ŏ	o	u	ŭ	i	ae	e	oe	ya	yŏ
개정안	a	eo	o	u	eu	i	ae	e	oe	ya	yeo
모음	ㅛ	ㅠ	ㅒ	ㅖ	ㅘ	ㅙ	ㅝ	ㅞ	ㅟ	ㅢ	
개정전	yo	yu	yae	ye	wa	wae	wo	we	wi	ŭi	
개정안	yo	yu	yae	ye	wa	wea	wo	we	wi	ui	
자음	ㄱ	ㄲ	ㅋ	ㄷ	ㄸ	ㅌ	ㅂ	ㅃ	ㅍ	ㅈ	ㅉ
개정전	k/g	kk	k′	t/d	tt	t′	p/b	pp	p′	ch/j	tch
개정안	g/k	kk	k	d/t	tt	t	b/p	pp	p	j	jj
자음	ㅊ	ㅅ	ㅆ	ㅎ	ㅁ	ㄴ	ㅇ	ㄹ			
개정전	ch′	s/sh	ss	h	m	n	ng	r/l			
개정안	ch	s	ss	h	m	n	ng	r/l			

그러나 그 동안 관행적으로 써 오던 인명이나 회사이름, 단체이름은 기존 표기대로 계속 쓸 수 있도록 했다. 세계적으로 이미 널리 알려진 삼성(SAMSUNG)이나 현대(HYUNDAI)와 같은 회사이름이나 김(KIM) 따위의 인명 표기등 이미 널리 알려진 표기는 기존의 표기법과 개정 로마자 표기법 둘 다를 사용할 수 있다. 그러나 종전 표기법에 의하여 설치된 교통 표지판은 2005년 12월 31일까지, 출판물은 2002년 2월 28일 까지 모두 개정하여야 한다.

<표 1>은 개정전과 개정후의 한글 로마자 표기법에 따른 자모의 변환을 비교한 표이다. 반달표와 어깨점으로 인한 컴퓨터를 이용하여 표기할 때의 어려움과 국어에 꼭 필요한 'ㄱ, ㄷ, ㅂ, ㅈ'과 'ㅋ, ㅌ, ㅍ, ㅊ'을 구별하도록 하였다[2,3].

2.2 페트리 넷의 정의

페트리넷에 관한 정의는 <표 2>와 같으며, 페트리넷의 역동적 성질에 대해서 모의 실험을 가능케 하여 주는 트랜지션(transition) 격발(fire)의 정의는 <표 3>과 같다[5,6].

<표 2> 페트리넷의 정의

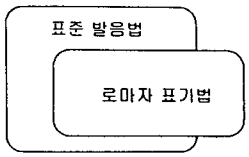
페트리넷 N은 5가지 요소로 구성된다. N=(p,t,f,w, M0) where : p={p1,p2, ... ,pm}은 플레이스라는 유한 집합. t={t1,t2, ... ,tm}은 트랜지션이라는 유한 집합. F: 화살선이라는 (P x T)와 (T x P)의 합 집합의 부분집합. W=F→{1,2,3, ... }는 화살선에 대입되는 가중치 M0:P→{0,1,2,3, ... }는 처음에 각 장소에 놓은 토큰의 수를 표현하며, 초기 marking이라고 함. 단, P와 T의 교집합은 공집합이고, P와 T의 합집합은 공집합이 아님.

<표 3> 격발의 정의

1. 하나의 트랜지션 t에 대하여, 입력 장소 p가 최소한 w(p, t) 만큼의 토큰을 갖고 있으면, t 는 장전되었다고 한다. 2. 장전된 t는 격발될 수도 있고, 아니 될 수도 있다. 3. t의 격발은 w(p ,t)만큼의 토큰을 각 입력 플레이스 p에서 제거하고 w(p ,t)만큼의 토큰을 각 출력 플레이스에 더하여 준다.

2.3 로마자 표기법의 페트리넷 모델링

1) 표준 발음법과 로마자 표기법의 집합관계



(그림 1) 표준 발음법과 로마자 표기법의 집합관계

(그림 1)은 표준 발음법과 로마자 표기법의 포함관계를 나타내는 그림이다. 로마자 표기법 중 1장 표기의 기본 원칙 제1항은 "국어의 로마자 표기는 국어의 표준 발음법에 따라 적는 것을 원칙으로 한다"로 명시되어 있다. 로마자 표기법은 표준 발음법에 기초하여 만들어졌으므로 규칙의 대부분이 발음법에 포함된다. 표준 발음법에서 로마자 표기법을 뺀 차집합 부분은 "종성 ㄹ"과 "초성 ㄹ"이 연속될 때 "rl"이 아닌 "ll"로 적는다 등의 로마자 표기시 예외 사항을 담고 있고 로마자 표기법에서 표준발음 법을 뺀 차집합 부분은 "된소리되기"등 로마자 표기법에서 표준 발음법의 항 중 반영되지 않는 부분을 담고 있다.

2) 페트리넷 구축 방안

(그림 2)의 로마자 표기법 제2장 제2항 붙임1은 파열음에 대한 일반항으로 모음앞에서의 표기와 자음앞이나 어말에서의 표기를 정하고 있다. 'ㅂ'이 "호법"에서 모음앞에 사용된 'ㅂ'은 'b'로 어말에서 사용된 'ㅂ'은 'p'로 사용됨을 의미한다. 이것을 표현하는 페트리넷은 22개('ㄱ, ㄷ, ㅂ 3개, 어말 1, 자음 18)의 입력 플레이스와 6개의 출력 플레이스, 66개의 트랜지션으로 구성된다.

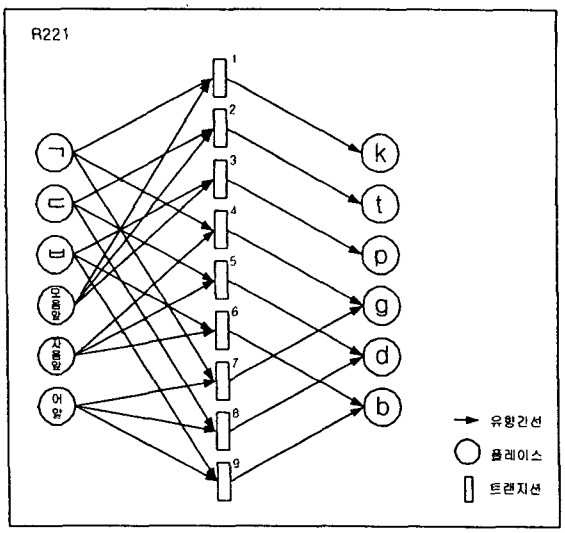
[붙임 1] 'ㄱ, ㄷ, ㅂ'은 모음 앞에서는 'g, d, b'로, 자음 앞이나 어말에서는 'k, t, p'로 적는다.([ ] 안의 발음에 따라 표기함.)

(보기)

구미 Gumi	영동 Yeongdong	백암 Baegam
옥천 Okcheon	합덕 Hapdeok	호법 Hobeop
월곶[월곧] Wolgot	벚꽃[벋꼳] beotkkot	
한밭[한받] Hanbat		

(그림 2) 로마자 표기법의 예(제2장 제2항 [붙임1])

(그림 3)은 (그림 2)의 로마자 표기법을 페트리넷으로 구성한 것이다. 플레이스 ㄱ에서 출발하는 3개의 간선은 각각의 트랜지션과 사상되며, 이 각각은 모음앞, 자음앞, 어말에서 출발한 간선과 함께 트랜지션의 입력 간선이 된다. 각각의 트랜지션이 격발될 조건은 입력 트랜지션에 토큰이 놓이면 각 간선이 활성화되고 이때 트랜지션의 격발조건이 만족되며, 트랜지션의 격발로 출력 플레이스 g, k에 토큰이 놓이게 된다.



(그림 3) 그림 2의 페트리넷 모델링

(그림 3)의 플레이스는 초기 마킹 Mo가 놓여지지 않은 상태이다. 플레이스에 놓여지는 초기 마킹은 입력되는 문자열의 형태소 분석 정보와 자모 분리 결과에 따라 동적으로 놓여진다. 예를 들어 입력 문자열로 "법"이라는 글자가 입력되었을 때 종성 'ㅂ'이 처리되는 과정은 형태소 분석과 자모 분리 결과로 '어말'과 'ㅂ'이 생성되고 각각 해당하는 플레이스에 초기 마킹 Mo가 놓여진다. 입력 플레이스 'ㅂ'과 입력 플레이스 '어말'에 토큰이 놓여지면 트랜지션 1에 연결된 유향간선이 활성화(enable)되고 격발하기 위한 조건이 충족되어 트랜지션 '221-1'이 격발하면 출력 플레이스 'p'에 토큰이 놓여진다.

(그림 3)의 페트리넷은 페트리넷의 구성을 설명하기 위하여 간략히 그린 것으로 플레이스 '자음앞'은 18개의 자음으로 확장되어야 한다.

(3) 표준 발음법 모델의 근접 행렬 변환

(그림 3)과 같은 페트리넷 모델의 통합과 분석을 용이하게 하기 위하여 페트리넷을 근접 행렬(incidence matrix)로 표현한다.

근접 행렬 C는 |P| x |T| 행렬이며, C의 일반항은 다음과 같이 정의한다[7,8,9].

cij = 1 if (tj, pi) ∈ F, -1 if (pi, tj) ∈ F, 0 otherwise

정의에 따라 (그림 3)의 페트리넷 모델을 근접 행렬로 변환하면 <표 4>와 같다. <표 4>는 이해를 돕기 위하여 10항을 간소화하여 모델링한 것으로 입력 플레이스 중 '자음앞'은 18개의 자음으로 확장되어야 한다. 다만 여기서는 근접 행렬의 구성을 설명하기 위하여 '자음앞'으로 나타내었다.

<표 4> 표준 발음법 10항의 페트리넷 근접 행렬

P221①	1②	2	3	4	5	6	7	8	9
Iㄱ ③	-1⑦			-1			-1		
Iㄷ		-1			-1			-1	
Iㅂ			-1			-1			-1
I모음앞	-1	-1	-1						
I자음앞				-1	-1	-1			
I어말④							-1	-1	-1
⑥									
Ok	+1⑧								
Ot		+1							
Op ⑤			+1						
Og				+1			+1		
Od					+1			+1	
Ob						+1			+1

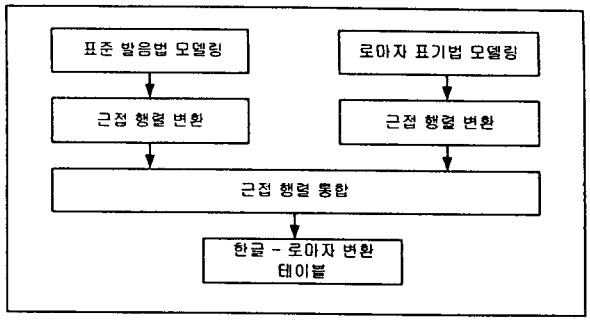
<표 4>의 근접 행렬에서 ①~⑧은 테이블을 설명하기 위한 주석 표시이며 각 주석이 표시하는 내용은 다음과 같다.

① 표준 발음법의 항 번호
② 항의 각 소항목을 나타내는 번호(트랜지션)
③ 장전 조건 1(입력 플레이스1)
④ 장전 조건 2(입력 플레이스2)
⑤ 격발 후 출력(출력 플레이스)
⑥ 각 조건 구분을 위한 공백
⑦ 해당 소항목의 조건부
⑧ 해당 소항목의 결론부

3장. 한글-로마자 변환을 위한 변환 테이블 생성

3.1 로마자 변환 테이블의 생성

(그림 4)는 로마자 변환 테이블을 생성하는 과정을 나타낸 것이다. 로마자 표기법중 실제 표기의 변환 과정과 관계되는 규칙을 모델링하고, 한글 표준 발음법중 로마자 표기 변환에 관계되는 규칙을 모델링 하여 각각을 근접행렬로 변환한다. 근접행렬을 통합하여 두 개의 근접 행렬에서 동시에 나타나는 규칙을 통합하고 각각 나타나는 항을 추가하여 통합된 근접 행렬을 구성한다. 이렇게 도합된 근접행렬을 종성과 초성으로 나누어 배열하여 한글-로마자 변환 테이블을 구성한다.



(그림 4) 한글-로마자 변환 테이블 생성 과정

4장. 구현 및 실험

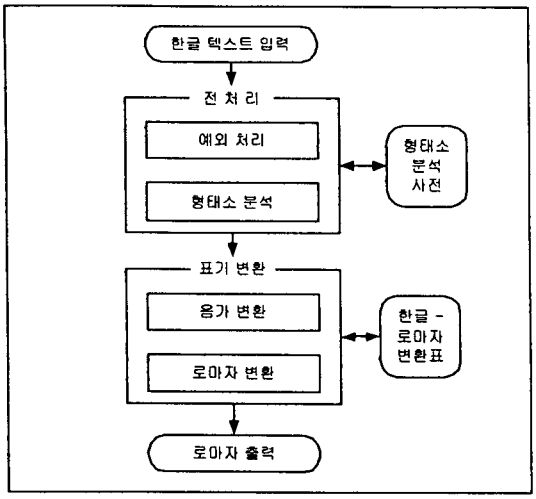
4.1 구현 환경

각 항별로 구성된 페트리넷 근접행렬을 통합하여 로마자 표기 변환표를 생성하는 모듈은 Solaris 2.4 운영 체제 환경과 GCC 컴파일러를 이용하는 ANSI C 언어로 구현하였으며, 생성된 변환표를 검증하기 위한 한글-로마자 표기 변환 시스템은 Pentium-III, Windows98 시스템 상에서 Visual Basic Professional Edition의 Visual Basic 6.0 언어로 구현하였다.

4.2 구현

(그림 5)는 한글-로마자 변환기의 전체 구조를 나타내고 있는 그림이다. 입력된 한글 텍스트는 예외 처리 단계와 형태소 분석의 전처리 단계를 거쳐 표기 변환 모듈로 전달된다. 예외 처리 단계에서는 한글 이외의 텍스트 및 완성되지 않은 한글 등을 예외 처리한다. 형태소 분석 단계에서는 음가 변환 단계에서 필요한 형태소 분석 정보를 생성하여 표기 변환 모듈로 전달된다.

표기 변환 모듈은 음가 변환 단계와 로마자 변환 단계로 구분된다. 한글은 표기와 음가(발음)이 서로 다르기 때문에 표기를 그대로 로마자로 변환시킬 경우 올바른 로마자 표기를 생성 할 수 없다. 음가 변환 단계에서는 한글 표준 발음법중 로마자 표기법에 해당하는 영역만을 모델링 하여 얻은 표기-음가 변환 내용과 로마자 표기법을 모델링 하여 얻은 한글-로마자 변환 표가 통합되어 있으므로 이를 이용하여 한글의 음운 변동 현상을 처리하여 입력된 텍스트를 음가로 변환시킨다.



(그림 5) 한글-로마자 변환 시스템 전체구조

이때 일어나는 한글의 자모 변화는 VisualBasic의 특성상 Unicode를 지원하므로 완성형과 조합형이 변환 없이 Unicode 상태에서 바로 자모의 변형이 일어날 수 있다.

(그림 6) 표기법 예제 실험 결과 화면

로마자 변환 단계에서는 변환된 음가열을 자모 분해하여 한글-로마자 표기 변환표를 바탕으로 로마자 표기 규칙에 따라 로마자로 변경한다.

이때, 음가 변환과 로마자 변환은 두 개의 모듈이 각각 동작하는 것이 아니고, 통합된 표를 이용하여 인접한 두 음절 사이에서 음가 변환이 필요한 경우는 음가 변환과 동시에 로마자 변환을 수행하고 음가 변환이 필요 없는 경우는 로마자 변환만을 수행하는 것으로 이 같은 과정이 음절과 음절의 경계에서 연속적으로 일어나 표기 변환 과정을 완료한다.

(그림 7) 용례사전 예제 실험 결과 화면

4.3 실험 결과

로마자 표기법의 규정에 예제로 나오는 84개의 단어를 대상으로 실험 한 결과 모두 정확한 변환 결과를 보였으며 경부선 철도 역명으로 실험한 결과 모두 정확한 변환이 되었다. (그림 6)은 한국어 로마자 표기법의 규정에 나타난 예제로 실험한 그림이며, (그림 7)은 로마자 표기 용례 사전의 경부선 철도역명을 실험한 그림이다.

(그림 6)에서 옵션 부분은 인명, 고유명사, 일반, 학술 응용, 행정구역 등 사용자가 표기를 원하는 영역에 따른 결과를 로마자 표기를 볼 수 있도록 하였다.

5장. 결론 및 향후과제

5.1 요약

본 논문에서는 한글 로마자 표기법의 규칙 변환을 위한 자연 언어의 수학적 모델링을 위하여 동적 모델링 도

구인 페트리넷을 이용하여 한국어 표준 발음법과 한글 로마자 표기법을 모델링 하였다.

동적 모델링을 정적으로 표현하기 위하여 각 항별 페트리넷 모델을 근접행렬로 변환하여 전체를 통합하여 로마자 표기 변환표 생성 방안을 제시하였다.

본 논문에서 작성된 한글-로마자 표기 변환표를 이용하여 한글 로마자 표기 변환표를 이용한 규칙 기반 한글-로마자 변환 시스템의 구현과 검증하였다.

5.2 결론

본 논문에서는 표기 언어인 한글을 한국어로 발음하는 규칙을 정의한 한국어 표준 발음법과, 외국인이 한글을 읽는 것을 가정으로 한 국어의 로마자 표기법을 페트리넷으로 모델링하고, 모델링된 페트리넷을 근접행렬로 변환하여 통합 후 한글-로마자 표기 변환표의 생성을 보였다.

규칙에 기반한 한글-로마자 변환을 위하여 발음법과 표기법을 모델링하여 로마자 변환에 필요한 규칙을 정형화하고 정형화된 규칙을 무결한 방법으로 다루어 최종 테이블을 작성하였으므로 본 연구에서 생성된 테이블은 무결성을 유지한다고 할 수 있다.

생성된 테이블의 검증을 위하여 국어의 로마자 표기법 고시본에 나타난 예제와 로마자 표기 용례 사전의 예제를 실험하여 올바른 변환 결과가 나타남을 보였다.

5.3 향후과제

본 논문의 목적은 자연 언어의 모델링과 이를 이용한 한글-로마자 표기 변환표의 생성이었다. 한국어는 알타이어족의 언어로서 두음법칙, 형태상 첨가적 성질, 명사에 성(gender)구분이 없음 등의 특징을 가지며 알타이 어족의 언어로는 한국어, 터키어, 몽고, 퉁구스어, 일본어, 핀란드어 등 유럽 동부 지역에서부터 중앙 아시아, 중국 서북부 및 동북부, 몽골, 시베리아 지역에 이르기까지 광범위하게 분포하고 있다. 본 연구에서 제시한 자연언어의 모델링 방법을 다른 알타이 어족의 언어에 적용하기 위한 연구가 요구된다.

참고 문헌

[1] 문화관광부, "국어 로마자표기법", 문화관광부 고시 제 2000-8호, 2000.

[2] 문화관광부, 국립국어연구원, "로마자표기법 이렇게 바뀌었습니다", 문화관광부, 2000.

[3] 문화관공부, 국립국어연구원, "로마자 표기 용례 사전", 동화서적, 2001.

[4] 문화교육부, "표준어 규정", 문교부 고시 제 88-2호, 1988.

[5] T.Murata, "Pertri nets :properties, analysis and applications," Proceeding of the IEEE, Vol77. no.4, pp.541-580, 7 April 1989.

[6] 임재걸, 이계영, "한글 받침 발음법의 페트리 네트 표현", 동국대학교 동국논집, 14집, pp. 155-167, 1995.

[7] 임재걸, 이계영, 김경징, "페트리넷을 이용한 표준 발음법 분석 시스템 디자인", 한국정보과학회 봄학술발표논문집, pp. 369-371, 1999.

[8] 임재걸, 이계영, 김경징, 김규식, "페트리넷을 이용한 표준 발음법 분석 시스템 구현", 한국정보처리학회 춘계학술발표논문집, pp. 609-612, 1999.

[9] 임재걸, 이계영, 김경징, "표준 발음법 페트리넷을 이용한 음운 변환기 설계", 한국멀티미디어학회 춘계학술발표논문집, 제2권 제1호, pp. 339-344, 1999.