Creation of the Conversion Table from Hangeul to the Roman Alphabet

KyoungJing Kim[1] and SangBurm Rhee[2]
[1] Department of Electronics and Computer Engineering,
Graduate School of Dankook University
San8, HanNam-Dong, YongSan-Gu, Seoul, Korea
Tel. +82-2-799-1102, Fax.: +82-2-797-5857
[2] Department of Electrical and Electronic and Computer Engineering, Faculty of Engineering,
Dankook University
San8, HanNam-Dong, YongSan-Gu, Seoul, Korea
e-mail : jjing@dankook.ac.kr, sbrhee@dankook.ac.kr

Abstract: For a rule-based conversion of Hangeul into the Roman alphabet rather than a word-for-word conversion, one must come up with a faultless model for the Korean standard pronunciation rules, which are the basis of the Romanization. It is on this foundation that the Korean-Roman alphabet conversion table can be created.

For linguistic modeling using PetriNet, modeling boundary and notation of modeling can be defined. In order to describe PetriNet, which is a dynamic modeling tool, as a static one, one can model the standard Korean pronunciation rules and the Hangeul-Roman alphabet notation by conversion into incident matrix.

Thus, this research attempts to develop a mathematical modeling tool for a natural language using PetriNet, and create a Korean-Roman alphabet conversion table

1. Introduction

The Korean language is composed of a spoken language, Korean, and a phonetic language, Hangeul. The notation rule of the Korean language using the Roman alphabet is for foreigners reading Hangeul, and it is based on the standard pronunciation rules of Hangeul[1,2].

For a rule-based conversion of Hangeul into the Roman alphabet rather than a word-for-word conversion, one must come up with a faultless model for the Korean standard pronunciation rules which are the basis of the Romanization. It is on this foundation that the Korean-Roman alphabet conversion table can be created.

The Korean language contains characteristics of both a phonetic language and a syllabic language. A modeling tool for the two languages can be developed through a mathematical modeling tool on Hangeul. Thus, this research attempts to develop a mathematical modeling tool for a natural language using PetriNet, and create a Korean-. Roman alphabet conversion table.

Exceptions in standard Korean pronunciation and standard Roman alphabet notation are for specific vocabulary or phrases. They are not considered in this research since they do not follow the general rules.

PetriNet[3] is a graphic and mathematical modeling tool which can be applied to various systems, and it is especially useful in describing and studying the information processing system[3-5]. As a graphic tool, it can be used in flow charts inducing visual communications, block diagrams, and networks. As a mathematical tool, it can set up state and logic equations, and mathematical modeling of system operation is also possible[6-10].

For a linguistic modeling using PetriNet, we will attempt in this paper to set the domain of the modeling and define the notation of the modeling. In order to present the PetriNet model, which is a dynamic modeling tool, as a static one, we will develop a model on the standard pronunciation rules and the Hangeul-Roman alphabet notation by conversing into incident matrix. A Hangeul-Roman alphabet conversion table will be created based on the modeling results.

2. Standard Pronunciation Rules and Roman Alphabet Notation using PetriNet

2.1 Boundary of the Standard Pronunciation Rule Modeling

(1) Relationship between the standard pronunciation rule and the Roman alphabet notation rule

The Hangeul-Roman alphabet notation is based on the standard Korean pronunciation rule[2]. This paper will explain the linguistic modeling and the creation of the Hangeul-Roman alphabet conversion table using PetriNet and give examples from the standard pronunciation rule that are more important in conversion of the Roman alphabet notation.

Diagram 1 shows the relationship between the standard pronunciation rule and the Roman alphabet notation rule[11]. The standard pronunciation rule is the base of the Roman alphabet notation rule, so most of the latter belong to the pronunciation rule[12-14]. The complementary set area contains exceptions in the Roman alphabet notation. For example, when the last sound "ㄹ[ri-eul]" is followed by the initial sound "ㄹ[ri-eul]", the sound should be not "rl" but "ll".

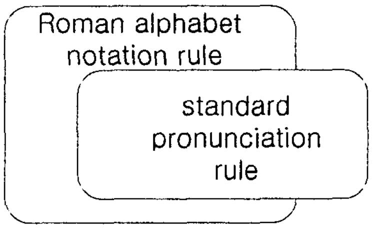

Diagram 1. Relationship between the standard pronunciation rule and

the Roman alphabet notation rule

(2) Modeling Boundary of Terms in the Standard Pronunciation Rule and Modeling of Character Change

Table 1. Modeling Boundary of Terms and of Characters

Term	Boundary
Monophthong	ㅏ ㅐ ㅓ ㅔ ㅗ ㅚ ㅜ ㅟ ㅜ ㅣ (10)
대 표 음	※< Reference [11] >
Fortis	ㄲ, ㄸ, ㅃ, ㅆ, ㅉ(5)
Consonant	ㄱㄲㄴㄷㄸㄹㅁㅂㅃㅅㅆㅈㅉㅊㅋㅌㅍㅎ(18)
Vowel	ㅏ,ㅐ,ㅑ,ㅒ,ㅓ,ㅔ,ㅕ,ㅖ,ㅗ,ㅘ,ㅙ,ㅚ,ㅛ,ㅜ,ㅝ, ㅞ, ㅟ,ㅠ,ㅡ,ㅢ, ㅣ(21)
Consonant under Vowel	ㄱㄲㄳㄴㄵㄶㄷㄹㄺㄻㄼㄽㄾㄿㅀㅁㅂㅄㅅ ㅆㅇㅈㅊㅋㅌㅍㅎ(27)
Twin Consonant under vowel	ㄲ, ㅆ (2)
Double Consonant under Vowel	ㄳ,ㄵ,ㄶ,ㄺ,ㄻ,ㄼ,ㄽ,ㄾ,ㄿ,ㅀ,ㅄ(11)
Regular Sound	ㄱ ㄷ ㅂ ㅅ ㅈ ㅎ (6)
Double Vowel	ㅑ,ㅒ,ㅕ,ㅖ,ㅘ,ㅙ,ㅛ,ㅝ,ㅞ,ㅠ,ㅢ(11)

In reference[11], using definitive term it is set a standard for standard pronunciation rule. However in process of modeling it is not set bounds of modeling according to the term when modeling based on the definitive term. So we set limits of modeling such as <table 1>

2.3 Definition of Modeling Notation

Table 2. Notation for Input and Output

Input	
Symbol	Meaning
I{C, V, J}	Input{first, middle, last sound}
I{M, MJ, MCV}	Input {Grammar Morpheme (GM), GM-last, GM-first and middle}
-1	enable input arc
Output	
Symbol	Meaning
O{C, J, N}	Output{first, last, no sound}
+1	Enabling output arc
*1	Enabling only one output arc

There are three cases where a character changes; vowel change, consonant change and legnth change. <Table 2> is summarized the changes acknowledged in the standard pronunciation rule and modeling notation.

Table 2.4. Modeling Notation of Character Change

Examples of Change	Modeling notation
젖[젇], 솥[솓], 넋[넉]	OJ
각하[가카], 좋던[조턴]	OJNC→OJN, OC
낳은[나은], 놓아[노아]	OJN
넋이[넉씨], 앉아[안자],닭을[달글]	OJC→OJ, OC

Category	Items	Reason
Not Related	1, 2, 3, 4, 5, 12.1, 17, 18, 21, 27	No direct relation or optional
Prearranged Exception	6, 7, 10(2), 16, 20, 26, 29	Exception

2.4 PetriNet Modeling of the Standard Pronunciation Rule and the Roman Alphabet Notation

(1) How to create PetriNet

Clause 10 of the standard pronunciation rule[1] *shown in* Diagram2.4 distinguishes the general case from the special case concerning '밟-' and '넓-'. The general case indicate that 'ㄳ' is pronounced "ㄱ" at the end of a character and in front of a consonant, and that "ㄵ" is pronounced "ㄴ" at the end of a character and in front of a consonant

Clause 10. Double consonants under vowels - 'ㄳ', 'ㄵ', 'ㄼ, ㄽ, ㄾ', 'ㅄ' – are pronounced as ㄱ, ㄴ, ㄹ, ㅂ at the end of a character and in front a consonant.

넋[넉] 넋과 [넉꽈] 앉다 [안따] 여덟 [여덜]
넓다[널따] 외곬[외골] 핥다[할따] 값[갑] 없다[업따]

However, '밟-' is pronounced '밥' in front of a consonant, and '넓-' is pronounced '넙' in the following cases:
(1) 밟다 [밥-따] 밟소 [밥-쏘] 밟지 [밥-찌]
밟는 [밥-는]→[방-는] 밟게 [밥-께] 밟고 [밥-꼬]
(2) 넓-죽하다 [넙쭈카다] 넓-둥글다 [넙뚱글다]

Diagram 2. Clause 10 of the Standard Pronunciation Rule

Examples of the special case is that '밟다' should be pronounced '밥[bap]' according to the special case rule even though it looks like it has to follow the general case rule. Likewise, "넓" should be pronounced '넙[neop]' if the word falls into the special rule category. Such exceptions are handles separately in the exception dictionary

Diagram 3 shows the PetriNet model combining transitions 1001 and 1002, and 1007 and 1008 in Clause 10.

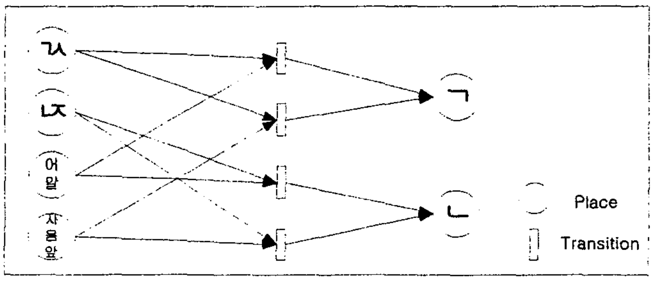

Diagram 3 PetriNet combining transitions 1001, 1002, 1007, 1008 in Clause 10

(2) Incidence Matrix Conversion of the Standard Pronunciation Rule Model

PetriNet is converted into the incidence matrix in order to make the combination and the analysis of PetriNet models shown in Diagram 2.3.

The incidence matrix C is |P| x |T| matrix, and the general cases of C are defined as follows: [3,17]

$$c_{ij} = 1 \text{ if } (t_j, p_i) \in F, -1 \text{ if } (p_i, t_j) \in F, 0 \text{ otherwise}$$

When the incidence matrix of Diagram 3 is converted according to the definition, it will look like Chart 1. Chart 1 has modeled Clause 10 after simplification to help the readers' understanding, and one of the input places, "in front of a consonant" should be extended to eighteen consonants that are included in the terms and boundaries defined in Table 2. Here, however, it is indicated as "in front of a consonant" to explain the structure of the incidence matrix. In the incidence matrix in Chart 1, ①~⑧ is symbols of notes explaining the table, and each note signifies the following:

① Clause number in the standard pronunciation rule
② Detailed clauses in each clauses (Transition)
③ Input condition 1 (input place 1)
④ Input condition 2 (input place 2)
⑤ Output after fire (output place)
⑥ Blank spaces for distinction of each condition
⑦ conditioning part of relevant detailed clauses
⑧ concluding parts of relevant detailed clauses

Chart 1. Incidence Matrix of Clause 10 of the Standard Pronunciation Rule

R10①	1001②	1002	1003	1004	1005	1006	1007	1008	1009
IJ ㄳ③	-1 ⑦						-1		
IJ ㄵ		-1						-1	
IJ ㄼ			-1						-1
IJ ㄽ				-1					
IJ ㄾ					-1				
IJ ㅄ						-1			
⑥									
IM-e④	-1 ⑦	-1	-1	-1	-1	-1			
IM-f							-1	-1	-1
⑥									
OJ ㄱ⑤	+1⑧						+1		
OJ ㄴ		+1						+1	
OJ ㄹ			+1	+1	+1				+1
OJ ㅂ						+1			

* Ime : IM ending, Imf- IM-in front of a consonant

3. Creation of the Hangeul-Roman Alphabet Conversion Table

3.1 Combination of the PetriNet Models

A combination of tables is necessary to combine the incidence matrices created for each item. The incidence table consists of four parts: transition numbers, the input place 1, the input place 2, and the output place. To combine each item, the first and the second item out of all the items intended for combination are compared, and elements in the second item that are already in the first item have markings "-1, +1" and a transition number at the end. When elements that are not in the first item appear, the new elements as well as the markings "-1, +1" and transition numbers are added. The next table is added to this table through the same method, and in the end a combination table is created.

3.2 Compression of the Combined Table

In the combined PetriNet incidence matrix of the standard pronunciation rule, the item number causing a phonological change is recorded as a label of a transition. Here, the transition numbers, i.e. the item numbers causing phonological change, are unnecessary data for phonological changes. These transition numbers can be omitted since they are not needed in the rule of phonetic value change. The number of the input place 1 and the input place 2 that are connected to transitions in PetriNet modeling is one, but the number of the output place can be two because sometimes only the initial sound changes and other times both the initial sound and the last sound change. Using property like this, numbering after removing the duplication of all labels possibility of appearance in the label of input place 1 and input place 2. These are used the first and the second item of two dimension matrix. Compact the table after recording in case of changing an initial sound and changing a final consonant in appearance of output place.

3.3 Transration of romanization table

Chart 2. A part of Hangeul-Roman alphabet Conversion table

	0	1	2	3		17	18	19		24	25	26	비고
0	-	-	-	-	...	-	-	-	...	-	-	-	null
1	k	k	ng	k	...	k	-	k	...	-	-	k	ㄱ
2	k	k	ng	k	...	k	k	k	...	-	k	k	ㄲ
3	k	k	ng	k	...	k	k	k	...	k	k	k	ㄳ
12	l	l	l	l	...	l	l	l	...	l	l	l	ㄽ
13	l	l	l	l	...	l	l	l	...	l	l	l	ㄾ
14	p	p	m	p	...	p	-	p	...	l	-	p	ㄿ
15	l	l	l	l	...	l	l	l	...	l	l	l	ㅀ
16	m	m	m	m	...	m	m	m	...	-	m	m	ㅁ
21	ng	ng	ng	ng	...	ng	ng	ng	...	ng	ng	ng	ㅇ
22	t	t	n	t	...	t	-	t	...	-	-	t	ㅈ
23	t	t	n	t	...	t	-	t	...	-	-	t	ㅊ
24	k	k	ng	k	...	k	-	k	...	-	-	k	ㅋ

The chart is part of the Korean alphabet to Rome conversion table for last sound. The length axis means the sub-code of Hangeul Uni-Code for last sound of front letter. Horizontal axis means the sub-code of Hangeul Uni-Code for last sound of front letter. The alphabet of the place where horizontal axis and the length axis intersect, By a front letter end sound and a rear letter first sound the Rome letter which is changed . Rome letter is that all change actual condition of the last sound which occurs from between last sound and the first sound of the Korean alphabet is contained. "-" is recording the Korean alphabet in Rome letter, by a phonological change actual condition it shows the letter which does not occur..

4. Conclusion and Future Research Areas

4.1 Summary

In this paper, we attempted modeling of the standard Korean pronunciation rule and the Hangeul-Roman alphabet notation using PetriNet, a dynamic modeling tool.

To express the dynamic modeling as static one, we converted each PetriNet model into an incidence matrix and then combined all the incidence matrices. The Hangeul-Roman alphabet conversion table can be created by deleting and compressing transition information from in the *combined incidence matrix.*

Using the modeling, combination and compression processes, the Hangeul-Roman alphabet conversation chart based on the standard Korean pronunciation rule and the Hangeul-Roman alphabet notation rule could be created.

4.2 Conclusion

1) Modeling of a Language

In this paper, *we set a modeling boundary and notation* for the mathematical modeling of a natural language, and PetriNet was used in the modeling. The PetriNet modeling was for the standard Korean pronunciation rule which provides guidelines for pronouncing Hangeul, a phonetic language, as a spoken language, and for the Hangeul-Roman alphabet notation rule which helps foreigners read Hangeul. We have then suggested ways to combine and compress the PetriNet models after converting them into incidence matrices.

2) Hangeul-Roman alphabet Conversion

For rule-based Hangeul-Roman alphabet conversion, we have come up with a pronunciation rule model and a notation rule, standardizing rules needed in the Roman alphabet conversion. After, the standardized rules are processed in *faultless ways to provide a final table.* Therefore, the table introduced by our research can be said to maintain its faultlessness.

4.3 Future Research Areas

The objective of this paper was a natural language modeling and the creation of the Hangeul-Roman alphabet conversion table using the modeling. The Korean language is an Altaic language, and has unique characteristics such as the initial sound rule, The characteristic of an agglutinative language, and no gender distinction in nouns. Many languages including Korean, Turkish, Mongolian, Tungus, Japanese and Finnish belong to the Altaic language family, and *it also has a wide distribution in East Europe*, Mongolia, Central Asia, southwestern and northeastern China and Siberia. A further research is called for to apply the method of a natural language modeling introduced in this research to other languages in the Altaic language family. In addition, system materialization and verification of the rule-based Hangeul-Roman alphabet conversion system is necessary, using the Hangeul-Roman alphabet conversion chart introduced in this paper.

References

[1] Ministry Of Culture & Education , *"Standard Pronunciations"*, Ministry Of Culture & Education Notification No.88-2, 1988.

[2] Ministry Of Culture & Tourism , *"Guidelines to the Romanization of Korean"*, Ministry Of Culture & Tourism Notification No.2000-8, 2000

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

[4] C.A. Petri, *"Communication with Automata."*Newyork:Griffiss Air Force Base. Tech.Rep.RADCTR-65-377, voll, Suppl. 1, 1996.

[5] F. Commoner, *"Deadlocks in Petri nets,"* Wakefield, Applied Data Research, Inc., Report #CA-7206-2311,1972.

[6] J.L Peterson, *Petri Net Theory and the Modeling of Systems.* Englewood Cliffs, NJLPrentice-Hall, Inc.j 1981.

[7] K.Voss, H,J.Genrich, and G.Rozenberg, Eds. Concurrency and Nets, Special volume in the series *"Advences in PetriNets."* New York : Springer-Verlag,1987.

[8] C.A Petri, *"Concurrency as a basis if systems thinking,"* St, Augustin Gesellschaft fur Mathematik und Datenverarbeitung *Bonn, Interner Bericht ISF-75-02.1 ,1978.*

[9] E.W.Mayr, *"An algorithm for the general Petri net reachability problem"*SIAM, J. Comput. Vol 13, no.3 pp.441-460, Aub.1984.

[10] K. Lautenbach, "Liveness in Petri nets," St, Augustin Gesellschaft fur Mathematik und Datenverarbeitung Bonn, Interner Bericht ISF-75-02.1 ,1975.

[11] Jaegeol Yim, Gyeyoung Lee, Kyoungjing Kim, *"A design of an analysis system for pronunciation natural Korean Grammar using Petri-Net"*, Proceeding of KISS Spring, pp. 369-371, 1999.

[12] Jaegeol Yim, Gyeyoung Lee, Kyoungjing Kim, GyuSik Kim, *"An Implementation of an Analysis System for Pronunciation Natural Korean Grammar Using Petri-Net"*, Proceeding of KIPS Spring, pp. 609-612, 1999.

[13] Jaegeol Yim, Gyeyoung Lee, Kyoungjing Kim *"A Design of a Phonology Translator Using Pronunciation Natural Korean Grammar Petri-Net"*, Reprocessing of dddddd Spring. Vol. 2, *No.1, pp. 339-344, 1999.*

[14] Gyeyoung Lee, Jaegeol Yim, Kyoungjing Kim, *" Consistency Verification and Priority Decision of Standard-pronunciation-rule "*, Proceeding of KIPS, Autumn, 1999.

[15] Gyeyoung Lee, Jaegeol Yim, Kyoungjing Kim, *" Compression of the Rules of Producing Phonetic Values "*, Proceeding of KISS, Autumn, 01-203 , 1999.

[16] Gyeyoung Lee, Jaegeol Yim, Kyoungjing Kim, *" Reprocessing of Compressed Rule Table of Producing Phonetic Values "*, Reprocessing of STEC, pp.170-174, 1999.

[17] Jaegeol Yim, Gyeyoung Lee, *" Hangeul consonant under a vowel Representation with Petri net "*, Dongguk Univ. Vol. 14., pp. 155-167, 1995.