동국대 경주 변정용 '92.10.

1992년도 제 4 회 한글 및 한국어 정보처리 학술발표 논문집

훈민정음 창제 원리와 한글코드 제정 원리 : 자소형 제안

1992. 10. 9

변 정 용

동국대학교 자연과학대학
전자계산학과

차 례

정보교환용 한글 코드의 관찰

○ 용　도 : 시스템과 시스템 또는 시스템과 주변 장치간 정보교환시 사용

○ 현행표준 : KS C5601-1987 --- 한글 2350, 한글 4888
-보조자판 : KS C5657-1989 --- 한글 1930, 옛글 1673
-표현형식 : 두바이트 코드계(8836자), 완성형 글자
-국제규격 : ISO 646IRV, ISO 2022-1983

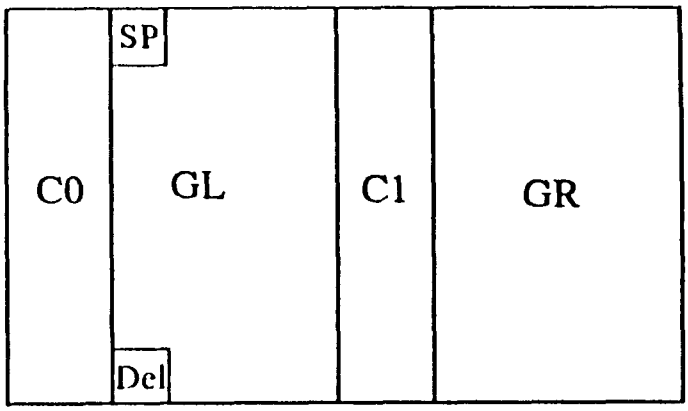
○ 긍정적 평가
- 단순 자료처리 응용에 무난
- 국내 표준화, 국제 표준화 수용
- 두바이트 코드계 소프트웨어 적용 쉬움(한자권)

○ 부정적 평가
- 한글 표현의 제약 --- 4/5불가, 2350/(11172+옛한글)
- 한국어 정보처리 응용의 제약 --- 자소정보 없음
- 정보교환방식의 이중성 --- 두바이트, 필코드
- 한 바이트 코드계 S/W 적용 어려움(ASCII권)
- 보조자판으로 알고리즘 복잡화

○ 원인
- 한글공학을 위한 한글과학 연구의 취약 --- 한글문자의 문자론 및 과학성의 오해 --> 완성형 집착

○ 해결 => 훈민정음 창제 원리로 부터 ...

정보교환용 미국표준코드(ASCII)



In-Use Table

- Nibble (4bits) : 크기 16
- BCD(6bits) : 크기 64
- EBCDIC(8bits) : 크기 256
- ASCII(7/8bits) : 크기 128/256
- NAPLPS : North American Presentation-Level-Protocol Syntax (for vediotex)

국제표준 규격

국제규격	내 용
ISO R 646-1967 ISO R 646-1976 ISO 646 IRV	정보교\환용 6,7비트 문자셑 정보교환용 7비트 문자셑
ISO 2022-1983	7,8비트계 코드 확장법 단수 바이트 94 두 바이트 94X94 = 8836 N 바이트 94^n
ISO 10646-1989 ISO 10646-1.2-1992 ISO 10646-1.2(개정)	4 Octets(bytes),제어문자 지역 배제 (256X256X256X256) UCS-4 4옥테트 정규형 명세 UCS-2 2옥테트 국제공용 글자판 (256 X 256 = 65536) 제어문자 지역 사용 유니코드와 완전통합
UNICODE	정보교환용 다국어 부호계 두 바이트계 (256 X 256 = 65536) 제어문자 지역 사용 Apple, MicroSoft, IBM, DEC등 참여

한글 코드의 기본 개념

○ 자소 --- 초성, 중성, 종성 (음소)
글자 --- <초성,중성,종성> (음절)
문자 --- 일반 호칭

○ 완성형 코드(글자/음절)
- 두 바이트 코드 : 94 X 94 = 8836
예) KS C 5601-1987, KS C 5619-1982
- 글자집단 : 한글 2350(빈도 0.001%), 한자 4888자

○ 조합형 코드(글자, 자소)
- 두 바이트 코드 : 1빝 + 5빝 + (2+3)빝 + 5빝
예) KS C 5601-1982
- 초성 19 X 중성 21 X 종성 28 = 11172자

○ 자모형 코드 : 낱자모형, 겹자모형
- 단수 바이트 : 94
예) KS C 5601-1974, 33자-자모형[변정용1989]
- 겹자모형 : 자음 30(4,5열)+모음 21(6,7열)=51자

○ 자소형 코드 : 낱자소형, 겹자소형 (옛한글 포함)
- 단수 바이트 : 94
예) KS C (없음), ISO 10646-1.2의 개정판
- 낱자소형 : 초성 17 + 중성 11 + 종성 17 = 45자
- 겹자소형 : 초성 91 + 중성 67 + 종성 82 = 240자
***'92년 3월 국내 ISO/IEC JTC1/SC2/WG2회의에서
변정용 정음형(자소형) 제안 한국안으로 채택

한글 코드 표준화 고찰

한글 부호계	내 용	가나 부호계
KS C 5601-1974	한글자모 51 도형문자열 2,3,4,5열	JIS C 6220
KS C 5620-1977	7단위 부호확장법	JIS C 6228
KS C 5714-1977	한자 7200자 (가나식 배열순서)	JIS C 6226
KS C 5619-1982	한글 1316,한자 1692 완성형 (94x94=8836)	JIS C 6226
KS C 5601-1982	2바이트 조합형 (32x32x32=32768) 19x21x28 = 11172자	없음
KS C 5601-1987	한글 2350자, 한자 4888자 완성형 (94x94=8836)	JIS C 6226
KS C 5657-1989	한글 1930,옛한글 1673 완성형 (94x94=8836)	

* 단수 바이트계 --> 복수 바이트계
** 자모형 --> 조합형 --> 완성형
*** < 훈민정음 > --> <구결, 이두>

훈민정음 창제 원리

훈민정음 요구도

○ 국민교화 (주체적 우리 말 표기체계의 확립)
○ 한자표기통일 (한자 표기체계의 표준화)
○ 선진문화섭취
○ 용비어천가 편찬

(문헌 인용)

- 천지자연의 소리가 있으면 곧 반드시 천지자연의 글이 있다.(해례본 정인지서)
- 사람으로 하여금 쉽게 익혀, 쓰기 편하게 하고져 한다. (예의본)
- 슬기로운 이는 아침전에 어리석은 이라도 열흘이면 깨우친다.(해례본 정인지서)
- 무슨(바람, 학울음, 개 짖음) 소리라도 적을 수 있다.(해례본 정인지서)
- 전환이 무궁하고, 간단 요긴하며, 정밀하고도 잘 통한다.(해례본 정인지서)
- 제자해, 초성해, 중성해, 종성해, 합자해(해례본)

훈민정음 구조 원리

○ 제자해 : 음양오행에 의한 글자 만드는 이치

○ 초성해 : ㄱ ㄴ ㄷ ㄹ ㅁ ㅂ ㅅ ㅿ ㅇ ㆁ ㆆ ㅈ ㅊ ㅋ ㅌ
ㅍ ㅎ (17자)

○ 중성해 : ㅏ ㅑ ㅓ ㅕ ㅗ ㅛ ㅜ ㅠ ㅡ ㅣ(11자)

○ 종성해 : 내중소리에는 다시 첫소리를 쓰느니라.
(終聲復用初聲)

○ 합자해 : 글자 구성법

● 연서법 : ㅇ 을 순음 아래 이어 쓰면 입술 가벼운 소리가 되느니라.(ㅱ ㅸ ㆄ ㆊ)

● 병서법 : 초성자를 어울려 쓰면 나란히 쓰라.
(각자병서:ㄲ ㄸ ..., 합용병서: ㅺ ㅽ...)

● 합용법 : 중성자의 합용은 왼쪽 부터 쓰라.

● 부서법 : 가로꼴 모음은 초성 아래 붙여 쓰고,
세로꼴 모음은 오른녘에 붙여 쓰라
종성을 쓰려면 초성과 종성에 잇대어 적는다.

● 성음법 : 무릇 모든 글자는 합하여야만 소리를 이루나니.(凡字必合而成音)

훈민정음 해석 : 문자론

○ 표음문자 : 음소문자 : 로마자
음절문자 : 가나, 한자
음소.음절문자 : 한글

○ 한글의 표현 :
. 일차원 표현 --- 음소문자
. 이차원 표현 --- 음절문자, 가독성(Readability)

4층	글자의 구성	부서법, (합자)성음법
3층	자모의 확장	연서법, 병서법, 합용법
2층	기본 음소	소리의 거셈, 발음형태에 따른 확장 (초성 17자, 중성 11자, 종성 17자)
1층	원소 음소	초성 5자 ㄱ ㄴ ㅁ ㅅ ㅇ (아설순치후) 중성 3자 . ㅡ ㅣ(천지인)

그림 3-1 한글문자의 계층구조

훈민정음 해석 : 집합론

[정의] 초성, 중성, 종성은 유한집합이며, C1, V, C2라 한다.
C1 = {ㄱ ,ㄴ ,ㄷ ,ㄹ ,ㅁ ,ㅂ ,ㅅ , ㅿ ,ㅇ , ㆁ ,ㅎ ,ㅈ ,ㅊ ,ㅋ ,ㅌ ,ㅍ ,ㆆ }
V = { ㆍ, ㅏ, ㅑ, ㅓ, ㅕ, ㅗ, ㅛ, ㅜ, ㅠ, ㅡ, ㅣ}
C2 ≡ C1 (종성부용초성)

[정의] Closure : $C1^*(V^*,C2^*)$
=> C1(V,C2)의 원소로 구성된 모든 문자열 집합

[정의] Concatenation : x.y 또는 xy
$x = a_1a_2a_3...a_n$, $y = b_1b_2b_3...b_n$에 대하여
$x.y = a_1a_2a_3...a_nb_1b_2b_3...b_n$

[정의] 집합 K : C^*의 부분집합,
K^*(Closure), K^+(Positive Closure)는?

$$K^* = \cup K_n = K_0 \cup K_1 \cup K_2 \ldots$$

$$K^+ = \cup K_n = K_1 \cup K_2 \cup K_3 \ldots$$

(단, $K_0 = \{\Lambda\}$, $K_{n+1} = K_n . K$)

따라서 연서법-병서법(C1) => $C1^+$
합용법(V) => V^+
$C1^+ \equiv C2^+$, $\{\Lambda\} + C2^+ => C2^*$

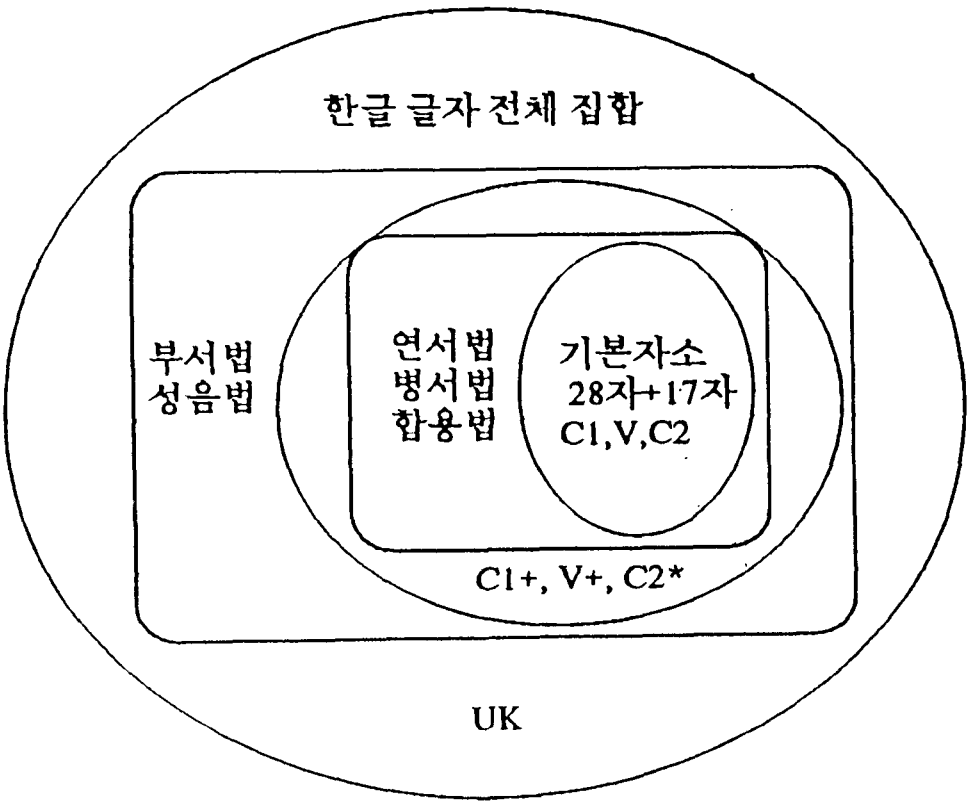
* 불완전한 글자 => $C1^*, V^*$
** 해례본의 "무궁", "무슨" ==> 무한, 불확정

[정의] <a1,a2, ... , an>를 n-순서쌍(Ordered n-tuple)

[정의] {K_1,K_2, ..., K_n}은 K_i, n>= i >0인 첨자집합의 집합,
곱집합(Catesian Product) $K_1xK_2x,...,xK_n$은
n-순서쌍 집합 {<k_1,k_2,...,k_n> | $k_i \in K_i$}이다.
따라서, 한글 글자 전체 집합 UK

$$UK = \{<c1,v,c2> \mid c1 \in C1^{+} \wedge v \in V^{+} \wedge c2 \in C2^{*}\}$$

● 글자집합 표현 : 원소나열법 아닌, 조건제시법
=> 훈민정음 요구도 만족



훈민정음 해석 : 지식 공학

○ 훈민정음의 구현 방법

○ 한글 글자 지식베이스 (HCKB) :
- 사실베이스(Fact Base) : C1, V, C2
- 규칙베이스(Rule Base) : 연서법, 병서법, 합용법, 부서법, 성음법

○ 글자의 생성 (≡언어 생성)
- 유한수의 자소로써 무한수의 글자 생성
(유한수의 어휘로써 무한수의 문자 생성)

한글 부호계의 요구도

- 훈민정음 구조 원리 만족 : 문자론, 집합론
- 한국어 정보처리 요구(자소 정보)의 만족
- PC에서 슈퍼컴퓨터까지 일관된 적용성
- 알고리즘의 효율성 제고
- 국제 표준(ISO 10646)에 적합성 : 문자집단의 크기
- 코드의 호환성 : super-set

자소집단의 선정

[제1안] 초성 17, 중성 11, 종성 17 = 45자
[제2안] 초성 22 : 17 +.(ㄲ ㄸ ㅃ ㅆ ㅉ),
중성 15 : 11 + (ㅒ ㅖ ㅙ ㅞ),
종성 19 : 17 + (ㄲ ㅆ) = 58자

...
[제K안] 초성 91, 중성 67, 종성 82 = 240자

...
[제N안] ???

○ [제 K,N 안]의 문제
- 훈민정음 구조 원리에 벗어난다.---조건제시법
- 모든 자모를 찾아내기 어렵다.
- 새로운 자모 발견시 추가가 어렵다.
- 복합자모의 분리가 불가하다.
- 자소집단 ≥ 240자 => ISO 646IRV 적용불가

○ [제 1,2 안]의 문제
- 글자 구성 자소 수의 가변성 : 한글의 개별성
- 공간의 낭비 : 평균 7.9% 낭비

○ 선정 : 요구도를 만족하면서,
현실적으로 ISO 2022 사용,
단음화 자소 수용
=> [제2안] < 94

자소형 배열 : ISO 646 IRV 및 ISO 2022

	8	9	a	b	c	d	e	f
0	Fc	ㅈ	Fj					ㅋ
1	ㄱ	ㅉ	.				ㄱ	ㅌ
2	ㄲ	ㅋ	ㅏ				ㄲ	ㅍ
3	ㄴ	ㅌ	ㅐ				ㄴ	ㅎ
4	ㄷ	ㅍ	ㅑ				ㄷ	
5	ㄸ	ㅎ	ㅒ				ㄹ	
6	ㄹ		ㅓ				ㅁ	
7	ㅁ		ㅔ				ㅂ	
8	ㅂ		ㅕ				ㅅ	
9	ㅃ		ㅖ				ㅆ	
a	ㅅ		ㅗ				ㅿ	
b	ㅆ		ㅘ				ㅇ	
c	ㅿ		ㅙ				ㅈ	
d	ㅇ		ㅚ				ㅊ	
e	ㅈ		ㅛ					
f	ㅇ		ㅜ					
			ㅝ					
			ㅞ					
			ㅟ					
			ㅠ					
			ㅡ					
			ㅣ					

IBM PC 용

겹자소형 배열 : ISO 10646-1.3(예정)

00 01 02 03 ..... fd fe ff

00 01	00 A-영역 알파벳, 기호
... ...	11 한글 자소(초성 91, 중성 67, 종성 82)
...	34 한글 KS C 5601-1987(2350자) 3D 한글 보조-A KS C 5657-1989 (1930자) 44 한글 보조-B (2376자)
	4E I-영역 CJK 통합한자
	A0 O-영역 미사용
	E0 R-영역 사용자 정의
... fe ff	기타문자

국제공용문자판(BMP)

현대 및 옛 한글의 내부표현

○ 일차원 표현 : <c1, v, c2>
○ 불완전 문자표현 : <Fc ㅏㄹ ><ㅇ Fm ㄹ>
○ 순경음 표현 : <ㅁ ㅇ ><ㅂ ㅇ ><ㅃ ㅇ ><ㅍ ㅇ >
<ㅇ ㅇ ><ㅎ ㅎ >등

기존 코드와 호환성

○ 국내표준 : 자소형 ⊃ 자모형 ⊇ 조합형 ⊃ 완성형

○ 국제표준 : ISO 646 IRV => ISO 10646
(00 .. FF) => 11 + (00 .. FF)

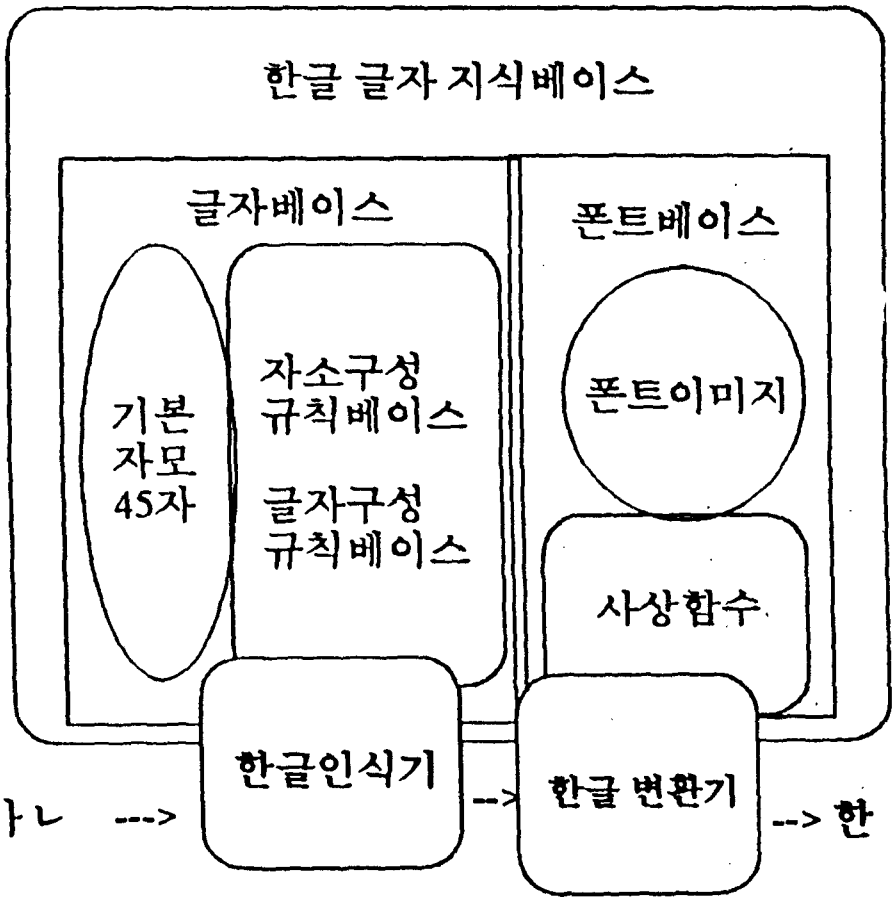
○ ISO 10646 - 1.3의 문제
- 한글 코드 체계가 5가지로 분할
. 겹자소형 1100-11FF
. 겹자모형 3130-318F
. 완성형 3400-3D2F
. 보조-A 3D2E-44B7
. 보조-B 44B8-4DFF
- Hangul Jamos Characters => Hangul Jasos Characters
- 문자열 정열 알고리즘 불가능
- 구현에 따라 한글 코드가 다를 가능성
- 국내표준과 호환성 결여
(해결) => 겹자소형 -> 낱자소형, 그외엔 모두 삭제

알고리즘 소개 및 평가

○ 자모형 옛 한글 자판

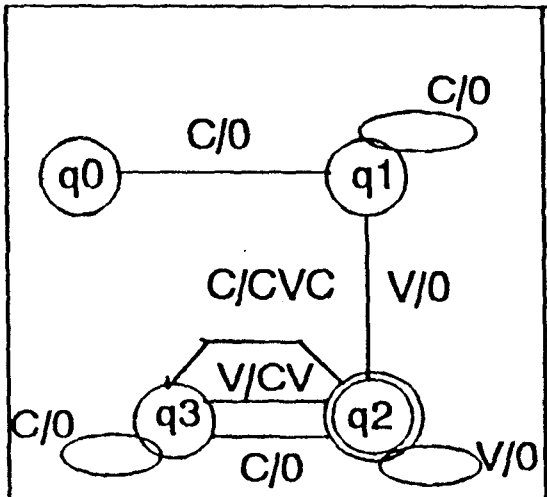
A	S	D ㅇ	F ㅿ	G ㆆ	H	J	K ·	L ‖
ㅁ	ㄴ	ㅇ	ㄹ	ㅎ	ㅗ	ㅓ	ㅏ	ㅣ

○ 한글 글자 KB

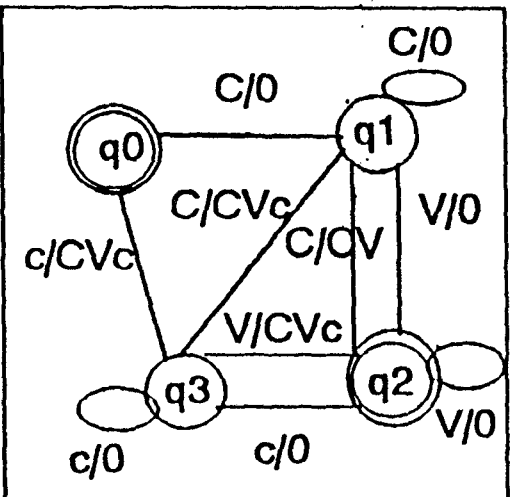


○ 한글 인식기(Accepter)

M = (Q, S, R, f, g, qi)
Q : 유한 상태 : (q0, q1, q2, q3)
S : 유한수의 문자 :(기본자소 45자)
R : 유한수의 출력 : $C^kV^mc^n$ (k,m,n = 1,2,3)
f : 상태천이 함수 f : Q X S ---> Q
g : 출력함수 g : Q X S ---> R
qi : 초기상태 qi ∈ Q



자모형 인식기



자소형 인식기

○ 한글 변환기(Transducer)

- 한글 글자 KB 참조

○ 문자열 정열 (한글 문자열 비교 알고리즘)

S T	예
1 초성 초성	ㄱㅏㄷㅏ ㅂㅗㄷㅏ
2 초성 중성	ㅂㅅㅏㄷㅏ ㅂㅏㄷㅏ
3 초성 종성	ㅁㅏㄹㄱㅗㄱㅣ ㅁㅏㄹㄱㅇㅡㅁ
4 중성 중성	ㅂㅗㄷㅏ ㅂㅏㄷㅏ
5 중성 종성	ㅂㅗㅏㄹㅏ ㅂㅗㄹㄱㅣ
6 종성 종성	ㅅㅗㅁ ㅅㅗㄹ

```
hstrcmp(char *s, char *t) {
  for ( ; *s == *t; s++, t++)
       if (*s == '\0')
            return (0);
  if (초성(*s) && 중성(*t))
     if (초성(*(s-1)) && 초성(*(t-1)))  return 10;
  if (초성(*t) && 중성(*s))
     if (초성(*(s-1)) && 초성(*(t-1)))  return -10;
  if (중성(*s) && 종성(*t)) return 10;
  if (중성(*t) && 종성(*s)) return -10;
  return (*s -*t);
}
```

○ 문자열 패턴 매치
- 영문자 알고리즘과 같음
- 단, 초성의 실패는 소우스의 다음 초성으로 건너 뜀
  에) S : ㅅ ㄷ ㅐ ㄹ ㅗㄴ ㅡㄴ
  P : ㄷ ㅐ

○ 커서 이동
- 글자 이동
- 자소 이동
* 삽입, 삭제, 대치의 글자, 자소 연산

○ 글자 수 계산
- 모음의 수 계산 (복모음은 하나의 단위)
- 모든 글자는 모음 포함 예)"<ㄱ Fm>은"

○ 문자 및 음성 인식의 후처리
- 인식율 90% 점후, 나머지 10% 후처리
- 불확정 획 또는 자소의 확인
- 자소의 완전 분리 => 낱자소형

요약 및 결론

○ 한글 코드 논쟁의 원인
- 한글문자 개별성의 인식론적 해석
- 한글문자의 문자론의 오해

○ 문제의 해결

- 훈민정음의 존재론적 접근
. 문자론 : 음소문자, 음절문자 특성
. 집합론 : 불확정 무한 글자 집합
. 지식공학적 구현

○ 결론
- 문자론적, 집합론적 사실의 적용
- 지식공학적 구현

=> 자소형 코드
. 한국어 정보처리의 모든 응용분야 만족
. 기존 코드와 호환성 제고
. 알고리즘의 효율성 및 간결성

=> 문자집단의 간결성
. ISO 646 IRV, ISO 2022, ISO 10646에 적합성