

한국자소의 분류와 연속 상관빈도

Classification of Korean Characters and Frequency of Continual Characters

김국*, 정병용**

ABSTRACT

Classification of Korean characters(alphabets) and frequency data of them are studied that is essential to information process of Korean. We defined a classification of characters using the concept of 'set of 2 parts' and 'set of 3 parts', and we researched frequencies about all combinations of continual two characters. These data would be important basic data to design input device of computer, for example.

Keyword: Korean characters, classification of Korean characters, frequency of continual two characters

* 서경대학교 산업공학과, 서울시 성북구 정릉동, E-mail: kkim@skuniv.ac.kr

** 한성대학교 산업공학과, 서울시 성북구 삼선동, E-mail: byjeong@hansung.ac.kr

1. 서론

한글은 기본적으로 소리글자이면서 모아쓰기를 하는 독특한 성격을 가진다. 한자의 영향으로 초성, 중성, 종성이 모인 한 음절 단위로 한 글자를 이룬다. 라틴 문자와 비교하면 한글은 2~5개의 알파벳이 모여서 한 문치의 글자를 이루는 특성이 있으며 이것이 한글의 정보화를 어렵게 하는 한 요인이 된다. 한글의 정보화에 관련된 주제들은 자판, 한글코드, 폰트의 설계와 같은 것들이다. 그러나 이는 소리글자이면서 뜻글자의 장점을 가지게 하고 있으며 보다 복잡한 중국, 일본에 비해서는 유리한 입장이다.

한글의 글자체 체계의 특징에는 분명한 사실이 있는데 즉 한 글자문치(보통 음절)는 세 부분, 즉, 초성(자음), 중성(모음), 종성(자음, 받침)으로 나눌 수 있다는 것이다. 이것이 로마자와 다른 점이다. 그리고 모음과 받침은 기본 자모의 결합으로 이루어지기도 하고 받침은 꼭 있는 것은 아니다. 이러한 점이 한글의 정보화에 관련되어 중요한 요소이다. 이후 종성은 받침이라고 부르기로 한다.

여기서 우리는 글자 문치에 있어서 한글의 자소(characters) 종류 및 한글 오토마타에 따른 자소집합을 정의, 제시하였으며 자소 간의 상관 빈도를 조사 연구하였다. 자소 분류는 문법적 관점이 아닌 입력장치 설계관점의 구분 방법을 제시한다. 한글의 단순 빈도조사는 여러 가지 연구가 있는데 '한국기계연구소(1985)', '한글기계화연구소(1975)', '송계범(1968)'의 조사가 있다. 어휘조사는 '김영채

(1986)'의 연구가 있다. 여러 공학설계에서의 한글 특성에 대해서는 '오길록 등(1995)'의 자료가 기초를 잘 제공한다. 우리는 두 개의 자소 간 상관관계의 빈도를 조사 연구하였다. 앞 자소의 입력 후 뒷 자소의 입력이 되는 관계를 말한다. 이것은 컴퓨터 시대에 한글의 정보화에 관련된 자판, 한글코드, 폰트의 설계와 같은 많은 연구 주제들의 기본 자료로서 중요한 것이다.

2. 한글 자소의 종류와 자소 집합

한글 자소는 현재 쓰이는 것으로만 대상으로 한다. 한글 자소의 기본은 자모(alphabet) 24 자로서 {ㄱ, ㄴ, ㄷ, ㄹ, ㅁ, ㅂ, ㅅ, ㅇ, ㅈ, ㅊ, ㅋ, ㆁ, ㅎ, ㅌ, ㅍ, ㅑ, ㅓ, ㅕ, ㅗ, ㅛ, ㅜ, ㅠ, ㅡ, ㅣ}과 같은데 이를 기본으로 모아쓰기에 의해 한글 음절을 구성하는 자소(character elements)는 표 1.과 같이 정리된다. 이 자소 분류체계는 문법적 또는 음

표 1. 한글의 자소 종류

구분	자소들	갯수	비고
자음	단자음 SA {ㄱ, ㄴ, ㄷ, ㄹ, ㅁ, ㅂ, ㅅ, ㅇ, ㅈ, ㅊ, ㅋ, ㆁ, ㅎ}	14	초성과 받침 양쪽에 쓰임
	쌍자음 SB {ㄲ, ㄸ, ㅃ, ㅆ, ㅉ}	5	{ㄲ, ㅆ}은 초성과 받침, {ㄸ, ㅃ, ㅉ}은 초성에만 쓰임
	겹자음 SC {ㄴㅇ, ㄷㅇ, ㄹㅇ, ㄹㅁ, ㄹㅂ, ㄹㅅ, ㄹㅇ, ㄹㅁ, ㄹㅂ, ㄹㅅ}	10	받침에만 쓰임
모음	단모음 VA {ㅏ, ㅑ, ㅓ, ㅕ, ㅗ, ㅛ, ㅜ, ㅠ, ㅡ, ㅣ}	10	
	겹모음 VB {ㅘ, ㅙ, ㅚ, ㅛ, ㅜ, ㅠ, ㅡ, ㅣ, ㅞ, ㅟ, ㅠ, ㅡ, ㅢ, ㅣ}	11	{ㅘ, ㅙ, ㅚ, ㅛ}는 단모음 글자꼴
계		50	

운학적 분류체계를 말하는 것이 아니고, 한글자음치의 구성 체계를 말한다. 예컨대 단순모음, 중모음의 개념이 아니고, 글자가 기본자소인지 합성된 것인지를 말한 것이다.

전에는 ㄹ 받침이 있었는데 쓰이는 곳은 'ㄹ'이라는 한 글자밖에 없었고, 1989년 표준말이 개정되면서 없어졌다. 이밖에 받침 ㄱ 같은 경우 ㄴ, ㄷ, ㄹ 세 개밖에 없다. 한글은 라틴 문자에 비해 모음이 많은데, 예를 들어 영문에서는 모음이 {a, e, i, o, u} 다섯 뿐인데 한글은 10개이다. 예컨대 ㅏ 는 음운으로는 | + ㅏ 와 같은 중모음이지만 글자체로는 단모음꼴이다. 영어와 비교하면 {ㅏ, ㅑ, ㅓ, ㅕ, ㅗ, ㅛ, ㅜ, ㅠ} 를 기본으로 {ㅓ, ㅕ, ㅗ, ㅛ} 가 단모음 꼴이고, {ㅓ, ㅕ, ㅗ, ㅛ} 를 포함하면 14 개가 단모음 꼴이다.

컴퓨터에서 한글의 모아쓰기를 자동으로 처리해 주는 것을 한글 오토마타라고 한다. 자소의 입력 단위에 따라 자소 집합의 구분을 두벌집합, 세벌집합으로 구분하자. 두벌집합은 앞의 표와 같이 자소를 자음, 모음의 두가지로 구분하여 한글을 구성하는 것이고, 세벌집합은 초성, 중성, 종성으로 구분하는 것이다. 두벌집합은 입력 단위가 적은 장점이 있으나, 입력과 출력 표시간의 정합성(compatibility)이 떨어지고 한글의 구성 원리와 부합하지 않는 단점이 있다. 인간공학적 정합성의 개념은 '박경수(1990)'을 참고한다. 사실 두벌집합은 컴퓨터화된 입력장치가 아니면 불가능하다.

한글 자소를 3벌집합으로 분류하면 초성자소 19개, 중성자소 21개, 종성자소 26개 도합 66개이다. 그 자세한 내용은 뒤에 나오는

표 2. 및 표 3.에서 보는 바와 같다.

모아쓰기를 위한 자소의 수에 따라 완전집합, 보통집합, 최소집합으로 구분하자. 선정된 자소 외의 자소는 한글 오토마타로 합성한다.

2.1 완전집합

선정된 자소집합으로서 초성이든, 중성이든, 받침이든 각 1 회 입력으로서 현대의 모든 한글을 표현할 수 있을 때 그러한 자소집합을 완전집합이라 하자. 두벌집합의 완전집합은 표 1.에서와 같이 자음 자소 29 개, 모음 자소 21 개, 도합 50 개이고, 세벌집합의 완전집합은 초성 자소 19 개, 중성 자소 21 개, 종성 자소 26 개, 도합 66개이다. '정승훈 등(1991)'의 연구에서는 '최대완전집합'이라고 불렀다.

2.2 보통집합

보통집합은 정의를 여러 가지로 할 수 있으나, 여기서 {ㅓ, ㅕ, ㅗ, ㅛ}을 제외한 나머지 모든 겹자모는 합성하도록 선정된 자소집합을 보통집합이라고 부르자. ㅓ는 ㄹ + ㅓ, ㅕ는 ㄹ + ㅕ + | 와 같다. {ㅓ, ㅕ, ㅗ, ㅛ}는 단모음 꼴로 간주되어 선정된다. 쌍자음의 경우, 두벌집합에서는 반드시 포함되어야 한다. 왜냐하면 쌍자음은 일반적으로 합성이 불가능하다. 겹자음은 받침에만 쓰이기 때문에 단어의 끝이 되거나, 다음에 초성의 자음이 반드시 등장하지만, 쌍자음은 해당 단자음이 초성과 받침에 다 쓰이기 때문에 거듭치기로 해결이 안된다.

예컨대 '느끼다'의 경우 '느+ㄱ+ㄱ+ㅣ+다'로 치면 '느끼다'처럼, '왼쪽'의 경우 '왼+ㅅ+ㅅ+ㅡ+ㄱ'으로 치면 '왼쪽'처럼 된다. 따라서 두벌집합에서는 보통집합이 자음 19 개, 모음 14 개, 합 33 개이다. 컴퓨터 키보드의 예처럼 이 집합이 보편성이 있다. 세벌집합에서는 쌍자음을 해당 단자음의 거듭 입력으로 합성 가능하다. 세벌집합에서는 초성 14 개, 중성 14 개, 받침 14 개로 도합 42 개가 보통집합이 된다.

완전집합과 보통집합 사이에서 일부의 겹차모를 하나의 입력자소 단위로 사용할 경우도 있다. 예컨대 자판 설계에서 공병우 자판이 그 예이다.

표 2. 자소 집합의 분류

* ()안은 자소의 수.

	개요	2벌집합			3벌집합			
		자음	모음	합	초성	중성	받침	합
완전 집합	완전한 자소종류 전부	SA(14) +SB(5) +Sc(10) =29	VA(10) +VB(11) =21	50	SA(14) +SB(5) =19	VA(10) +VB(11) =21	SA(14) +{ㅍ, ㅋ, ㆁ} +SC(10) =26	66
보통 집합	모든 겹차모 합성. ㅈ, ㅊ, ㅋ, ㆁ 선정. 쌍자음은 두벌집합에서는 선정	SA(14) +SB(5) =19	VA(10) +{ㅈ, ㅊ, ㅋ, ㆁ}=14	33	SA(14)	VA(10) +{ㅈ, ㅊ, ㅋ, ㆁ}=14	SA(14) =14	42
최소 집합	오토마타가 성립되는 최소자모. ㅈ, ㅊ, ㅋ, ㆁ도 합성	상동	VA(10) =10	29	상동	VA(10)	SA(14) =14	38

2.3 최소집합

한글 오토마타가 성립하는 필수적인 최소의 자소 집합을 최소집합이라 하자. 단모음 풀인 {ㅈ, ㅊ, ㅋ, ㆁ}는 각각 {ㅈ+ㅣ}, {ㅊ+ㅣ}, {ㅋ+ㅣ}, {ㆁ+ㅣ}로 합성 할 수 있다. 전자수첩과 같이

키보드의 수가 적은 제품에서 그러한 처리방식이 있다. 두벌집합에서는 최소집합이 자음 19 개, 모음 10 개 합 29 개이다. 세벌집합에서는 초성 14 개, 중성 10 개, 받침 14 개로 도합 38 개가 최소집합이 된다.

이상을 정리하면 표 2.와 같다.

3. 자소 상관 빈도

다음으로 자소의 빈도에 대해 조사를 하였다. 단순히 자소별 빈도뿐 아니라 자소간의 관계에 따른 빈도를 조사하였다.

한글 자소 빈도는 연구자와 조사 텍스트에 따라 여러 가지가 있다. 빈도 조사의 기초는 초성, 중성, 받침 단위가 된다. 한 예는 '송계범(1968)'의 자료와 같다. 이를 빈도순으로 그래프로 보인 것이 그림 1.과 같다. 현 맞춤법을 기준으로 하여 받침 ㄹ은 제외하였다. 한글에서 모음보다 자음의 빈도나 연속되는 경우가 많은데 이점은 영문에서도 film, strike처럼 마찬가지이다.

자소 빈도 자료의 기초는 초성, 중성, 받침의 모든 자소 즉, 3벌집합의 완전 집합의 경우가 된다.

우리는 소설, 신문, 전문서적, 백과사전의 4 가지에 대해 자료를 샘플링하여 조사하였다. 연구조원을 통하여 이들 자료를 임의 선정 수집하고 수식, 영문자, 특수문자를 제외하고 한글 음절만을 텍스트 모드로 입력하여 원재료로 사용하였다. 백과사전은 두산세계대백과사전을 사용하여 임의 발췌하여 표제어를 제거하고 텍스트로 삼았다. 표제어는 특정 자

+빈칸, 빈칸+ㅎ, ㅎ+ㅏ, ㅏ+ㄴ, ...' 과 같은 자소 관계가 나타나고 이 모든 조합의 빈도를 조사하였다. 66 개의 초중종 자소 외에 빈칸, 마침표, 쉼표, 숫자, 기타(영문, 기타부호)의 다섯 가지의 부호를 포함하였다. 부호 중 한글 자소의 상관관계와 관련된 의미가 있는 것은 마침표, 쉼표, 빈칸이라고 생각하였다.

3별집합의 보통집합의 경우는 초성의 쌍자음은 분배하고, ㅈ, ㅊ, ㅋ, ㆁ을 제외한 겹모음 분리 배분하고, 그리고 받침의 겹자음도 분리 배분한다. 최소집합의 경우는 보통집합에서 다시 ㅈ, ㅊ, ㅋ, ㆁ을 분리 배분한다. 보통집합, 최소집합은 분석 프로그램을 개정하여 추가로 연구 조사할 계획이다. 두별집합을 기준으로 할 경우도 역시 분석 프로그램의 변형에 의해 분석 가능하다.

상관 빈도가 아닌 각 자소별 빈도는 표의 행(가로방향) 별로 합계를 낸 것을 사용할 수 있다. 그 결과는 표 3.과 같이 정리된다.

조사 텍스트의 총 자소수는 초성, 중성, 종성, 부호(쉼표, 빈칸, 마침표)를 포함하여 약 275,400 개다. 빈칸이 2 개 이상 계속되는 것은 1 개로 간주하였다. 초중종성의 비율은 그림 2.와 같다.

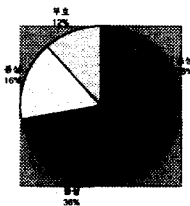


그림 2. 초중종성 및 부호의 비중

표 3. 자소별 빈도

초성	중성	중성	부호				
[ㄱ]	12846	[ㅏ]	21709	[ㅇ]	4336	[.]	31777
[ㄴ]	738	[ㅓ]	4363	[ㅗ]	57	[,]	103
[ㄷ]	6831	[ㅕ]	640	[ㅜ]	17	[;]	74
[ㄹ]	9144	[ㅗ]	10	[ㅡ]	14343		
[ㅁ]	715	[ㅛ]	9958	[ㅝ]	29		
[ㅂ]	7438	[ㅜ]	4615	[ㅟ]	352		
[ㅅ]	4736	[ㅠ]	4788	[ㅠ]	261		
[ㅇ]	3820	[ㅡ]	475	[ㅢ]	8938		
[ㅋ]	135	[ㅣ]	10325	[ㅣ]	100		
[ㆁ]	8597	[ㅑ]	1773	[ㅤ]	69		
[ㅃ]	273	[ㅓ]	72	[ㅥ]	38		
[ㅅ]	23692	[ㅕ]	1222	[ㅧ]	1		
[ㅈ]	7855	[ㅗ]	867	[ㅩ]	3		
[ㅊ]	192	[ㅛ]	6751	[ㅫ]	27		
[ㅋ]	2172	[ㅜ]	513	[ㅭ]	2563		
[ㆁ]	688	[ㅠ]	15	[ㅱ]	1453		
[ㅃ]	1199	[ㅡ]	525	[ㅳ]	321		
[ㅅ]	1138	[ㅣ]	579	[ㅵ]	1346		
[ㅎ]	7024	[ㅑ]	13299	[ㅷ]	2567		
		[ㅓ]	2114	[ㅹ]	6941		
		[ㅕ]	14781	[ㅻ]	150		
				[ㅽ]	143		
				[ㅿ]	18		
				[ㅿ]	279		
				[ㅿ]	203		
				[ㅎ]	307		
소계	99233	소계	99394	소계	44862	소계	31954
합계	275443						

자소 66 개를 빈도가 높은 것부터 정렬하면 20 개(자소 수의 30%)가 84 % 를 차지하며 25 개가 90 %를 차지한다. 이것을 파레토 그림으로 그려보면 그림 3.과 같다. 자음 중에 *표가 붙는 경우는 받침을 초성과 구분하여 식별하기 위한 것이다.

한편 그림 4.는 초성, 중성, 종성 별 각각 빈도순 그래프이다. '송계범(1968)'의 자료와 순위가 다른 곳이 있다. 초성의 경우 빈도 상위순위에서 ㅅ, ㄴ이 각 8.7%, 6.9%로서 송계범의 자료와 순위차이가 발생하였다. ㄹ, ㅎ의 순위도 달라졌다. 모음에서 ㅏ, ㅓ의 순위와 ㅓ, ㅕ의 순위가 송계범과 다르다. 모음



그림 3. 자소 전체 빈도순 및 파레토 그림

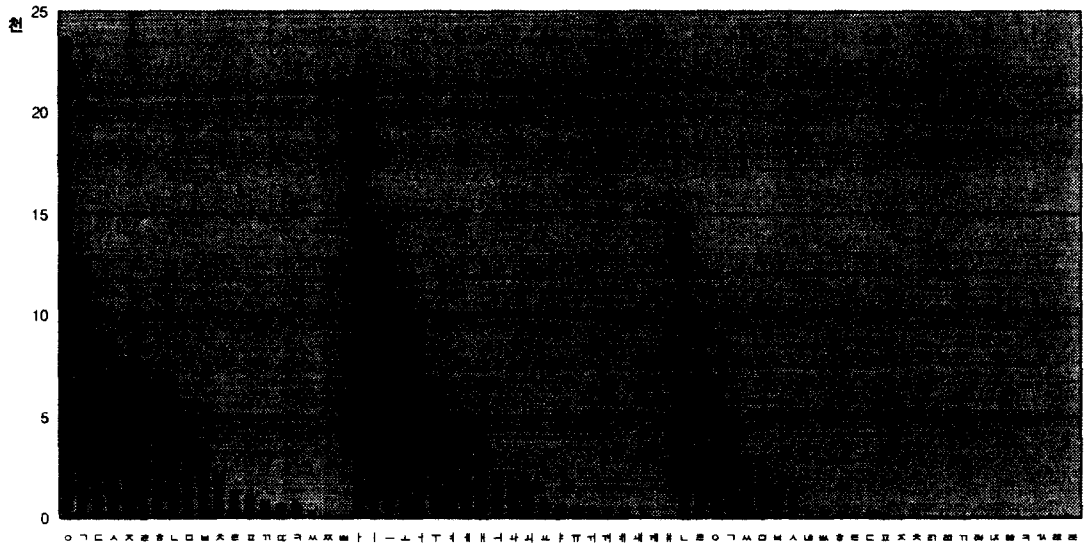


그림 4. 초성중성 받침별 빈도순

과 받침에서는 이들은 편차가 크게 달라지진 않았다. 빈도가 극히 적은 곳에서의 차이점은 통계적 오차의 범위이다. 이러한 차이의 발생은 샘플링한 자료의 차이에 인한 것으로서, 시대의 변화와 텍스트에 따라 달라지는 것으로 판단되나 이러한 차이의 분석은 이 논문에서 자세히 다루어지지 않았다.

그러나 예컨대 초성에서는 ㄷ 이 ㅋ보다 빈도가 높고, 받침에서는 ㅍ 이 다른 자소에 비

해 월등히 높은 위치에 있다. 이런 것은 현재의 키보드 설계가 어떤 면에서 최적이지 아니라는 것을 의미한다.

이 연구의 중요한 부분인 초, 중, 종성간의 상관 빈도는 표 4, 5, 6, 7에서 보는 바와 같다. 그 조합은 초성-중성, 중성-받침, 중성-초성, 받침-초성의 4 가지가 의미가 있다. 'ㄱ'은 자음으로서 ...'와 같이 초성 다음 초성의 경우는 여기서 고려하지 않았다.

표 4. 초성-중성의 상관빈도

	[ㄱ]	[ㅋ]	[ㆁ]	[ㄴ]	[ㄷ]	[ㄸ]	[ㄹ]	[ㄺ]	[ㄻ]	[ㄼ]	[ㄽ]	[ㄾ]	[ㄿ]	[ㅀ]	[ㅁ]	[ㅂ]	[ㅃ]	[ㅄ]	[ㅅ]	[ㅆ]	[ㅇ]	계
[ㄱ]	2576	244	0	0	1211	675	661	301	2243	766	4	29	156	859	66	2	37	71	1604	0	1241	12846
[ㅋ]	199	21	0	0	40	32	8	0	68	4	3	5	0	133	2	0	1	0	90	0	132	738
[ㆁ]	1668	431	54	0	148	100	410	0	325	3	0	8	6	139	4	0	3	14	2878	5	635	6831
[ㄴ]	3649	763	1	0	335	221	2	0	1631	0	38	773	0	246	1	0	69	0	1316	0	99	9144
[ㄷ]	162	220	0	0	122	9	0	0	112	0	0	0	0	12	0	0	22	0	46	2	8	715
[ㄸ]	1003	230	70	0	731	102	565	24	1617	0	0	38	41	119	2	0	4	177	1427	0	1288	7438
[ㄹ]	1315	106	0	0	161	46	923	0	613	0	0	0	14	1038	12	0	0	3	50	0	455	4736
[ㄺ]	925	189	0	0	342	25	230	0	785	6	0	2	0	940	0	0	0	1	22	0	353	3820
[ㄻ]	37	14	0	0	10	0	2	0	3	0	0	0	1	55	0	0	0	0	7	0	6	135
[ㄼ]	1834	559	4	0	1963	338	40	3	712	0	5	6	2	898	1	0	28	2	674	0	1528	8597
[ㄽ]	51	0	0	0	58	0	0	0	6	0	0	0	0	4	0	0	0	0	74	0	79	273
[ㄾ]	1893	59	430	8	2335	2144	1215	126	569	458	10	86	503	732	379	4	288	293	4231	2076	5853	23092
[ㄿ]	1490	187	1	0	1768	498	129	0	609	17	0	10	9	881	9	0	10	0	148	0	2049	7855
[ㅀ]	29	37	0	0	16	0	0	0	51	0	0	1	0	8	0	0	0	0	4	0	46	192
[ㅁ]	416	99	0	0	329	198	79	1	149	4	0	56	0	355	8	0	30	0	69	0	379	2172
[ㅂ]	123	11	0	0	37	22	37	0	148	0	6	0	0	8	0	0	0	0	178	0	118	688
[ㅃ]	233	115	0	0	160	54	0	0	288	1	0	22	0	48	0	0	3	0	227	0	38	1199
[ㅄ]	214	49	0	0	30	25	136	10	143	0	0	0	93	129	0	0	0	4	150	0	155	1138
[ㅅ]	3805	1028	77	0	144	21	334	9	244	510	3	182	42	153	27	1	22	13	96	29	284	7024

표 5. 중성-받침의 상관 빈도

	[ㄱ]	[ㅋ]	[ㆁ]	[ㄴ]	[ㄷ]	[ㄸ]	[ㄹ]	[ㄺ]	[ㄻ]	[ㄼ]	[ㄽ]	[ㄾ]	[ㄿ]	[ㅀ]	[ㅁ]	[ㅂ]	[ㅃ]	[ㅄ]	[ㅅ]	[ㅆ]	[ㅇ]	계						
[ㄱ]	777	36	0	3324	27	334	74	1440	44	10	14	0	0	1	817	226	8	46	260	1589	112	2	0	186	90	29	9406	
[ㅋ]	258	0	0	57	0	0	14	0	0	0	0	0	0	0	48	3	0	31	315	628	5	0	0	0	0	0	1359	
[ㆁ]	77	0	0	2	0	1	0	1	0	1	2	0	0	2	1	3	0	0	0	230	0	0	0	4	0	2	355	
[ㄴ]	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	2	
[ㄷ]	695	8	4	1406	0	4	38	528	6	29	15	0	0	0	372	399	291	731	578	989	5	1	2	2	9	128	6139	
[ㄸ]	22	0	3	54	2	2	28	63	15	15	0	0	3	3	21	0	0	27	66	8	2	0	0	1	0	1	336	
[ㄹ]	379	4	0	1292	0	1	2	325	0	0	5	0	0	0	66	95	1	4	313	838	1	32	9	10	14	1	3392	
[ㄺ]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	18	0	0	0	0	1	0	0	0	19	
[ㄻ]	739	2	7	471	0	0	36	356	0	10	1	0	0	2	109	64	0	280	0	1305	3	39	1	2	53	96	3575	
[ㄼ]	76	1	0	330	0	0	0	46	0	0	0	0	0	0	0	1	0	69	154	0	0	1	0	0	1	0	679	
[ㄽ]	0	0	0	7	0	0	1	0	0	0	0	0	1	0	0	0	0	3	18	0	0	0	0	0	0	0	30	
[ㄾ]	31	0	2	224	0	0	0	58	0	0	0	0	0	0	16	7	0	2	0	5	0	0	0	0	0	1	346	
[ㄿ]	18	0	0	2	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	281	
[ㅀ]	634	4	0	1096	0	0	22	864	4	3	0	1	0	4	209	16	0	45	0	456	0	0	0	24	9	1	3392	
[ㅁ]	0	1	1	291	0	0	0	108	0	0	0	0	0	0	0	0	0	1	17	1	0	0	0	0	0	0	420	
[ㅂ]	1	2	0	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	9	
[ㅃ]	3	0	0	5	0	0	7	1	0	0	0	0	0	0	0	20	0	21	0	1	0	0	0	0	0	0	5	
[ㅄ]	74	0	0	24	0	0	0	60	0	0	0	0	0	0	2	0	0	0	0	29	0	0	0	0	0	0	189	
[ㅅ]	242	0	0	4273	0	11	33	4029	20	0	0	0	0	0	484	304	0	95	0	480	8	0	0	40	1	0	10000	
[ㅆ]	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	6	
[ㅇ]	319	1	0	1497	0	0	27	1945	10	1	0	0	0	0	17	429	32	0	41	931	65	16	69	1	12	67	0	4869

다음으로 부호와 한글자소와의 관계에서 자소 다음 부호로 끝나는 것을 조사하였다. 부호는 우선 빈칸, 마침표, 쉼표에 대해 조사하였으며 추후 물음표, 따옴표에 대해서도 조사할 계획이다. 숫자, 영문의 경우는 전후로 빈칸이 있는 것으로 가정하여 자소의 빈도조사와 직접 관련이 없다고 보았다. 빈칸으로 끝나는 자소, 마침표로 끝나는 자소, 쉼표로 끝나는 자소는 표 8.과 같다. 이 세가지는 단어나 문장의 끝을 의미하며 중성 또는 중성 다음에 오게 된다. 예를 들어 '-다.'는 'ㅏ+'의 순서이다. 마침표로 끝나는 것은 '-다'로 끝나는 것이 대부분이므로 'ㅏ'가 일순위임을 쉽게 예상할 수 있다.

4. 결 론

한글은 기본 자모 24 자의 모아쓰기에 의해 글자체가 구성된다. 한 음절을 구성하기 위한 자소 집합에 있어서 초성, 중성, 받침으로 구분하는 3별집합과 자음, 모음으로 구분하는 2별집합으로 분류할 수 있다. 우리는 한글 오토마타를 위해 선정되는 자소에 따라 완전집합, 보통집합, 최소집합을 정의하여 제시하였다. 최소집합은 다른 자소의 합성으로 가능한 것을 제외한 최소한의 자소집합을 말한다. 이 제안은 한글정보화에 유익하게 사용될 것이다.

다음으로 여러 종류의 문헌, 즉 소설, 신문, 전문서적, 백과사전의 다양한 샘플링을 통한 자료 텍스트의 자소 빈도를 조사하였다.

이것은 초성, 중소, 받침의 모든 자소를 기

표 8. 빈칸, 마침표, 쉼표로 끝나는 자소

	[]	[]	[]	계	[]	[]	[]	계	
[]	2003	2272	102	4377	[]	666	9	47	722
[]	535	7	53	595	[]	2	0	0	2
[]	149	11	6	166	[]	0	0	0	0
[]	0	0	0	0	[]	7091	34	130	7255
[]	1237	29	33	1299	[]	0	0	0	0
[]	2105	46	46	2197	[]	5	0	0	5
[]	564	6	87	657	[]	22	2	0	24
[]	46	2	3	51	[]	3872	5	37	3914
[]	2978	59	205	3242	[]	6	0	0	6
[]	591	3	15	609	[]	5	0	0	5
[]	12	1	0	13	[]	2	0	0	2
[]	22	1	46	69	[]	0	0	0	0
[]	32	50	9	91	[]	0	0	0	0
[]	594	12	16	622	[]	0	0	0	0
[]	29	6	0	35	[]	292	18	19	329
[]	0	0	0	0	[]	108	15	12	135
[]	45	0	1	46	[]	0	0	0	0
[]	22	2	2	26	[]	140	4	5	149
[]	418	3	16	437	[]	6	0	0	6
[]	1798	0	2	1800	[]	647	20	53	720
[]	3009	40	137	3186	[]	4	0	0	4
					[]	79	0	1	80
					[]	6	0	0	6
					[]	2	0	1	3
					[]	6	0	0	6
					[]	2	0	0	2
계	16189	2550	779	19518	계	12963	107	305	13375

초로 하였다. 또한 중요하게 자소간의 상관 빈도, 즉 첫 입력 자소 다음의 입력 자소 간의 조합에 따른 빈도를 조사한 것이다. 아울러 빈칸, 쉼표, 마침표로 끝나는 자소, 처음 시작하는 초성의 빈도 등도 분석 정리하였다.

우리는 이 연구과정에서 텍스트 원재료 작성에 사용한 문헌 샘플의 명확한 특성을 함께 준비하지 않았는데 이는 문제점이라고 생각한다. 연속된 자소의 분석이라는 관점에서 프로그램 설계와 분석에 중점을 두었다. 이 결과를 바탕으로 다음 기회에 더 많은 샘플을 사용하여 분석을 하고자 한다.

이 연구 결과는, 예를 들어 키보드와 같은

입력장치의 설계 등 한글의 정보화, 인간공학 적 설계에 관한 주제에 유용한 기초자료로 중요하게 쓰일 것이다. 자소의 상대 빈도와 연속 되는 상관은 키보드 구성에 유용하게 적용되는데 그 가이드라인과 배열 시안은 본 연구 범위를 초과하는 것으로서 저자들이 연구를 계속하고 있다. PDA나 핸드폰의 글자 할당에도 적용될 수 있을 것으로 보이는데 이 경우는 단순히 빈도만이 아니라 자모의 순서 및 자모의 유사성에 의한 정합성도 고려되어야 할 것이다.

참고 문헌

김영채, "한국어 어휘 빈도 조사", 한국심리학회지, 5, 1986.
 김우철 등, 일반통계학, 영지문화사, 1986.
 박경수, 인간공학, 영지문화사, 1990.
 송계범, 한글타자기의 건반 배열에 관하여, 1968.
 오길록, 최기선, 박세영, 한글공학, 대영사, 1995.
 이만영, 표준한글자판 문제 해결을 위한 정책결정 모형의 개발, 국어정보학회, 1992.
 정승훈, 박진우, 이일병, "컴퓨터 모의실험에 의한 자판배열의 성능평가", 제3회 한글 및 한국어 정보처리 학술대회 논문집, 1991.
 한국기계연구소 연구보고서, 단순빈도조사(통계 활용자료, 수필, 단편소설, 신문사설, 공문서), 한국기계연구소, 1985.
 한글화기계연구소 연구보고서, 한글 글자의 찾기 조사, 한글기계화연구소, 1975.

저자 소개

◆ 김 국

서울대학교 산업공학과 졸업
 한국과학기술원 산업공학박사
 국방과학연구소 시스템종합 및 기획담당 경력
 서경대학교 교수
 한국설비보전공학회 이사
 설비관리 용어사전 편찬
 관심분야: 물류관리 및 정보화, 감성공학, 신뢰성공학

◆ 장병용

고려대학교 산업공학과 졸업
 한국과학기술원 산업공학박사
 한성대학교 교수
 삼성전자 인간공학적 개선연구
 관심분야: 인간공학 및 응용, 산업안전관리

논문접수일 (Date Received): 2002/12/20
 논문게재승인일(Date Accepted): 2002/04/30