발음사전 표제어중의 음소의 통계적 성질 - 음성 DB용 단어선정을 위하여 -

°이 용주, 김 경태, 조 철우*, 이 태원*

한국전자통신연구소 *고려대학교

On the statistics of Korean Phonetic Dictionary

- Basic Survey to make corpus of Korean Speech DB -

Y. J. LEE, K. T. KIM, C. W. JO*, T. W. RHEE*

ETRI * Korea Univ.

Abstract

Statistical information about spoken Korean was obtained. The data are the results of analyzing the Korean phonetic dictionary.
This is one of the basic survey to make phoneme ballanced corpus of Korean Speech Data Base(KSDB).

1. 서 론

불특정 발성자의 대어휘, 연속어인식이나 규칙합성을 위해서는 각종환경하에서의 음운변동을 파악할 필요가 있으므로 다수 발성자의 다양한 음운환경을 가진 음성데이타를 음소와 같은 미세단위로 labelling한 음성데이타베이스(DB)가 필요하다. 따라서, 각국에서는 본격적인 음성연구의 기반으로서 이 음성 DB의 구축을 조직적으로 시도하고 있다. (1) (2) (3)

필자들도 한국어 음성 DB 의 필요성을 절감하고 이에 대한 기초적인 검토를 수행중에 있다. (1) (4) (5)

일반적으로 다양한 음운환경을 가진 음성 DB용 대상 단어를 선정하기 위해서는 사전의 표제어, 중요어, 고빈도 단어 등을 대상으로 2 음소열 또는 3음소열의 종류수, Entropy 등을 메개로한 여러가지 통계적인 선정방법을 사용하고 있다. (6) (7)

특히 사전의 표제어를 대상으로 할 경우에는 먼저 이 표제어를 발음표기하여 컴퓨터에 입력한후 음소 또는 음소열의 발생빈도등 각종 통계적인 자료를 도출해야만 한다. (13)

이러한 통계적 자료는 음성 DB 용 대상단어 선정을 위해서만이 아니고 자연언어처리, 음성인식 및 합성에 기초자료로서도 의미가 있으며 사전 데이타베이스(8) 자체를 우리말의 대량 데이타로 다루므로서 언어학적, 음성학적으로 우리말이 가진 성질에 관해서 유익한 자료를 얻을 수도 있다. 본고에서는 음성 DB용 단어 선정을 위한 기초조사로서 우선 8500단어 규모의 우리말 발음사전의 표제어에 대한 음소 또는 음소열 관련의 통계적인 조사결과에 대하여 논하였다.

2. 분석자료 및 방법

o 대상자료: 한국어 발음사전(9) (8575 단어)

o 입력방법: 발음사전의 표제어와 함께 한글로 발음표기된 단어를 2byte code로 컴퓨터에 입력.

o 음소 및 음소열의 선정:

음소는 일반적인 음소분류(10)를 따랐으나, 중모음은 반모음과 단모음으로 분리하였고 음절의 어두자음과 어말자음은 음소분류에 상관없이 서로 구분하여 다루었다. (표1참조) 2음소열은 위와같이 분류된 음소단위의 쌍으로 구성하였고 3 음소열은 단어의 첫머리와 끝에오는 경우도 같이 고려하였다.

(예: bCV는 단어첫머리에 오는 CV음절로서 공백을 음소처럼 취급하여 3음소열로 설정하였다)

o 제한사항:

본 조사에서는 장모음은 고려되지 않았고, 어두의 /의/가 /으이/로 발음표기 되어있어서 이를 그대로 이용하였으므로 별도의 음소로 다룰수 없었음을 먼저 밝혀둔다.

표1. 음소 구분

일반적인 분류	본 조사에서의 분류
모 음 10	모 음 9
중모음 12	반 모 음 2
반모음 2	어두자음 19
자 음 19	어말자음 7
계 43 종류	계 37종류

3. 결과 및 검토

조사대상으로한 발음사전의 규모는 8575단어에 19785 음절, 음소단위로 보면 61655개로 구성되어 있다.

그림1과 그림2는 음소 및 음절길이별 단어분포를 보인것이다. 2~3음절어 및 5~6음소어가 가장 많다. 단어당 평균 음절수는 2.3개, 평균 음소수는 6개이다.

먼저 각 음소열의 종류를 규칙에 의해 생성한 경우와 본 자료에 의한 조사결과를 표2에 보였다.

이 분류기준에 의하면 2음소열은 이론적으로 770종류를 구성할 수 있는데 여기서는 이중 약 80% 정도가 출현하고 있다. 나머지 20%는 음소결합의 제약성을 나타낸다고 볼 수 있다. 3음소열의 경우 전체 실현율은 1/5 정도로 대부분은 위와같은 제약성 때문이라고 볼수있으나 음소열에 따라 상당히 차이가 있는것으로 보아 원자료의 크기와도 상관이 있는것으로 보인다. 그림3에 규칙에 의한 생성과 실제조사된 종류수 그리고 음소열의 전체 빈도수를 대비하여 보았다. 음소열의 발생빈도로 볼때는 CVC, bCV, CCV순이고, 종류수의 경우는 CVC, VCV, CCV순이다. 특히, CVC 와 VCV 가 전체 실현 음소열 종류의 60%를 차지하고 있으며 어두의 CV와 어말의 VC가 종류수에 비해 많은 빈도를 보이는 것이 특징으로 지적될 수 있다.

표3에 각 음소의 발생빈도를 보였다. 자음과 모음이 6:4의 비율로 나타나며 모음에서는 /ㅏ/, /ㅓ/가, 자음.반모음에서는 /j/, /ㄱ/, /W/, /ㅅ/이, 어말자음에서는 /o/, /ㄴ/이 각각 절반씩을 차지하고 있다.

전체적으로 보면 /ㅏ/, /j/, /ㅓ/, 어말의 /o/, /ㅗ/, 어말의 /ㄴ/, /ㅣ/순서이며, 이들 7개의 음소가 또한 전체빈도의 절반을 차지한다. 2 음소열의 빈도분포를 표4에 보였고 이들 2음소열중에서 앞 또는 뒤 음소계열이 전체의 절반에 해당하는 고빈도 음소군을 뽑아 표5에 정리하였다.

이 표에서 2음소열중 앞의 음소로는 단모음, 반모음, 자음 등 고루 나타나나 후속음소는 어말자음 /ㅇ/과 /ㄴ/을 제외하고는 모두 모음이다. 즉, 후속음소는 모음일 경우가 비교적 많다는 것을 나타내고 있다. 실제로 후속음소의 모음 및 자음간의 전체 비율을 구하여 보니 음소의 음소수가 9개로 자음에 비해 적지만 거의 비슷한 분포를 보였다.

표2. 음소열의 종류

음 소 열	규칙에 의한 생성	본 자료에 출현한 종류
2 음소열		
CV	28 x 9 = 252	185 (73.4)
VC	9 x 28 = 252	220 (87.3)
CC	7x21 + 19x2 = 185	144 (77.8)
VV	9 x 9 = 81	59 (72.8)
계	770	608 (79.0)
3 음소열		
bCC	19 x 2 = 38	19 (50.0)
bVC	9 x 28 = 252	116 (46)
bCV	21 x 9 = 189	164 (86.8)
VVb	9 x 9 = 81	56 (69.1)
CVb	28 x 9 = 252	154 (61.1)
VCb	9 x 7 = 63	50 (79.4)
CVC	28 x 9 x 28 = 7056	1698 (24.1)
VCV	9 x 28 x 9 = 2268	773 (34.1)
CCV	26x 21 x 9 = 4914	659 (13.4)
CVV	28 x 9 x 9 = 2268	185 (8.2)
VVC	9 x 9 x 28 = 2268	139 (6.1)
VVV	9 x 9 x 9 = 729	9 (1.2)
계	20378	4022 (19.7)

( )는 실현된 음소열 종류의 규칙생성에 대한 백분율

표3. 음소의 발생빈도

모음	빈 도 (%)	자음. 반모음	빈 도 (%)
/ㅏ/	12.76	/j/	12.76
/ㅓ/	5.71	/ㄱ/	4.35
/ㅗ/	4.56	/W/	3.55
/ㅣ/	4.53	/ㅅ/	3.32
/ㅐ/	3.85	/ㅎ/	2.90
/ㅜ/	3.56	/ㄴ/	2.73
/ㅡ/	1.34	/ㅁ/	2.72
/ㅔ/	0.84	/ㄷ/	2.54
/ㅚ/	0.34	/ㅁ/	2.20
	37.49	/ㅂ/	2.11
		/ㅊ/	1.87
		/ㅋ/	0.99
		/ㅍ/	0.87
어말자음		/ㄲ/	0.76
		/ㅉ/	0.65
/ㅇ/	4.80	/ㅌ/	0.56
/ㄴ/	4.53	/ㅆ/	0.46
/ㅂ/	2.74	/ㅃ/	0.42
/ㄱ/	2.24	/ㅄ/	0.35
/ㅁ/	1.46	/ㅋ/	0.07
/ㅅ/	0.78		45.06
/ㄷ/	0.29		
	16.84		

표5. 2음소열의 전후음소별 고빈도 음소군

(누적 백분율)

전		후	
/ㅏ/	10.5 %	/ㅏ/	12.6 %
/ㅓ/	18.3 %	/ㅓ/	20.7 %
/ㄱ/	24.5 %	어말/ㅇ/	27.6 %
/j /	30.1 %	어말/ㄴ/	34.1 %
/ㅅ/	34.9 %	/ㅗ/	40.5 %
/ㅣ/	39.1 %	/ㅣ/	46.7 %
/ㅎ/	43.3 %	/ㅜ/	51.7 %
/ㅜ/	47.4 %		
/ㄴ/	51.3 %		

4. 맺음말

지금까지 음성 DB용 대상 단어 선정을 위한 기초자료로서 소규모의 우리말 발음사전을 대상으로 음소 및 음소열 등의 통계적 분포에 관하여 검토하였다. 대상어휘수가 적어 통계적 자료로서 다소 부족함을 느끼지만 단어선정 알고리즘의 개발 및 평가에는 나름대로 유용한 자료로 생각한다.

앞으로 장모음등 음운환경의 보다 자세한 검토 및 대상어휘의 확대로 DB용 대상 단어 선정만이 아니라 minimal pair 나 음운결합에 따른 각종 신뢰성 있는 통계자료를 얻기 위해서 10만 단어이상이 수록된 발음사전을 컴퓨터에 입력할 계획이다.

감사의 글

본 연구는 ETRI의 기초연구의 일환으로 이루어진 것이다. 사전입력을 위해 수고해준 이 성구군과 조 화신양에게 감사의 뜻을 표한다.

참고 문헌

1. 이용주, 김경태, "연구 및 평가용 음성데이타베이스의 개발 동향과 제안"
전자통신 8권 3호(86.10)
2. 이용주, 김경태, "음성이해 연구의 동향"
전자통신 9권 1호 (87.3)
3. 일본전자공업진흥협회, "일본어 정보처리 표준화에 관한 조사연구보고서 (85.)
4. 강철희 외, "ETRI의 음성정보처리분야 연구 활동 소개"
음성통신 및 처리기술 Workshop논문집(86.5.31)
5. 이용주 외, "단어음성데이타의 수집 및 DB 구성 시스템"
전자공학회 추계학술대회 논문집
Vol.9 No.2 (86.12)
6. 鹿野 , "일본 음향학회 강연 논문집 3-3-10 (84.3)
7. 速水 외, "일본 음향학회 강연논문집 2-4-7 (85.10)
8. 横山 외 , 일본 전자통신학회 논문집
Vol J68-D No.12 (85.12)
9. 전영우, 표준한국어 발음사전, 한국방송사업단(84.2)
10. 허웅, 국어음운론, 정음사(85.2)
11. 문헌10 P.213 의 참고문헌을 재 인용
정철; "국어 음소 배열의 연구
12. 남광우외, 한국어 표준 발음사전
한국 정신문화연구원 (84.11)
13. Denes, "On the statistics of spoken English"
JASA 35-6 (63.6)

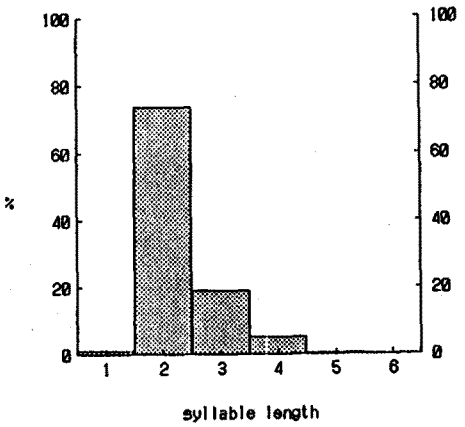

그림1. 음절길이별 단어분포

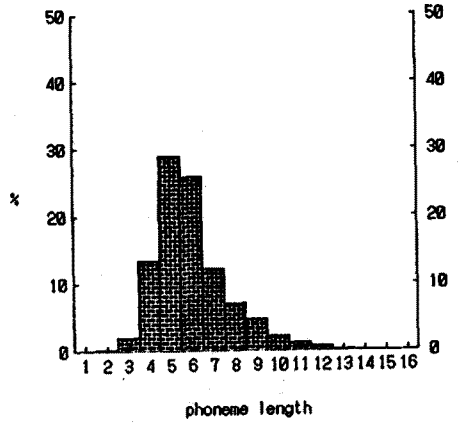

그림2. 음소길이별 단어분포

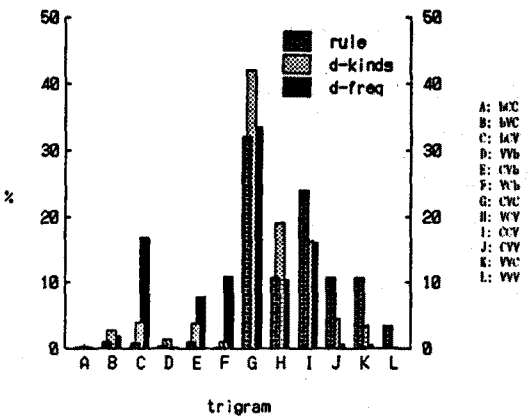

그림3. 3음소열, 종류수 및 발생빈도

표4. 2음소열의 빈도분포