한국어 음소분포에 대한 계량언어학적 연구
- 『소』와 『고도를 기다리며』를 중심으로 -

A Quantitative Study for the Distribution of Korean Phonemes in the two parts: *The Ox and Waiting for Godot*

배 희 숙* · 구 동 욱* · 윤 영 선* · 오 영 환*
Hee-Sook BAE · Dong-Ook Koo · Young-Sun Yun · Yung-Hwan Oh

Abstract

The goal of quantitative linguistics is to show the quantitative behavior of linguistic units. There are several studies which examine the frequency of Korean phonemes, which are important in comprehending the internal function of the linguistic units. However, the frequency information, from the pure phonological level without any consideration of rhythmic group, cannot adequately represent linguistic phenomena. Therefore, to provide the effective information, the phonological transcription must be carried out on the level of rhythmic group. In this paper, we made the transcription to analyze Korean phonology. We were not satisfied with merely investigating the frequencies of the phonemes, but also examined whether the distribution of Korean phonemes show the binomial distribution within linguistic constraints.

Keywords : quantitative linguistics, phoneme, frequency, binomial distribution

서 론

음소에 대한 계량언어학적 연구는 영어나 프랑스어의 경우 이미 오래 전부터 이루어져 왔다.[1] 계량언어학적 연구가 음운구조를 이루는 구성원들의 기능에 대해 명확한 자료를 제시할 수 있기 때문이다. 그러나 국내 계량언어학의 역사는 매우 짧으며,[2] 현재 이루어지고 있는 대부분의 연구는 어휘에 대한 빈도연구에 치중되어 있고, 한국어 음소에 대한 계량적 연구는 대단히 미미하다. 현재까지 발표된 한국어 음소의 기능부담량에 대한 몇몇 작업은 음소 기저형에 대한 순수음운론적 연구이다. 그러나 본 연구는 운율구를 고려한 발화음운론적[3] 차원에서 이루어졌으며, 바로 이점에서 앞선 연구와 차별성을 갖는다. 이

* 한국과학기술원 음성언어연구실

1) 계량언어학의 역사에 대해서는 Marc Hug(1980: 371-374)에 자세하게 설명되어 있다.

2) 임칠성(1997: 1-36) 참조.

3) "[...] 그러기에 이들에 대한 연구는 순수한 음운론의 임무가 아닌 것으로 볼 수도 있으

연구는 국어음운론 자체뿐만 아니라, 한국어 음성과학 분야의 기초작업을 위해서도 중요한 자료가 될 것이다.

음소 빈도수 분포를 연구하기 위해서는 우선 텍스트나 녹음된 발화체의 발음변환이 이루어져야 하고, 올바른 발음변환 작업을 위해서는 일관성 있는 규칙을 적용해야 한다. 이 규칙에 따라 음소의 수는 영향을 받을 것이고, 본 연구는 그 수에 대한 계량적 연구이기 때문이다. 본 논문의 1장에서는 본 연구가 의거한 발음변환규칙에 대해 명시하고, 2장에서 음소 빈도수의 분포를 조사할 것이다. 음소 빈도수 분포에 대한 연구는 서로 다른 음운규칙 적용에 따른 결과를 비교·분석하는 방법, 코퍼스 내의 여러 표본들을 비교하여 문체나 테마를 분석하는 방법 등이 있을 것이나, 본 연구에서는 음소의 기본 빈도수에 입각해서, 한국어 음소가 이항분포를 따르는지 조사하는데 그 목적을 두고자 한다.[4] 코퍼스는 두 개의 텍스트로 구성된다. 유치진의 희곡 『소』[5]와 사뮈엘 베케트(Samuel Beckett)의 『고도를 기다리며 En attendant Godot』의 한국어 번역본[6]이 그것이다.

1. 발음변환 규칙

발음변환은 일반적으로 음성학적 발음변환과 음운론적 발음변환으로 대립된다. 그런데 본 연구를 위한 발음변환은 순수 음운론적 발음변환이 아니라, 운율구 단위로 이루어진 발화음운론적 차원의 발음변환이다. 사실, '국물'을 기저형 /ㄱㅜㄱㅁㅜㄹ/이 아니라 음운론적 조음현상을 고려한 [ㄱㅜㅇㅁㅜㄹ]로 발음변환하였다. 하지만, 음운론적으로 이 음소환경에는 /ㄱ/가 올 수 없다. 이 위치에서는 항상 /ㅇ/이 될 수밖에 없으며, 같은 음소환경에서 /ㄱ/를 제시하는 다른 단어도 없다. 또한, 우리말 '코'에서의 [ㅋ]와 '키'에서의 [ㅋ]는, 전자는 후자보다 구강의 뒷부분에서 발음될 것이므로, 음성학적으로 다른 두 개의 음이다. 그러나 이 두 음은 하나의 음소 /ㅋ/로 대표된다. 본 연구는 異音이 아니라 그 모든 음의 대표인 음소만을 다루었으므로 발화음운론적 차원에서 발음변환된 음소에 대한 연구인 것이다.

한국어는 표음문자이므로 한국어의 발음변환은 인도유럽어에 비해 상당히 편리한 것이 사실이다. 원칙적으로 표기대로 발음하면 되기 때문이다. 표기와 발음이 일치하지 않는 경우에 대해서는 92년 발표된 국립국어연구원의 ≪표준발음법≫에 30개 항에 걸쳐 훌륭하게 정리되어 있다. 그러나 다음에 열거되는 몇 가지 문제에 있어서는 조정이 필요하였다.

나, 음적 현상이 관여하는 것은 사실이어서 넓게는 음운론, 특히 발화음운론의 임무라 할 수 있다."(이병근 & 최명옥, 1999: 7)

4) 프랑스어 음소의 경우, 음소가 이항분포를 따른다는 연구 결과가 1979년 Marc HUG에 의해 발표된 이후, 프랑스 통계언어학 분야에서는 이를 일반적 사실로 받아들이고 있다.

5) 텍스트 유치진의 『소』는 http://user.chollian.net/~javanet/play/so.zip에서 내려 받기한 것임을 밝힌다.

6) 사뮈엘 베케트의 『고도를 기다리며』는 오증자 역으로 www.kcaf.or.kr/hyper/Kdrama_main. html에서 내려받기한 텍스트를 사용했음을 밝힌다.

1.1 음소목록

한국어 음소목록을 정하는 데 있어서, 가장 큰 문제는 이중모음에 대한 견해차이다. 한국어 이중모음을 반자음과 단모음의 결합으로 본다면, 음소목록은 분명히 줄어든다. 그리고 국제 음성기호로 표기하면 이중모음은 분명히 두 개의 음소로 이루어져 있다. 그러나 국립국어연구원을 비롯한 한국 음운론계에서 제시하는 한국어 음소목록은 이중모음을 고유한 음소로 간주하고 있다. 또한, /ㅚ/와 /ㅟ/가 단모음인지 이중모음인지에 대한 이견도 간과할 수는 없는 일이다.

본 연구를 위한 발음변환은, 우선 한국과학기술원 음성언어연구실의 TTS시스템[7]을 이용하여 자동 변환한 다음, 운율구 단위로 일일이 수정하였다. 이 시스템은 종성에 올 수 있는 일곱 가지 자음을 따로 구분하고 있다. 종성의 내파음들은 그 위치가 고정되어 상보적 분포를 보이는 이음들일 뿐 의미에 관여하지는 않는 것이지만, 음성합성을 위한 이 변환기가 이러한 이음들을 따로 구분한 것은 타당하다. 자동 변환기에 의해 분리 조사된 이음들은 통계 작업시 조정되어 계산될 것이다.

1.2 이중모음화

음절의 연결부위에서 /ㅣ/모음이나 /ㅜ/모음 등에 연이어 단모음이 발음될 때 청각적으로 반자음을 감지하게 된다. 예를 들어 텍스트『소』에는 '불시에'라는 어절이 있다. 각 음절을 발음하면 [불], [시], [에]이지만, 세 음절을 연달아 발음하면 [불씨예]가 된다. [ㅖ]에서 [ㅖ]로 이중모음화가 일어난 것이다. 이는 두 모음의 충돌을 피하기 위해 반자음이 삽입되는 것으로 이해할 수 있을 것이다. 하지만, 두 개의 목표음 사이에 전이부가 있는 영어의 이중모음과는 달리, 우리말의 이중모음은 전이부와 종결부분으로만 이루어진 비음소적 이중모음이다. 사실, 이중모음은 천천히 발음하면 두 개의 모음이지만 빨리 발음하면 반자음이 개입되면서 이중모음으로 인식된다. 양병곤 교수는 음향학적 실험연구 결과[8]에 따라 한국어 이중모음을 지속시간에 입각해서 판단할 때, 이중모음이라기 보다는 [자음+모음]으로 보는 것이 타당할 것이라는 조심스런 결론을 내리고 있다.

'불시에'의 경우, 반자음의 첨가 현상으로 처리해야 할까? 아니면 단모음 /ㅖ/가 이중모

7) 본 연구에서는 KAIST 음성언어 연구실의 TTS시스템을 위한 음소변환 프로그램을 사용하였다. 프로그램에 적용된 자세한 규칙은 이상호(1997: 33-39)를 참조할 수 있다. 본 연구를 위해 이 프로그램을 이용한 것은 텍스트의 음소변환을 가능하면 자동으로 처리하기 위한 것이다. 그러나 언어학적 차원에서 몇 가지 문제가 있는 까닭에 자동 변환 이후에 수동으로 일일이 수정함으로써 정확성을 기하였다.

8) "이중모음의 지속시간을 모음별로 평균하여 본 결과 110 msec이고 표준편차의 평균은 33 msec이었다. 그 범위는 모음 [여]의 99 msec에서 모음 [예]의 122 msec로써 23 msec의 범위에서 발성했다. 또한 [w]그룹을 평균하여 보면 112 msec 이고 [j]그룹의 평균은 111 msec로써 두 그룹간에 뚜렷이 구분하여 유형화할 차이는 없는 것으로 여겨진다. 이러한 지속시간은 Liberman(1956) 등의 영어 지각실험에서 200-230 msec 근처에 [자음+모음]의 경계가 나타난 것에 비추어볼 때 매우 짧은 시간이고, 따라서 이중모음이라기보다는 [자음+모음]으로 보는 것이 타당할 것이다. 하지만 한국인들의 지각체계가 다를 수도 있으므로, 이것은 앞으로 지각실험을 통해 밝혀야 할 과제이다." (양병곤, 1993: 3-26)

음 /ㅖ/로 변한 것일까? 이에 대한 국어학계의 입장과 음성학계의 입장은 차이를 보인다. 국어학계에서는 /ㅑ/를 하나의 모음으로 다루고, 음성학계에서는 분명히 두 개 음소의 결합인 [ja]로 본다. 본 연구는 이에 대한 결정을 유보하고 이중모음화에 대해 고려하지 않았음을 밝힌다. 따라서 개별 음소의 빈도수 계산 결과를 분석할 때, 이중모음의 상대적 빈곤에 대한 고려가 필요하다.

1.3 운율구 결정

국립국어연구원의 표준 발음법은 형태음운론적 단계에서만 다루어지고 있으나, 우리가 다루어야 할 텍스트들은 어절의 단순한 연결이 아니라 문장들이다. 따라서 운율구를 결정한 후 발음변환이 이루어져야 한다. 간단히 어절단위 발음변환과 운율구 단위 발음변환의 결과를 비교해 보자.

죽을 놈이 [주글 노미] / [주글로미]
이웃 사람 [이욷 사람] / [이욷싸람]
이것 봐요 [이걷 봐요] / [이걷뽜요]
사갈 사람이 [사갈 사라미] / [사갈싸라미]
세상 일은 [세상 이른] / [세상니른]
아들 낳기는 [아들 나키는] / [아들라키는][9]

이 예들에서 알 수 있는 바와 같이 "같은 단어라도 운율구 내에서의 위치에 따라 발음이 달라질 수 있다"[10]. 따라서 텍스트의 발음변환은 꼭 운율구 단위로 이루어져야 한다.

프랑수아 비올랑(François Wioland)에 따르면, 프랑스어의 경우 일상 대화체에서 평균 2.5음절로 운율구가 이루어지고 6음절 이상은 예외적이다(Wioland, 1983: III). 한국어의 경우, KAIST 음성언어연구실의 음성파일에 의거해 계산해 보면, 교과서를 읽어 나가는 문체에서 운율구당 평균 음절수는 6음절이며 11음절 이상은 조사된 전체 운율구의 4.5% 내로 예외적이다. 물론 일상 대화체를 분석하면 분명 이와는 차이점을 보일 것이다.

텍스트를 읽을 때 운율구를 결정하는 주요 요인은 문장성분과 음절수이다. 운율의 기능이란 의사전달이 정확하게 이루어지도록 하는 것이기 때문에 문장을 구성하는 성분들의 구조는 운율구 결정에 직접적인 영향을 준다.[11] 본 연구에서는 이를 수동으로 처리하였다.

1.4 어절 꼬리에서 불파음 [ㄱ], [ㄷ], [ㅂ]의 존재

한국어 표준 발음법 23항과 30항에 관련하여 음성학적·음운론적으로 논란의 여지가 있다. 표준 발음법 23항은 "받침 'ㄱ(ㄲ, ㅋ, ㄳ, ㄺ), ㄷ(ㅅ, ㅆ, ㅈ, ㅊ, ㅌ), ㅂ(ㅍ, ㄼ, ㄿ,

9) 이 예들은 모두 유치진의 『소』에서 발췌된 것이다.
10) "Un meme mot peut se prononcer differemment selon sa position dans l'unite rythmique : il n'existe pas en tant que tel dans ce cadre" (Fr. Wioland, 1983: III).
11) 음향적 특성과 문법정보와의 관계를 통해 운율구의 특성을 파악함으로써 운율구를 자동으로 추출하려는 연구로는 성철재(1996), 김선미(1997) 등이 있다.

ㅄ)' 뒤에 연결되는 'ㄱ, ㄷ, ㅂ, ㅅ, ㅈ'은 된소리로 발음한다"라고 규정하고 "국밥[국빱], 깍다[깍따]" 등 20개의 예를 제시하고 있다. 그러나 '학교', '익고', '먹고' 등과 같이 받침과 뒤에 연결되는 자음이 모두 'ㄱ'인 예는 언급되지 않고 있다.[12] '학교'는 조음 위치가 같은 장애음이 계속되면 앞 장애음이 탈락한다는 음운규칙에 따라 [하꾜]로 발음해야 할까? 아니면 '학교'와 같은 음소환경에서 의미전달에 관여적인 경우들을 고려하여 [학꾜]로 발음해야 할까? 한편, 표준발음법 30항 1은 "'ㄱ, ㄷ, ㅂ, ㅅ, ㅈ'으로 시작하는 단어 앞에 사이시옷이 올 때는 이들 자음만을 된소리로 발음하는 것을 원칙으로 하되, 사이시옷을 [ㄷ]으로 발음하는 것도 허용한다."고 규정하고 있다. 즉 사이시옷을 [ㄷ]로 발음해야 하는지 뒤의 경음화된 자음에 합쳐진 것으로 고려해야 하는지에 대해서는 화자의 선택으로 남기고 있는 것이다. 예를 들어 '콧등'은 [코뜽]으로 발음해도 되고 [콛뜽]으로 발음할 수도 있다.

청각적으로는 구분이 어렵지만,[13] 본 연구의 음소변환에서는 이들 종성자음의 흔적을 인정하였다. 사실, '있다'와 '이따'에서 종성자음은 변별적 자질을 가지며, 본 연구는 음운론적 차원의 발음변환을 하였기 때문이다.

이렇게 발음변환 작업에서 제기되는 운율구, 이중모음화에 따른 반자음 첨가, 경음화되는 초성 앞의 종성 파열음 존재 여부 등에 대한 문제점들을 해결하기 위해서는 음성학적 실험을 거쳐야 한다. 본 연구에서 이를 명확하게 해결하지 않고 조정·적용하는데 만족하였으며, 이는 아쉬움으로 남는다. 하지만 계량언어학적 연구에서는 어떤 규칙을 적용하느냐의 문제보다는, 얼마나 일관성 있게 문제들을 조정·적용하였으며, 적용된 규칙이 무엇인가를 명확하게 제시해 주는 것이 더욱 중요하다.

2. 빈도수

이제 두 텍스트는 음소로 변환되었다. 그 결과로 얻어진 음소코퍼스는 자음과 모음의 비율, 평균 빈도수, 분산, 음소간 상관관계, 텍스트에서 추출된 2 음소, 3 음소의 출현 확률 등 한국어 음소연구를 위한 기초 자료를 제공할 것이다. 또한 이를 기초로 하여 한국어 음소가 언어적 제약에도 불구하고 이항분포를 따르는지에 대해 조사할 수 있다.

2.1 음소의 빈도수 순서

도표 1에서 제시되는 개별 음소의 빈도수는 발화음운론적 차원에서 음소들이 한국어의 발음 구조에서 양적으로 어떤 기능을 담당하고 있는지 그 중요성을 한눈에 볼 수 있게 한다.

---

12) "뒷음절의 초성이 된소리나 거센소리로 날 때 앞 음절의 (...)내파음은 표기하지 않았다"(표준한국어발음대사전: xviii). 그러나 국립국어연구원의 표준발음법 23항의 예 중에 '있던'의 발음은 [일떤]으로 제시된다.

13) There are phonological and morphological contrast between a single and double plosives as in "아끼다"[akida], "악기다"[akkida] "this is an instrument". However, many Koreans are likely to perceive clusters [pp, tt, kk] as single consonant, thus making no distinction between a single and double plosives. (Lee Hyun-Bok, 1999: p. 3)

도표 1. 음소의 빈도수

	소			고도를 기다리며		
	절대빈도	상대빈도(%)	순위	절대빈도	상대빈도(%)	순위
/ㅂ/	1371	2.176	15	1692	1.986	15
/ㄷ/	2818	4.474	9	4858	5.703	7
/ㄱ/	4297	6.822	5	6249	7.336	4
/ㅈ/	2214	3.515	12	2700	3.170	11
/ㅃ/	162	0.257	33	170	0.200	34
/ㄸ/	784	1.245	20	845	0.992	21
/ㄲ/	735	1.167	21	1061	1.246	19
/ㅉ/	373	0.592	26	530	0.622	26
/ㅍ/	299	0.475	29	455	0.534	28
/ㅌ/	451	0.716	25	626	0.735	25
/ㅋ/	310	0.492	28	634	0.744	24
/ㅊ/	600	0.952	23	743	0.872	23
/ㅅ/	1872	2.972	13	2356	2.766	13
/ㅆ/	680	1.079	22	757	0.889	22
/ㅎ/	1003	1.592	18	1381	1.621	17
/ㄹ/	5270	8.365	3	7664	8.997	3
/ㄴ/	6382	10.131	2	8296	9.739	2
/ㅁ/	2919	4.634	8	3805	4.466	10
/ㅇ/	1336	2.121	16	1326	1.557	18
/ㅏ/	6518	10.347	1	9564	11.227	1
/ㅓ/	3888	6.172	6	4524	5.311	8
/ㅗ/	2663	4.227	10	4106	4.820	9
/ㅜ/	2262	3.591	11	2414	2.834	12
/ㅡ/	3516	5.581	7	5566	6.534	6
/ㅣ/	4668	7.410	4	6187	7.263	5
/ㅔ/	1705	2.707	14	2131	2.502	14
/ㅐ/	1226	1.946	17	1419	1.666	16
/ㅑ/	342	0.543	27	513	0.602	27
/ㅕ/	889	1.411	19	1055	1.238	20
/ㅛ/	489	0.776	24	278	0.326	30
/ㅠ/	55	0.087	36	56	0.066	37
/ㅖ/	18	0.029	38	29	0.034	39
/ㅒ/	27	0.043	37	58	0.068	36
/ㅘ/	252	0.400	30	380	0.446	29
/ㅝ/	143	0.227	34	222	0.261	31
/ㅟ/	165	0.262	32	175	0.205	33
/ㅢ/	190	0.302	31	180	0.211	32
/ㅙ/	100	0.159	35	143	0.168	35
/ㅞ/	3	0.005	39	36	0.042	38
	62995			85184		

도표 2.

소	자음	ㄴ, ㄹ, ㄱ, ㅁ, ㄷ, ㅈ, ㅅ, ㅂ, ㅇ, ㅎ, ㄸ, ㄲ, ㅆ, ㅊ, ㅌ, ㅉ, ㅋ, ㅍ, ㅃ
	모음	ㅏ, ㅣ, ㅓ, ㅡ, ㅗ, ㅜ, ㅔ, ㅐ, ㅕ, ㅛ, ㅑ, ㅘ, ㅚ, ㅟ, ㅝ, ㅞ, ㅒ, ㅖ, ㅠ, ㅢ, ㅙ
고도	자음	ㄴ, ㄹ, ㄱ, ㄷ, ㅁ, ㅈ, ㅅ, ㅂ, ㅎ, ㅇ, ㄲ, ㄸ, ㅆ, ㅊ, ㅋ, ㅌ, ㅉ, ㅍ, ㅃ
	모음	ㅏ, ㅣ, ㅡ, ㅓ, ㅗ, ㅜ, ㅔ, ㅐ, ㅕ, ㅑ, ㅘ, ㅛ, ㅝ, ㅚ, ㅟ, ㅐ, ㅖ, ㅠ, ㅢ, ㅔ

도표 2는 빈도수의 내림순으로 음소들을 나열한 것이다. 1930년대에 한국어로 쓰여진 작품과 1990년대의 번역 작품이라는 차이에도 불구하고, 두 텍스트에서 가장 많이 사용된 음소나 가장 적게 사용된 음소의 목록은 대체로 비슷했다. 도표를 관찰하면 두 텍스트 모두에서 가장 많이 쓰인 여섯 개의 모음은 /ㅏ/, /ㅣ/, /ㅓ/, /ㅡ/, /ㅗ/, /ㅜ/이고, 자음은 /ㄴ/, /ㄹ/, /ㄱ/이다.[14] 이 결과는 중세국어의 기본자가 '아', '이', '오', '우', '으', '어'였다는 점을 상기할 때 더욱 흥미롭다. 바로 이 기본자와 같은 6모음이 두 텍스트 모두에서 가장 높은 빈도수를 보이는 것이다. 『소』에서는 이 기본 6모음이 전체 모음 빈도수의 80.77%를 차지하고 『고도를 기다리며』에서는 82.90%를 차지한다. 이중모음을 반자음과 단모음의 결합으로 생각한다면, 이 기본 모음의 비율은 더욱 높아질 것이다.

이제, 두 텍스트의 빈도수에 따른 음소의 순서를 가지고 스피어만(Spearman) 상관계수[15]를 구해보자. 전체 음소의 경우 +0.9899이다. 자음과 모음을 구분해서 상관계수를 구하면, 자음은 +0.9895, 모음은 +0.9988이다. 이는 음소들이 그 내부에서 담당하는 역할이 텍스트와 관계없이 위의 빈도수 순서에서 크게 벗어나지 않음을 의미한다. 두 텍스트의 차이를 생각할 때, 더욱 그러하다.

음소의 빈도수 순서처럼, 음소들의 빈도수도 두 텍스트에 유사하게 분포되어 있을까? 이에 대해 카이제곱검정을 통해 알아보면, 757.313의 χ^2 값을 얻게 된다. 자유도 38에서 가우시안 변환편차는 30.2579이다. 이 분포가 우연에 근거할 확률은 거의 0에 가까움을 알 수 있다. 이는 전체적으로 한국어 구조에서 음소가 담당하는 기능은 텍스트에 관계없이 같은 양상을 보이지만, 개별 음소의 빈도수 자체는 텍스트의 성격에 따라 역동적으로 반응한다는 것을 의미한다. 구체적으로 두 텍스트에서 이론빈도수와 큰 차이를 보이는 음소들의 분포는 분명 텍스트의 문체나 테마와 밀접한 관계가 있을 것이고, 이는 등장인물 별 연구를 통한 문체연구에서 다루어져야 할 것이다.

2.2 자음과 모음 비율

한국어 담론 구성에서 자음과 모음은 어떤 비율로 이루어졌을까? 러시아어의 경우, 유명한 수학자 마르코프(Markov)가 대문호 프쉬킨(Sergeevitch Pushkin)의 걸작시 『에프게니 오네긴(Evgenij Onegin)』에 쓰인 20,000개 음소를 연구한 결과, 모음은 8,638개 자음은 11,362개로 약 43% 대 57%의 비율을 보였다.[16] 프랑스어의 경우, 1939년 지프와 로저스

14) 참고로 프랑스어의 경우, 가장 많이 사용되는 자음은 순서대로 [r], [s], [t]이고 모음은 [a], [e/E], [i]이다.

15) 스피어만 상관계수는 공식 $\frac{6\sum d^2}{n(n^2-1)}$에 의해 구해지며, 여기서 'd'는 비교된 두 텍스트의 음소순서의 차이이고, 'n'은 조사된 음소의 수이다.

(Zipf & Rogers)의 연구는 44.06%의 모음비율을, 1979년 비올랑(Wioland)의 연구는 43.45%의 모음비율을 보여주고 있으며, 위그(Hug, 1979: 21-27)의 연구 결과에 의하면 텍스트에 따라 모음 비율은 42%에서 45% 사이에서 변화를 보인다. 필자의 연구 결과에 따르면 프랑스어의 모음 비율은 41.76%에서 43.77%의 변화를 보였으며, 이는 적용된 음소변환 규칙에 따른 변화이기보다는 문체나 텍스트의 성격에 따른 변화였다(배희숙, 1997: 263-275).

도표 3. 자음과 모음의 비율

	총 음소	자 음		모 음	
소	62,995	33,882	53.785	29,113	46.215
고도를 기다리며	85,184	46,088	54.174	39,036	45.826

도표 3을 보자. 두 작품에서 자음과 모음의 비율 차이는 크지 않다. 두 텍스트 간의 모음과 자음의 수에 대해 카이검정을 해보면, 1d.d.l.에서 $\chi^2 = 1.857$로 나타난다. 이는 자음과 모음의 분포에 있어서 두 작품이 유사 분포를 보여주고 있음을 의미한다. 이 현상이 일관성이 있는지를 보기 위해 『소』와 『고도를 기다리며』를 대사와 지문으로 나누어 각각의 집단에서 모음과 자음 비율을 조사하였다. 그 결과, 『고도를 기다리며』의 지문을 제외하고는 모두 54%와 46%의 비율을 나타냈다. 『고도를 기다리며』의 지문에서는 자음의 비율이 54.75%로 상대적 상승을 보이지만 그 차이의 폭은 그리 크지 않다. 따라서 대체로 한국어 텍스트에서 자음과 모음의 비율은 평균 54%와 46%정도이며 변화 폭은 대단히 작다는 것을 알 수 있다.

텍스트의 문체에 따라서 적지 않은 변화를 보였던 프랑스어에 비추어 볼 때 이러한 결과는 한국어 음절구조와 관계된 특성이라는 예측이 가능하다. 젊은이들이 사용하는 언어에서 이 구조가 깨진다는 지적도 있긴 하지만, 현대 한국어 음절구조는 (C)V(C)로 단 네 가지 경우밖에 없다. 이에 대한 결론을 내리기 위해서는 다양한 텍스트를 통해 모음과 자음의 비율, 한국어 음절구성에 대한 연구, 그리고 운율구조 내에서 음절구조 특성의 약화가 이끌 수 있는 음운현상에 대한 연구 결과를 필요로 한다. 아울러 한국어 음소의 자음과 모음의 조화를 유지하기 위한 내적 구조장치가 어디에 있는 것인지 알아보는 것도 의미있는 일일 것이다.

2.3 音素群의 빈도수

2.3.1 조음방식에 따른 자음군 분포

한국어 자음을 조음방식에 따라 분류하면 크게 폐쇄음(/ㅂ, ㅍ, ㅃ/, /ㄷ, ㅌ, ㄸ/, /ㄱ, ㅋ, ㄲ/), 마찰음(/ㅅ, ㅆ, ㅎ/), 파찰음 (/ㅈ, ㅊ, ㅉ/), 비음 (/ㄴ, ㅁ, ㅇ/), 설측음 (/ㄹ/)이 있다. 자음의 경우, 이렇게 조음방식에 따라 분류하는 것은 소리의 성질을 잘 반영하므로 조음방식에 따라 분류된 자음군의 빈도수가 어떠한지 조사하는 것은 소리의 성질에 따른 음소 빈도수 분포를 알게 해준다. 훈민정음에서 조음 방식에 따라 자음을 '전청(평음), 차

16) 이승명(1984)의 pp. 41-42 참조.

청(유기음), 전탁(경음), 불청불탁(유성음)'으로 나눈 것도 조음방식에 따른 분류가 소리의 느낌을 잘 분류하고 있음을 보여준다.

도표 4

	소			고도를 기다리며		
	전체	지문	대사	전체	지문	대사
/ㄱ, ㄷ, ㅂ/	25.045	25.483	26.464	27.737	28.295	27.375
/ㄲ, ㄸ, ㅃ/	4.961	5.342	4.966	4.499	3.368	5.203
/ㅍ, ㅌ, ㅋ/	3.129	2.939	3.394	3.716	5.066	2.963
/ㅈ, ㅊ, ㅉ/	9.406	9.332	10.850	8.609	8.812	8.830
/ㅅ, ㅆ, ㅎ/	10.492	10.470	11.309	9.739	9.706	9.723
/ㄹ/	15.554	15.576	16.140	16.607	19.127	15.068
/ㄴ, ㅁ, ㅇ/	31.394	30.858	34.402	29.096	26.255	30.838

도표 4는 두 텍스트와 각 텍스트의 지문과 대사에 조음방식에 따라 구분된 자음군이 어떤 비율로 분포되어 있는지 백분율로 보여준다. 이 자음군의 빈도수 분포의 전체적인 양상, 즉 빈도수 순서는 여섯 개의 집단에서 모두 유사하다. 가장 많은 비율을 차지하는 음소군은 폐쇄음 계열이다. 폐쇄음 내부에서 음소의 분포를 살펴보면, 평음이 폐쇄음 전체에서 차지하는 비율은 60% 이상이다. 프라그망(fragment)[17]에 따라 조금씩 다르기는 하지만, 경음과 격음은 큰 차이를 보이지 않는다. 폐쇄음군 다음으로 많은 비율을 차지하는 음소군은 비음이다. 아홉 개의 음소로 이루어진 폐쇄음군에 비해 단 세 개의 비음이 한국어 음소체계에서 차지하는 비율은 매우 높다. 마찰음과 파찰음이 모두 합하여 단 하나인 유음의 수와 비슷한 것도 눈여겨보아야 할 것이다.

조사된 음소 목록에는 없는 이음이지만, 일곱 개의 종성자음의 비율을 조사하면, 『소』에서 26.35%, 『고도를 기다리며』에서 25.13%이다. 그리고 종성자음의 반 이상을 [ㄴ]과 [ㄹ]이 차지하고 있다. 일곱 개의 종성자음 중 가장 높은 빈도를 보이는 것은 [ㄴ]로 『소』에서 34.6%, 『고도를 기다리며』에서 39.4%를 차지했다.

2.3.2 조음위치에 따른 모음군의 분포

도표 5. 조음위치에 따른 모음군의 상대빈도수

	소			고 도		
	전체	지문	대사	전체	지문	대사
/ㅣ, ㅔ, ㅐ/	26.10	25.84	26.54	24.94	23.39	25.85
/ㅓ, ㅡ, ㅏ/	47.82	47.21	48.03	50.35	52.70	47.83
/ㅜ, ㅗ/	16.92	18.69	15.44	16.70	17.97	16.00
나머지	9.16	8.26	9.99	8.01	5.93	9.21

17) 프라그망이란 무작위로 추출된 표본과는 달리, 텍스트 자체의 성질에 의해 자연적으로 구분되는 부분집단에 의해 구성된 표본을 말한다.

모음의 소리성질은 구강의 앞쪽에서 나는 소리일수록 그리고 위쪽에서 나는 소리일수록 날카롭고, 뒤쪽에서 나는 소리일수록 그리고 아래쪽에서 나는 소리일수록 둔탁하다. 이러한 소리의 성질에 따라 조음위치별로 모음을 구분하여 그 빈도수를 조사하면 도표 5와 같다.

한국어 모음에서 가장 높은 빈도수를 보이는 음소군은 중설모음 /ㅏ/, /ㅓ/, /ㅡ/이다. 전체 모음의 47%에서 52%를 이 중설모음이 차지한다. 다음은 전설모음이다. 후설모음이 차지하는 비율은 단모음 중에서 가장 낮다. 중세국어에 대한 계량언어학적 연구자료가 없어 비교를 할 수는 없는 까닭에 현대 한국어의 전설모음화 현상이 있는지는 알 수 없지만, 위 도표에 따른 빈도수 분포는 분명히 후설모음이 상대적으로 빈약한 양상을 보여준다.

2.3.3 이중모음

이중모음과 단모음의 분포를 보면, /ㅟ/와 /ㅢ/를 이중모음으로 분류할 때 『소』에서는 모음의 약 91.03%가 단모음이고 약 9%만이 이중모음이다. 『고도를 기다리며』에서는 92.03%의 단모음과 8%의 이중모음으로 구성되었다.

도표 6.

	소	고도	TOT
[j]계	1820	1989	3809
[w]계	853	1136	1989
TOT	2673	3125	5798

이중모음을 [j]계와 [w]계로 분류하여 그 빈도수를 관찰하면 [j]계가 68.09%와 63.65%이고 [w]계가 31.91%와 36.35%이다. [w]계에 단모음으로도 분류되는 /ㅟ/와 /ㅢ/가 포함되어 있음을 상기하면 [w]계의 빈도수는 [j]계열에 비해 매우 낮은 편이다. 한편, 전체 음소 중에서 가장 낮은 빈도수를 보여주는 음소는 [w]계의 /ㅢ/이다. 『소』에 비해 『고도를 기다리며』에서 이 음소의 빈도수가 높은 것은, 『고도를 기다리며』에 '의자'라는 단어가 반복적으로 나타나기 때문이다. 일반적으로 모음 /ㅢ/가 이렇게 드물게 나타나는 것은, 조사로 쓰인 '의'는 [ㅔ]로, 첫음절이 아닌 다른 경우에는 [ㅣ]로 발음될 수 있기 때문이다. 즉 이 음소는 기타 다른 음소로 흡수되고 있는 것이다. 사실, 모음 /ㅢ/는 한국어 이중모음 중에서 단 하나 남아있는 하향성 이중모음으로서 매우 불안정한 양상을 보이는 음소인 것이다.

2.4 한국어 음소 분포는 이항법칙을 따르는가?

한국어 음소가 이항분포를 따를까? 음소의 분포가 주머니에 여러 가지 색깔의 공을 넣고 임의로 공을 하나씩 추출하여 얻는 분포와 유사할까? 실제 언어 연쇄에서 처음에 모음 /ㅏ/가 나왔다면 연이어 또 모음 /ㅏ/가 나오는 일은 매우 드물 것이고, 모음 /ㅏ/가 연이어 세 번이나 나타나는 일은 더욱 드물 것이다. 한국어의 경우, 감탄적 쓰임이 아니라면

[ㅏㅏ]나 [ㅏㅏㅏ]가 나타나는 일은 거의 없을 것이다. 이것이 바로 언어적 제약이다. 이러한 언어적 제약을 고려하면 음소의 추출이 주머니에서 빨간 공을 뽑는 것과 같을 수는 없다. 그러나 이러한 언어적 제약 없이 음소들이 단조롭게 분포되어 있다는 가정을 한다면, 주머니에서 임의로 음소를 추출할 때 음소 /ㅏ/를 뽑을 확률은 주머니에서 서로 다른 색깔을 지닌 공들 중에서 빨간 공을 꺼낼 확률과 같을 것이다.

두 텍스트를 구성하는 음소들이 언어적 제약 없이 단조롭게 분포되어 있을 것이라는 귀무가설을 세우고, 이항법칙에 근거하여 모델을 세워보면, 음소들이 나타날 확률은 이론적으로 $P_k = C_n^k p^k q^{n-k}$이다. 이 확률값에 조사된 블럭의 총수를 곱하면, 이론적으로 해당 음소가 n개 중 k개 나타날 블럭의 수가 된다. 음소연쇄를 n개씩 끊어 블럭을 구분하고 그 블럭들 안에서 음소의 출현 빈도수가 k번인 모든 경우의 수를 조사하여 모델과 비교하면, 한국어 음소분포가 이항분포에 비해 어떤 분포를 보이는지 알 수 있다.

도표 7.

x_i	소 (총 629 블럭)			고도를 기다리며 (총 851 블럭)		
	r_i	c_i	χ_i^2	r_i	c_i	χ_i^2
1	0	0.1310	*	0	0.0724	*
2	1	0.7486	*	0	0.4535	*
3	4	2.8222	*	1	1.8736	*
4	7	7.8987	0.0138	6	5.7461	0.1611
5	12	17.5027	1.7300	18	13.9326	1.1874
6	28	31.9835	0.4961	28	27.9390	0.0001
7	46	49.5685	0.2569	37	47.4487	2.3009
8	67	66.5041	0.0037	74	69.7590	0.2578
9	88	78.4591	1.1602	93	90.1837	0.0879
10	91	82.4014	0.8973	107	103.7892	0.0993
11	81	77.8097	0.1308	106	107.3952	0.0181
12	60	66.6028	0.6546	108	100.7341	0.5241
13	58	52.0333	0.6842	84	86.2380	0.0581
14	40	37.3183	0.1927	70	67.7755	0.0730
15	16	24.6933	3.0605	40	49.1431	1.7011
16	10	15.1401	1.7451	35	33.1075	0.1082
17	13	8.6339	2.2079	18	20.6328	0.3360
18	5	4.5948	0.3039	13	12.0322	0.0778
19	1	2.2886	*	5	6.5673	0.8797
20	0	1.0697	*	3	3.3638	*
21	0	0.4703	*	1	1.6206	*
22	1	0.1949	*	0	0.7360	*

N.B. 도표의 첫 행 x_i는 100개의 음소 덩어리로 이루어진 블럭에 /ㅏ/가 나타나는 수이다. r_i는 x_i에 해당되는 실제 블럭의 수, c_i는 모델에 따른 블럭의 수를 나타낸다.

$$\chi_i^2 = \frac{(r-c)^2}{c} \quad \chi^2 = \Sigma \chi_i^2$$

음소 중에서 가장 높은 빈도수를 보이는 /ㅏ/를 예로 들어보자. 『소』에서 이 음소의 상대빈도수는 0.10347이다. 따라서 p는 0.10347이고 q는 1-p이다. 그리고 100개의 음소로 이루어진 블럭에 /ㅏ/가 단 한 번 나타날 확률은 공식 $P_1 = C^{1}_{100} p^{1} q^{99}$에 대입하여 얻을 수 있고, 이 확률에 조사된 블럭의 총수를 곱하면, /ㅏ/가 단 한 번 나타나는 블럭의 수가 된다. 도표 7은 이런 방식으로 음소 /ㅏ/에 대해 계산하여 얻은 결과와 실제 분포를 모두 보여준다. 도표에 따르면, 음소 /ㅏ/의 실제 분포는 이항분포에 따른 모델과 큰 차이를 보이지 않는다. 그러나 정확성을 기하기 위해 이론값과 실제값 사이의 차이에 대한 카이제곱검정을 해보자. 이론빈도수가 지나치게 작은 경우들은 결합하면서 카이제곱검정을 하면, 『소』의 경우 χ^2는 13.5377이다. 자유도 14에서의 0.20의 확률에 해당되는 카이값은 18.1508이다. 따라서 귀무가설은 기각될 수 없다. 『고도를 기다리며』에서 이론값과 실제값의 차이에 대한 χ^2값은 7.8706으로 더욱 낮다.

같은 방법으로 모든 음소에 대해 조사하였다. 『소』의 경우, 카이제곱값에 해당하는 확률이 0.05보다 작은 음소는 자음 /ㄹ/과 모음 /ㅛ/, /ㅡ/였다. 이 세 개의 음소 외에는 모두 귀무가설을 기각할 수 없었다. 반면에 『고도를 기다리며』에서 실제 분포와 이항법칙에 따른 모델과 유의한 차이를 보이는 음소는 훨씬 많았다. 구체적으로 네 개의 자음 /ㅃ/, /ㅍ/, /ㅋ/, /ㅆ/에서 유의한 차이를 보였는데, 이 자음들은 모두 빈도수가 낮은 음소들이었다. 모음의 경우, 좀 더 많은 /ㅓ/, /ㅔ/, /ㅐ/, /ㅛ/, /ㅘ/, /ㅟ/에서 모델과 유의한 차이를 보였다. 이 모음들이 모델보다 불규칙한 분포를 보인 정확한 이유를 찾기 위해서는 문체나 테마연구와 같은 또다른 조사를 필요로 한다.[18]

이제, 이러한 조사 결과를 근거로, 한국어 음소는 언어적 제약에도 불구하고 전반적으로 이항분포를 따른다고 말할 수 있다. 특히 『소』에서는 거의 대부분의 음소가 이항분포를 따른다. 이와는 달리, 번역된 작품인 『고도를 기다리며』에서는 11개의 음소에서 유의한 차이를 보였다. 그러나 이 11개의 음소가 기각 수준에서 그렇게 멀리 떨어진 것은 아니었다.

분명히 언어적 제약은 존재한다. 그럼에도 불구하고 음소 분포가 주머니에서 붉은 공을 꺼내듯이 이항분포를 따른다는 결과는 어떻게 해석해야 할 것인가? 가장 먼저 언어적 제약이 미치는 범위의 문제를 생각할 수 있다. 구체적으로 말하자면, 언어적 제약이라는 것은 음운론적으로 바로 인접한 음소, 즉 최소 좌우 3음소 정도에 강하게 존재할 것이고, 약하게는 좀 더 멀리까지 존재할 것이다. 본 실험에서처럼 블럭의 크기가 100 정도되면, 이런 언어적 제약을 벗어나기에 충분한 크기가 될 수 있는 것이다. 사이즈를 50부터 1,000까지 다양하게 적용했을 때, 모두 유사한 결과가 나왔지만, 언어적 제약에 대한 구체적인 조사를 하기 위해서는 훨씬 더 작은 시퀀스에 대한 연구가 필요할 것이다.

---

18) 프랑스어 판 『고도를 기다리며』의 음소분포에 대한 연구(BAE Hee-Sook, 1989)와 어휘분포에 대한 연구(BAE Hee-Sook, 1990)를 통해, 베케트의 『고도를 기다리며』가 음소와 어휘 구조에 있어서 프랑스어로도 매우 독특한 분포현상을 보이고 있으며, 이는 작품의 문체와 테마에 직접적으로 연결되어 있음을 밝힌 바 있다.

결 론

발화음운론적 발음변환에 의해 얻어진 음소코퍼스에 기초해서, 음소의 빈도수를 조사하였다. 그리고 개별 음소 빈도수에 입각해서 자음과 모음의 비율, 조음방식에 따른 자음군의 빈도수와 조음 위치에 따른 모음군의 빈도수도 조사하였다. 이는 한국어 담론의 음소 구조 내에서 음소 혹은 음소군들이 담당하는 양적 역할을 알 수 있게 해준다. 또한 한국어 음소들이 언어적 제약에도 불구하고 전반적으로 이항법칙에 따라 분포된다는 사실도 밝혀냈다. 이러한 연구 결과는 한국어 음소에 대한 직관적이고 이론적인 추측을 넘어 실제 분포에 대한 명확한 통계 자료를 제시하는 것으로 한국어 음소 연구에 중요한 기초가 될 것이다.

본 논문에서는 다루지 않았으나 음소코퍼스에 기반하여 다양한 연구가 이루어질 수 있다. 예를 들면, 음소 간 상관관계를 통해 조음기관이나 조음방법의 차이가 실제로 음소의 분포에 미치는 영향을 조사할 수 있고, 한국어 음소 구조에서 2 음소 3 음소의 출현확률도 조사될 수 있다. 나아가 한국어 음소가 이항분포를 따른다는 결과는 한국어로 이루어진 담론에 대한 음성문체론적 연구의 토대가 되어, 언어학과 문학을 연결해 주는 중요한 가교 역할을 할 수 있을 것으로 기대된다.

참 고 문 헌

[1] 유치진, 『소』, http://user.chollian.net/~javanet/play/so.zip.
[2] 사뮈엘 베케트, 『고도를 기다리며』, www.kcaf.or.kr/hyper/Kdrama_main:html.
[3] 표준한국어발음대사전, KBS 편저, 김석득/이현복/유재원 감수, 어문각, 1993.
[4] 표준국어대사전, 국립국어연구원, 두산동아, 1999.
[5] 김선미, 『한국어의 리듬단위와 문법구조: 음성 합성에서 리듬 구현의 자연성 향상을 위한 음성·언어학적 연구』, 서울대학교 언어학과 박사학위 논문, 1997.
[6] 문양수 외, 『현대언어학』, 서울, 한신문화사, 1985.
[7] 성철재, 『한국어 리듬의 실험음성학적 연구: 시간 구조와 관련하여』, 서울대학교 언어학과 박사학위 논문, 1996.
[8] 양병곤, "한국어 이중모음의 음향음성학적 연구", 말소리 25권 6호, pp. 3-26, 1993.
[9] 이병근, 최명옥 공저, 『국어음운론』, 서울, 한국 방송대 출판부, 1999.
[10] 이상호, 『한국어 TTS시스템을 위한 운율의 트리 기반 모델링』, 박사학위논문, 한국과학기술원 전산학과, 1999.
[11] 이승명, 『심리 언어학』, 계명대학교출판부, 1984.
[12] 임칠성, 水野俊平, 北山一雄 공저, 『한국어 계량연구』, 전남대학교 출판부, 1997.
[13] BAE Hee-Sook, *Structures lexicales, syntaxiques et phonétiques dans deux pieces de J. Tardieu*, Thèse de Doctorat, Strasbourg, 1997.
[14] BAE Hee-Sook, *La distribution du lexique dans En Attendant Godot de S. Beckett*, Mémoire de D.E.A., Strasbourg, 1990.
[15] BAE Hee-Sook, *La distribution des phonemes dans En Attendant Godot de S. Beckett*, Mémoire de la Maîtrise, Strasbourg, 1989.
[16] Hyun Bok Lee, "A Linguistic Phonetic Approach to Speech Science" in *ICSP '99*.

[17] Marc Hug, *La distribution des phonemes en francais (Etudes statistique)*, Slatkine 1979.
[18] Marc Hug, "La Statistique linguistique en France", in *SELF*, 1980.
[19] F. Wioland, *Le rythmique du francais parle*, Strasbourg, 1987.

접수일자: 2000. 10. 31.
게재결정: 2000. 11. 23.

▲ 배희숙
대전시 유성구 어은동 한빛아파트 101동 202호
한국과학기술원 인공지능센터 음성언어연구실
Tel: +82-42-869-8720, H/P: 016-361-2965
E-mail: hsbae@bulsai.kaist.ac.kr

▲ 구동욱
대전시 유성구 구성동 979-1 한국과학기술원 전산학과
한국과학기술원 인공지능센터 음성언어연구실
Tel: +82-42-869-3556
E-mail: dokoo@bulsai.kaist.ac.kr

▲ 윤영선
대전시 유성구 구성동 979-1 한국과학기술원 전산학과
한국과학기술원 인공지능센터 음성언어연구실
Tel: +82-42-869-8720
E-mail: ysyun@bulsai.kaist.ac.kr

▲ 오영환
대전시 유성구 구성동 979-1 한국과학기술원 전산학과
한국과학기술원 전산학과 교수
Tel: +82-42-869-3516
E-mail: yhoh@bulsai.kaist.ac.kr