한국어 의미학 69, 49-104 (2020. 9. 30.) http://dx.doi.org/10.19033/sks.2020.9.69.49

한국어 단어의 어휘성에 영향을 미치는 표기음절 요인 탐구*

-무선 단어생성 실험 연구-

이은하 · 남기춘**

(고려대학교 지혜과학연구센터 연구교수 · 고려대학교 교수/고려대학교 지혜과학연구센터 소장)

<Abstract>

Lee Eun-Ha, Nam Ki-Chun, 2020. An investigation on syllabic features having an effect on the lexicality of Korean words: An random word generation study. *Korean Semantics, 69.* This study investigates whether three syllabic features, which are the classes of words, the types of words, and the positions in words from which orthographic syllables are drawn, have an effect on the probability for a pair of the sublexical items to be a legitimate Korean word when they are combined at random. As a result of three computational experiments, three main findings were obtained as follows: (1) the combination of a pair of syllables generated more real words when their positions in words were preserved than not; (2) the combination of a pair of syllables from a homogeneous word type produced more real words than from heterogeneous word types and the combination of a pair of sino-Korean syllables showed the higher productivity than of native Korean ones as well; and (3) the combination of a pair of syllables from a homogeneous word class created more real words than from heterogeneous word classes and the combination of a pair of noun syllables was more productive than of verb ones as well. These

* 본 논문은 2019년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(NRF-2019R1I1A1A01062999).

** 제1 저자: 이은하, 교신 저자: 남기춘

findings suggest that the word classes, the word types, and the positions in words have a significant impact on the probability of inter-syllabic combinations in Korean word formation, and that the syllables as sublexical units play an important role in organization of lexical representations in the mental lexicon of native speakers of Korean.

핵심어: 한국어 음절(Korean syllables), 하위어휘 단위(sublexical units), 한국어 심성어휘집(Korean mental lexicon), 어휘성(lexicality), 무선 단어생성(random word generation), 단어형성(word formation), 어종(word types), 품사(word classes), 세종 말뭉치(Sejong Corpus), 컴퓨터 기반 실험(computational experiments)

1. 서론

1.1. 연구의 목적과 필요성

모어화자는 시각을 통해 제시된 단어가 자신의 모어에 실제로 존재하는 단어인지 여부를 매우 신속하게 판단할 수 있다. 단어가 제시되면 일단 머릿속 사전 — 심성어휘집(mental lexicon) — 을 검색하여 목표단어에 대응하는 단어가 발견된다면 실제 단어로 판별될 것이고, 그렇지 않다면 비단어(nonword)로 판별될 것이다(Forster 1989). 모어화자가 이러한 능력을 발휘할 수 있는 이유는 개별 단어에 대한 철자적[1] 정보(상호활성화 모형[interactive activation model])(McClelland & Rumelhart 1981; Rumelhart & McClelland 1982) 또는 모어의 철자 조합 패턴에 대한 통계적 지식(연결주의 모형[connectionist model])(Harm & Seidenberg 2004; Plaut 1997)을 지니고 있기 때문이다. 모어화자가 시각적으로 제시된 단어의 어휘성(lexicality) — 모어화자의 심성어휘집에 실제로 존재하는 단어인지 아닌지 여부 — 을 판단할 때 어떤 암시적 철자지식을 사용하는지

1) 본고에서 “철자/철자적(orthographic)”은 시각적으로 제시되는 모든 하위어휘 단위(sublexical units: 예를 들면, 자소[graphemes] 또는 표기음절[syllables])를 포괄하여 가리킨다. 독자의 혼동을 방지하기 위해 개별 자소(예를 들면, ‘ㄱ’, ‘ㄴ’, ‘ㅏ’, ‘ㅜ’)는 “자소” 또는 “문자(letter)”로 따로 구분하여 부르기로 한다.

알 수 있다면, 심성어휘집의 체계와 시각어휘 처리 기제를 규명하는 데 큰 도움이 될 것이다.

일반적으로, 모어화자의 심성어휘집에 단어들이 어떻게 표상되어 있는지를 고찰하는 데는 두 가지 접근법이 있다. 하나는 실제 모어화자를 대상으로 말뭉치에서 추출한 어휘자극에 바탕을 둔 실험(예를 들면, 어휘판단과제[lexical decision task] 또는 단어명명과제[word naming task][2])을 통해 모어화자의 어휘표상 체계를 경험적으로 규명하는 것이다. 다른 하나는 다수의 모어화자가 생산한 대규모 텍스트 자료(예를 들면, 문어 또는 구어 말뭉치)에 나타난 언어사용 패턴을 분석함으로써 모어화자의 어휘표상 체계를 간접적으로 유추하는 것이다.

언어심리학에서는 주로 첫 번째 접근법을 통해 심성어휘집의 체계와 시각단어 재인의 원리를 고찰해왔다. 이때 가장 자주 사용되는 방법론이 바로 어휘판단과제이다. 어휘판단과제를 수행할 때 모어화자는 시각 또는 청각을 통해 제시된 문자열이 모어에 실제로 존재하는 단어인지 여부에 대해 가능한 한 빨리 네/아니오로 응답해야 한다. 이때 실험자극으로 실제 단어와 비단어가 함께 사용된다(Balota & Chumbley 1984). 많은 언어심리학 연구들이 시각적으로 제시된 단어의 어휘성을 판단할 때 어떤 암시적 철자지식을 사용하는지를 살펴보기 위해 어휘판단과제를 사용해왔다(예를 들면, 권유안 · 이윤형 2015; 김제홍 · 이창환 · 남기춘 2018; 이광오 · 배성봉 2009b; 이창환 · 김제홍 2018;Álvarez, Carreiras, & Taft 2001; Álvarez, Carreiras, & de Vega 2000; Conrad, Carreiras, Tamm, & Jacobs 2009; Perea & Carreiras 1998).

최근 들어 대규모 텍스트 자료 분석을 통해 한국어 모어화자의 어휘표상 체계를 유추하고자 하는 시도들 또한 속속 선보이고 있다(예를 들면, 말뭉치 분석: 김미란 · 최재웅 · 홍정하 2014; 남성현 · 김선회 2018; 신지영 2010; 이용은 2016; 이은하 · 남기춘 2020; 기계학습: 박나영 2014, 2015). 이들 연구는 세종말뭉치(강범모 · 김흥규 2009)나 표준국어대사전(국립국어원 2008) 같은 대규모 텍스트 자료를 바탕으로 단어 내 음소 간 결합 또는 음절이 어떤 확률적

2) 시각적으로 제시된 단어 또는 비단어를 가능한 한 빨리 소리 내어 읽도록 설계된 언어심리학 실험 기법이다(Schiller 2004).

분포를 나타내는지를 고찰함으로써 한국어 모어화자의 심성어휘집이 지닌 음운적 · 철자적 특성을 파악하는 데 초점을 둔다. 따라서 어휘판단과제를 사용하지 않고도 모어화자가 단어의 어휘성을 판단할 때 어떤 암시적 철자지식을 활용하는지 추론할 수 있다. 또한 실험 참여자를 필요로 하지 않으므로 인간 대상 실험에 대한 시의적절한 대안이 될 수 있다.

본 연구는 말뭉치 분석에 바탕을 둔 음절 무선조합 단어생성 실험을 통해 한국어의 어휘성에 영향을 미치는 표기음절 변수를 규명하는 데 목적을 둔다. 이때 대규모 말뭉치와 프로그래밍 기법을 바탕으로, 음절들을 무작위로 조합하여 한국어 단어를 만든다면 어떤 언어적 기준을 적용할 때 실제 단어를 더 많이 생성할 수 있을지에 대해 계량적으로 접근하고자 한다. 이를 통해 모어화자가 지닌 한국어 음절 분포 및 음절 조합 패턴에 대한 직관의 실체를 체계적으로 고찰하고자 한다. 이에 단어형성 시 음절 간 조합에 영향을 미칠 수 있는 주요 언어적 변수로 음절이 유래한 품사, 음절이 유래한 어종 그리고 음절의 단어 내 출현위치를 설정, 이들 변수에 따라 음절들을 무선 조합했을 때 실제 단어가 생성될 확률이 어떤 양상을 나타내는가를 살펴보고자 한다.

1.2. 선행연구 검토

1.2.1. 심성어휘집을 구성하는 하위어휘 단위로서의 음절

어휘는 언어의 가장 보편적 구성요소 중 하나이다. 개별 단어가 심성어휘집에 어떻게 표상되어 있으며 어떤 과정을 통해 처리되는가는 언어심리학의 주요 논제 중 하나이다. 가장 영향력 있는 시각단어 재인 모형 중 하나인 상호활성화 모형과 그 변종인 이중경로 모형에 의하면, 문자표상(letter representations)과 어휘표상(lexical representations) 사이에 하위어휘 표상(sublexical representations)이 존재하며, 하위어휘 표상은 둘 사이에서 어휘접속(lexical access)을 매개하는 역할을 담당한다. 따라서 모어화자가 시각적으로 제시되는 단어를 처리할 때 하위단위로 분해하는 절차가 개입되며, 이 과정에서 철자 간 또는/그리고 하위어휘 단위 간 조합에 대한 암시적 지식이 작용하게 된다.

현재까지 제안된 대표적인 하위어휘 단위로는 음절(syllables: Spoehr & Smith 1973; Taft & Forster 1976), 기본 표기음절 구조(Basic Orthographic Syllabic Structure, 이하 BOSS: Taft 1979, 1987),[3] 형태소(morphemes: Carlisle & Stone 2005; Taft & Forster 1975) 그리고 두 낱자 쌍(letter bigrams: (Andrews 1992; Seidenberg 1987)[4] 등이 있다. 이들 중에서도 가장 주목을 받고 있는 하위어휘 단위가 바로 음절이다(한국어: 권유안 2012; 권유안 · 이윤형 2014; 신하선 · 남기춘 2019; 태진이 · 남예은 · 이윤형 · 김태훈 2015; 영어: Macizo & Van Petten 2007; Schiller 1999); 독어: Conrad & Jacobs 2004; Conrad, Stenneken, & Jacobs 2006; 프랑스어: Chetail, Colin, & Content 2012; Doignon-Camus, Bonnefond, Touzalin-Chretien, & Dufour 2009; 스페인어: Álvarez et al. 2000; Perea & Carreiras 1998). 특히 음절의 철자적 경계와 음운적 경계가 일치하는 스페인어나 한국어처럼 음절 간 경계가 분명한 언어 — 표기심도가 낮은 언어(shallow languages) — 일수록 음절은 시각단어 재인에 중요한 영향을 미치는 것으로 나타났다. 이는 한국어 단어가 음절 단위로 심성어휘집에 표상되어 있을 가능성을 시사한다. 따라서 상술한 이론적 근거를 토대로 본 연구에서는 무선 단어 생성을 위한 기본단위로 표기음절을 선택하고자 한다.

1.2.2. 음절 및 음절 간 결합의 확률적 분포에 영향을 미치는 변수

그간 계량적 접근법을 토대로 음소 간 결합의 분포적 특성이나 음소 간 결합에 영향을 미치는 요인을 밝히려는 시도는 꾸준히 이어져왔다(말뭉치 분석: 김미란 외 2014; 남성현 · 김선회 2018; 이용은 2016); 기계학습: 박나영 2014, 2015). 그러나 계량적 접근법을 통해 표기음절의 분포적 특성이나 표기음절 간 결합에 영향을 미치는 요인을 규명하고자 시도한 본격적 연구는 이은하 · 남기춘(2020) 정도뿐이다. 이에 따라 본 연구에서는 계량언어학, 언어심리학, 국어

3) 단어의 첫 음절에 후속 자음 한 개가 더해진 하위어휘 단위(예를 들면, teapot의 teap)를 가리킨다.

4) 표기잉여 가설(orthographic redundancy hypothesis)에 따르면, vodka는 vo, od, dk, ka이라는 낱자 쌍으로 이루어져 있으며 이 중에서도 가장 빈도가 낮은 낱자 쌍인 dk를 음절의 경계로 인식하게 된다(Seidenberg & McClelland 1989).

학 등 다양한 관련 학문의 성과를 두루 참고하여 음절 및 음절 간 조합의 확률적 패턴에 영향을 미칠 수 있는 언어적 요인들을 검토하고자 한다.

한국어는 초성 19자, 중성 21자, 종성 27자를 조합하여 음절을 표기할 수 있다. 따라서 받침 없는 음절은 19 × 21 = 399자, 받침 있는 음절은 399 × 27 = 10,773자가 만들어지므로, 이를 모두 더하면 이론상으로 생성 가능한 표기음절은 모두 11,172자가 된다. 만약 모어화자가 사용하는 단어들이 임의의 음절이 일정한 조건 없이 무작위로 조합된 결과라면, 11,172개의 음절이 모두 동등한 확률로 단어형성에 참여해야 한다. 그러나 현실적으로는 2,110개(이은하 · 남기춘 2020)[5]의 음절만이 단어에 사용되고 있다. 이는 단어형성 시 모든 개별 음절이 균등한 가능성을 가지고 조합되지 않을 가능성을 강력하게 시사한다. 뿐만 아니라 1,500만 어절 규모 세종 형태의미 분석 말뭉치에 나타난 한국어 음절의 빈도와 분포를 분석한 남기춘 · 이은하(2020)에 따르면, 품사나 단어 내 출현위치에 상관없이 적은 가짓수의 음절이 출현형(token) 및 유형(type) 빈도[6]를 독점하는 경향(Zipf 1935) — 멱함수 곡선(power function curve)(Newman 2005)으로 시각화되는 — 을 나타냈다. 이렇듯 음절의 출현형/유형 빈도가 심각한 편포를 이루고 있다는 사실은 실제 사용되는 2,110개의 음절조차 무작위가 아닌 일정한 조건에 따라 선택 · 조합되는 것임을 암시한다.

본 연구에서는 단어형성에 참여하는 음절과 이들 음절 간 결합 분포에 영향을 미칠 수 있는 요인으로 음절의 단어 내 출현위치, 음절이 유래한 단어의 어종 그리고 음절이 유래한 단어의 품사를 제안하고자 한다. 그 이유는 다음과 같다. 첫째, 다수의 언어심리학 연구에 따르면 단어의 첫 음절은 시각단어 재인에서 매우 중요한 역할을 담당한다. 첫 음절을 공유하는 단어 — 음절 이웃[7] — 의 개수(음절 이웃 크기[syllable neighbor size])와 관련된 음절효과로는 음절 이웃 크기 효과(syllable neighborhood size effects)가 있다. 이는 첫 음절을 공유

5) 이는 1,500만 어절 규모 세종 형태의미 분석 말뭉치(강범모 · 김흥규 2009)에 1회 이상 출현한 단어에 사용된 음절의 가짓수를 가리킨다. 50만 표제어 규모 표준국어대사전에 등재된 표제어에 사용된 음절의 가짓수는 이보다 조금 더 많은 2,475개이다(김미란 외 2014).

6) 출현형 빈도(token frequency)는 해당 언어단위가 출현한 횟수를, 그리고 유형 빈도(type frequency)는 해당 언어단위의 가짓수를 가리킨다.

7) 예를 들면, '사과'와 첫 음절을 공유하는 음절 이웃으로는 '사람', '사유', '사진기' 등이 있다.

하는 음절 이웃이 많은 단어를 인식할 때 그렇지 않은 단어를 인식할 때보다 반응시간과 오류율이 감소하는 현상 — 촉진효과(facilitation effects) — 을 가리킨다(권유안 2012; Carreiras & Perea 2002; Conrad, Carreiras, & Jacobs 2006; Perea & Carreiras 1998).

아울러 첫 음절 공유 단어 누적 출현빈도와 관련된 음절효과로는 음절 빈도효과(syllable frequency effects)가 있다. 이는 어휘판단과제를 수행할 때 첫 음절의 출현빈도가 높은 단어가 그렇지 않은 단어보다 반응시간과 오류율이 증가하는 현상 — 억제효과(inhibition effects) — 에 해당한다(권유안 2012; 신하선 · 남기춘 2019; Álvarez et al. 2000; Conrad et al. 2006). 특히 권유안(2012)와 Álvarez 외(2000)에서는 첫 음절을 공유하는 실제 단어 이웃을 많이 거느린 비단어일수록 NO 반응이 지연되는 — 실제 단어에 가깝게 보이는 — 경향을 나타냈다. 이는 음절의 위치별 빈도분포가 어휘성 결정에 중요한 역할을 담당하며, 모어화자가 음절의 위치별 빈도분포에 대한 암시적 지식을 지니고 있을 가능성을 강하게 시사한다.

1500만 어절 규모 세종 형태의미 분석 말뭉치에 나타난 위치별 음절 분포를 살펴본 이은하 · 남기춘(2020)에서도 이러한 가능성과 관련된 흥미로운 현상이 관찰된다. 그에 따르면 체언[8] 표제어 유형의 경우, 상위빈도 100개 음절 가운데 첫째 또는 둘째 자리에만 나타나는 음절이 28개에 달했다. 이를테면, 체언 표제어 첫째 자리에 자주 쓰이는 음절 중 '오', '초', '한', '저', '예' 등은 상대적으로 둘째 자리에서 잘 쓰이지 않았다. 그리고 빈도에 관계없이 체언 표제어의 첫째 또는 둘째 자리에서만 발견되는 음절도 279종이나 되었다.

뿐만 아니라 용언[9] 표제어 유형 역시 상위빈도 100개 음절 중 첫째 또는 둘째 자리에만 쓰이는 음절이 35개나 되었다. 예를 들면, 용언 표제어 첫째 자리에 주로 나타나는 음절 중 '불', '되', '뒤', '노', '후' 같은 음절은 상대적으로 둘째 자리에 자주 나타나지 않았다. 아울러 빈도와 무관하게 용언 표제어의 첫째 또는 둘째 자리에만 발견되는 음절 또한 289종이 확인되었다. 이는 같은

8) 본래 체언은 일반명사, 대명사, 의존명사 및 수사를 모두 아우르는 개념이나, 본고에서는 편의상 일반명사를 지칭한다.

9) 본고에서 용언은 동사와 형용사를 모두 아우른다.

음절일지라도 표제어의 첫째 자리와 둘째 자리에 항상 동등한 확률로 고르게 선택되는 것은 아님을 나타낸다. 이로 미루어볼 때 한국어 모어화자는 상술한 표제어 내 위치별 음절의 언어적 특성과 분포정보에 관한 암시적 지식을 가지고 있으며, 주어진 음절연쇄의 어휘성을 판단할 때 이를 활용할 가능성이 높다.

둘째, 어종 — 고유어 또는 한자어 — 에 따라 단어를 구성하는 음절의 언어적 특성과 분포양상이 뚜렷한 차이를 나타낸다(권인한 1997; 김유범 2016; 신지영 2009; 안소진 2009). 일단 한자어의 자소체계는 고유어보다 훨씬 간소해서 ‘ㄱ’, ‘ㄲ’, ‘ㄴ’, ‘ㄷ’, ‘ㄹ’[10] ‘ㅁ’, ‘ㅂ’, ‘ㅅ’, ‘ㅆ’, ‘ㅇ’, ‘ㅈ’, ‘ㅊ’, ‘ㅋ’, ‘ㅌ’, ‘ㅍ’, ‘ㅎ’까지 총 15개만 초성으로 쓰일 수 있다. ‘ㄸ’, ‘ㅃ’, ‘ㅉ’는 초성으로 쓰일 수 없으며, ‘ㄲ’, ‘ㅆ’, ‘ㅋ’은 매우 제한적으로 나타난다(권인한 1997; 남성현 · 김선회 2018; 박나영 2015; 신지영 2009; 안소진 2009). 중성의 사용양상 역시 고유어와 일정한 차이를 보인다. 한자어는 고유어보다 이중모음이 훨씬 다양하게 사용되며, 고유어에 거의 사용되지 않는 ‘ㅕ’, ‘ㅘ’도 제법 자주 나타난다. 그러나 ‘ㅐ’가 포함된 한자어는 찾아볼 수 없으며, ‘ㅔ’의 사용빈도도 매우 낮다. 뿐만 아니라 ‘ㅢ’는 ‘의’와 ‘희’에만 존재하며, ‘ㄷ’, ‘ㅌ’, ‘ㄹ’, ㅁ, ‘ㅂ’, ‘ㅍ’ 뒤에 ‘ㅘ’, ‘ㅙ’, ‘ㅝ’, ‘ㅞ’ 등의 이중모음이 결합된 음절은 존재하지 않는다. 종성의 범위 역시 고유어보다 제한적이어서, ‘ㄷ’, ‘ㅅ’, ‘ㅈ’, ‘ㅊ’, ‘ㅌ’, ‘ㅋ’, ‘ㅌ’, ‘ㅍ’과 고유어에 흔한 겹받침 역시 사용되지 않는다(권인한 1997; 신지영 2009; 안소진 2009).

게다가 한자 음절[11]은 한글 음절[12]보다 생산성이 압도적으로 높다. 어종이 확인된 표준국어대사전 표제어(n = 363,082)만 놓고 보면, 한자어의 비율이 53.04퍼센트에 달하는 반면 고유어의 비율은 20.89퍼센트에 지나지 않는다. 고유어/외래어와 한자어의 혼종어 비율 역시 20.46퍼센트로, 한자어가 포함된 표제어를 모두 합치면 그 비율은 73.50퍼센트나 된다(국립국어원 2020). 이는 한자 음절의 높은 생산성을 시사하는 단적인 증거이다.

---

10) 한국어에서 ‘ㄹ’은 한자어의 첫 음절 초성으로 쓰일 수 없다.

11) 본 연구에서 한자 음절이란 한자어(예를 들면 ‘현재’의 ‘현’과 ‘재’)와 혼종어(예를 들면 ‘공부하-’의 ‘공’과 ‘부’)에 사용된 모든 한자 유래 음절을 통틀어 가리킨다.

12) 본 연구에서 한글 음절이란 고유어(예를 들면 ‘사람’의 ‘사’와 ‘람’)와 혼종어(예를 들면 ‘욕심쟁이’의 ‘쟁’과 ‘이’)에 사용된 모든 한글 유래 음절을 통틀어 가리킨다.

뿐만 아니라 모어화자의 심성어휘집에 고유어와 한자어가 분리되어 표상되어 있을 가능성을 시사하는 언어심리학 연구들도 적지 않다(권유안 · 남기춘 2011; 이광오 2003; 이광오 · 배성봉 2009a; 이광오 · 정진갑 · 배성봉 2007; Kim & Na 2000). 한국어 모어 실서증(失書症[dysgraphia]) 환자에 대한 사례연구인 Kim and Na(2000)에 따르면, 고유어에 대한 받아쓰기 정확도가 한자어에 비해 현저히 낮았다. 고유어, 한자어, 외래어 세 어종별 어휘판단 시간을 비교한 이광오(2003)에서는 고유어에 대한 반응시간이 가장 짧았으며 한자어와 외래어가 그 뒤를 이었다. 또한 한자어에 대한 어휘판단 시 한자어 음절 이웃 — 동일한 한자어 형태소를 첫 음절로 공유하는 단어들 — 을 많이 거느린 단어일수록 반응시간이 짧았다(권유안 · 남기춘 2011). 상술한 결과들은 어종의 심리적 실재성을 시사하는 강력한 증거이다. 나아가 어종에 따라 단어를 구성하는 음절들이 언어적 · 분포적으로 분명한 차이를 나타내는 것으로 미루어볼 때, 모어화자가 지닌 어종에 대한 인식은 어종별 음절특성에 대한 암시적 지식과도 밀접한 관련이 있을 것으로 추측된다.

마지막으로, 품사에 따라서도 단어를 구성하는 음절이 분포적 · 언어적으로 차이를 나타낸다. 한국어 품사는 크게 체언, 용언, 수식언, 독립언 그리고 관계언으로 나뉜다. 이 가운데 가장 자주 사용되는 품사가 체언과 용언으로, 둘은 기능, 형태, 의미 면에서 확연히 구분된다. 특히 용언은 체언에 비해 형태의 변화(가령, 불규칙 활용, 탈락 또는 삽입 등)가 훨씬 다양하다. 또한 이은하 · 남기춘(2020)에 따르면 용언 음절과 체언 음절의 음운론적 구성에 일정한 차이가 있었다. 용언 표제어는 체언 표제어에 비해 자음 + 모음 음절의 비율(용언 65.60퍼센트, 체언 40.20퍼센트)이 더 큰 반면, 자음 + 모음 + 자음 음절의 비율(용언 27퍼센트, 체언 45.70퍼센트)은 훨씬 작았다. 그리고 용언 표제어 출현형 빈도 상위 100개 음절과 체언 표제어 출현형 빈도 상위 100개 음절 간 중복음절 개수는 48개에 불과하다(이은하 2020). 이러한 사실은 단어를 구성하는 음절 및 음절조합 패턴이 품사별로 일정한 차이를 지닐 가능성을 시사한다.

게다가 체언과 용언의 어종구성에도 큰 차이가 있다. 체언의 경우 한자어가 전체의 69.12퍼센트를 차지하는 반면, 고유어는 13.44퍼센트밖에 되지 않는다. 하지만 용언의 경우 한자어는 존재하지 않으며, 고유어가 35.35퍼센트에 혼종

어가 나머지 64.65퍼센트를 점유한다(국립국어원 2020). 즉 용언의 고유어 비율은 체언의 고유어 비율보다 무려 2.63배나 높은 반면, 용언의 한자어 비율은 사실상 0이다. 이은하 · 남기춘(2020)에서도 체언 표제어는 생산성이 높은 한자어 어근 음절이 높은 유형 빈도를 나타냈지만, 용언 표제어는 고유어 어간 또는 접사 음절의 유형 빈도가 높게 나타났다. 이로 미루어볼 때 품사와 어종이 단어를 구성하는 음절의 분포적 · 언어적 특성에 미치는 영향을 검증하고자 한다면, 반드시 두 변수를 통계적으로 분리하는 작업이 필요하다.

상술한 바와 같이, 기존의 언어심리학 연구들이 한국어 음절이 모어화자의 어휘접속과 어떤 관계를 맺고 있는지에 대해 의미 있는 통찰을 제공한 것은 사실이다. 그럼에도 불구하고 음절정보가 심성어휘집에 어떻게 표상되어 있는지에 대해서는 알려진 것이 많지 않다. 물론 계량언어학적 방법론을 바탕으로 음소 간 결합 또는 음절의 분포양상을 조사한 연구들을 통해서도 한국어 단어 형성에서 음절이 수행하는 역할에 대한 유용한 정보를 얻을 수 있다. 하지만 이들 연구는 특정한 음절연쇄를 실제 한국어 단어로 만들어주는 음절의 통계적 · 언어적 속성이 무엇인가란 질문에 만족스러운 해답을 제공하지 않는다.

이에 본 연구에서는 특정한 음절연쇄가 한국어 모어화자의 심성어휘집에 실제로 존재할 확률을 결정하는 표기음절 변수에 대해 계량언어학적 방법론을 바탕으로 고찰하고자 한다. 언어심리학, 계량언어학, 국어학 등 다양한 관련 학문의 성과를 토대로, 단어형성 시 음절 간 조합에 영향을 미칠 수 있는 주요 변수로 음절의 단어 내 출현위치, 음절이 유래한 단어의 어종 그리고 음절이 유래한 단어의 품사를 설정하고자 한다. 그리하여 대규모 한국어 말뭉치 — 세종 형태의미 분석 말뭉치와 표준국어대사전 표제어 목록 — 에서 추출한 고빈도 음절을 이들 변수에 따라 두 개씩 무선 조합했을 때 실제 단어가 생성될 확률이 어떤 양상을 나타내는가를 통계적으로 검증하고자 한다. 본 연구의 결과는 한국어 단어 형성과 밀접한 관련을 맺고 있는 음절변수는 무엇이고, 한국어 모어화자의 심성어휘집에 음절정보가 어떻게 표상되어 있는지를 규명하는 데 기여할 것으로 기대된다. 본 연구의 연구문제와 연구가설은 다음과 같다.

연구문제 1: 고빈도 음절 목록을 무작위로 조합하여 2음절 연쇄를 생성할 경우,

음절이 유래한 단어 내 출현위치가 실제 한국어 단어가 만들어질 확률에 어떤 영향을 미치는가?

연구가설 1: 음절이 유래한 단어 내 출현위치를 고려하여 음절쌍을 무선 조합했을 때[13], 음절이 유래한 단어 내 출현위치를 고려하지 않고 음절쌍을 무선 조합했을 경우[14]보다 실제 한국어 단어를 더 많이 만들어낼 것이다.

연구문제 2: 고빈도 음절 목록을 무작위로 조합하여 2음절 연쇄를 생성할 경우, 음절이 유래한 단어의 어종이 실제 한국어 단어가 만들어질 확률에 어떤 영향을 미치는가?

연구가설 2: 동일한 어종으로부터 유래한 음절쌍[15]을 무선 조합했을 때, 서로 다른 어종으로부터 유래한 음절쌍[16]을 무선 조합했을 경우보다 실제 한국어 단어를 더 많이 만들어낼 것이다. 그리고 한자 음절끼리 무선 조합했을 때, 한글 음절끼리 무선 조합했을 경우보다 실제 한국어 단어를 더 많이 만들어낼 것이다.

연구문제 3: 고빈도 음절 목록을 무작위로 조합하여 2음절 연쇄를 생성할 경우, 음절이 유래한 단어의 품사가 실제 한국어 단어가 만들어질 확률에 어떤 영향을 미치는가?

연구가설 3: 동일한 품사로부터 유래한 음절쌍[17]을 무선 조합했을 때, 서로 다른 품사로부터 유래한 음절쌍[18]을 무선 조합했을 경우보다 실제 한국어 단어를 더 많이 만들어낼 것이다. 그리고 체언 음절끼리 무선 조합했을 때, 용언 음절끼리 무선 조합했을 경우보다 실제 한국어 단어를 더 많이 만들어낼 것이다.

1.3. 연구개요

본 연구에서는 고빈도 음절을 두 개씩 무작위로 조합했을 때 실제 단어가

13) 단어의 첫째 자리에 자주 쓰이는 음절 + 둘째 자리에 자주 쓰이는 음절 간 조합을 가리킨다.
14) 단어 내 출현위치에 관계없이 자주 쓰이는 두 음절 간 조합을 가리킨다.
15) 한글 음절 + 한글 음절 또는 한자 + 한자 음절 간 조합을 가리킨다.
16) 한글 음절 + 한자 음절 또는 한자 + 한글 음절 간 조합을 가리킨다.
17) 체언 음절 + 체언 음절 또는 용언 + 용언 음절 간 조합을 가리킨다.
18) 체언 음절 + 용언 음절 또는 용언 + 체언 음절 간 조합을 가리킨다.

생성될 확률이 음절의 단어 내 출현 위치, 음절이 유래한 단어의 어종 그리고 음절이 유래한 단어의 품사에 따라 어떤 양상을 나타내는지를 알아보았다. 앞서 지적했듯이 음절이 유래한 단어의 품사와 어종은 서로 밀접하게 연관된 까닭에, 음절의 위치 · 어종 · 품사 변수가 실제 단어 생성률에 미치는 영향을 각기 분리하여 검증하기 위해 총 세 번의 실험을 실시했다.

실험 1은 음절의 위치 변수[19]와 품사 변수[20]의 효과를 확인하는 데 목적을 두었다. 실험 2는 체언이 용언에 비해 상대적으로 한자어 비율이 높고 고유어 비율이 낮은 것을 고려하여, 품사 변수 효과의 혼입을 방지하기 위해 단일 품사(체언 또는 용언)에서 추출한 음절을 대상으로 어종 변수[21]의 효과를 확인하는 데 목적을 두었다. 마지막으로 실험 3은 실험 2와 같은 이유로 어종 변수 효과의 혼입을 방지하기 위해 단일 어종(고유어 또는 한자어)에서 추출한 음절을 대상으로 품사 변수[22]의 효과를 확인하는 데 목적을 두었다.

각 실험의 절차는 기본적으로 다음의 4단계로 구성되었다. 1단계는 단어생성에 사용될 음절 추출 단계이다. 이때 세종 형태의미 분석 말뭉치 또는 표준국어대사전으로부터 음절을 추출하여 출현형 빈도를 산출하고, 품사별(체언과 용언), 어종별(한글과 한자), 위치별(첫째 음절과 둘째 음절) 음절목록을 구축했다.[23]

2단계는 무선 단어생성 단계이다. 이때 구축된 음절목록 내 상위빈도 음절 100개로 구성된 음절 세트 두 개를 무작위로 조합함으로써 100개의 2음절 연쇄를 생성했다. 100개짜리 2음절 연쇄 100세트를 500회에 걸쳐 반복 생성, 총 500만 개(= 100개× 100세트 × 500번)의 2음절 연쇄를 만들어냈다.

3단계는 단어 생성률 산출 단계이다. 이때 무선 생성된 100개의 2음절 연쇄를 표준국어대사전 기반 표제어 목록과 대조한 뒤 실제 단어와 일치하는 음절 연쇄의 비율을 측정했다. 2단계에서 2음절 연쇄 100세트를 500회에 걸쳐 반복

19) 위치별 고빈도 음절을 차례대로 배치한 경우와 위치에 대한 고려 없이 추출된 고빈도 음절을 무작위로 배치한 경우의 실제 단어 생성률을 비교한다.
20) 단일 품사에서 추출한 음절 무선조합과 이종 품사에서 추출한 음절 무선조합의 실제 단어 생성률을 비교한다.
21) 동일 어종 간 음절 무선조합과 이종 어종 간 음절 무선조합의 실제 단어 생성률을 비교한다.
22) 단일 품사 간 음절 무선조합과 이종 품사 간 음절 무선조합의 실제 단어 생성률을 비교한다.
23) 각 실험에 사용된 음절목록에 대한 구체적인 정보는 이은하(2020)을 참조하라.

생성했으므로, 총 5만 개(= 100세트 × 500번)의 실제 단어 생성률 측정치가 산출되었다.

마지막 4단계는 통계분석 단계이다. 이때 음절의 위치 · 어종 · 품사 변수가 무선조합 음절연쇄 표본의 실제 단어 생성률에 미치는 영향을 검증하기 위해 기술통계 분석, 다요인분산분석(factorial analysis of variance) 및 후속검정을 500번 실시하고, 통계량과 효과 크기를 계산한 뒤 평균을 산출했다.

2. 실험 1

시각단어 재인 시 단어의 첫 음절은 어휘 후보군을 활성화하고 둘째 음절은 활성화된 어휘 후보군 중에서 해당 단어를 선택하는 역할을 담당한다(권유안 2012; Álvarez et al. 2000). 뿐만 아니라 단어 내 위치별 음절의 형태론적 · 음운론적 특성이 각기 다른 패턴을 나타내는 것으로 미루어볼 때(이은하 · 남기춘 2020), 같은 음절일지라도 단어형성 시 표제어의 첫째 자리와 둘째 자리에 항상 동등한 확률로 선택되지 않을 가능성이 높다. 따라서 실험 1은 두 개의 고빈도 음절을 무작위로 조합했을 때 음절이 유래한 품사와 음절의 단어 내 위치가 실제 단어 생성률에 미치는 영향을 통계적으로 검증하는 데 목적을 둔다. 이때 단어 내 출현위치를 고려하여 음절쌍을 무선 조합했을 때, 단어 내 출현위치를 고려하지 않고 음절쌍을 무선 조합했을 경우보다 실제 한국어 단어를 더 많이 만들어낼 것이라는 가정을 확인하고자 한다.

2.1. 연구방법

2.1.1. 연구자료

실험 1에서는 품사별 · 위치별 한국어 음절 정보 추출을 위해 1,500만 어절 규모의 세종 형태의미 분석 말뭉치를 사용했다. 이때 무선 조합된 2음절 연쇄가 실제 단어인지 여부를 확인하기 위해 50만 표제어 규모의 표준국어대사전 수록 2음절 단어들과 대조하므로, 어간 출현형 빈도 상위 100개 음절을 실험재

료로 채택했다. 어간 출현형 빈도 추출 시 문법적 의미의 어간 원형이 활용/곡용 시에도 그대로 유지되는 어절만을 사용했으며, 어간과 조사/어미 간 경계가 불분명한 불규칙 활용이나 준말 어절은 제외했다.24)

실험 1에 투입될 음절의 품사로는 세종 말뭉치에서 가장 높은 비중을 차지하고 있으며, 한국어 어휘처리 연구에서도 가장 널리 사용되고 있는 품사인 체언(일반명사)과 용언(동사와 형용사)을 채택했다. 음절의 출현위치와 단어형성 간의 관련성을 검증하기 위해 품사별 음절의 빈도는 단어 내 출현위치에 따라 위치고려 음절 빈도와 위치 비고려 음절 빈도 두 층위로 구분했다. 위치고려 음절 빈도는 어간 출현형 내에서 첫째 자리에 쓰인 음절의 빈도와 둘째 자리에 쓰인 음절의 빈도로 나뉜다. 따라서 동일한 음절이라도 첫째 자리에 쓰인 횟수와 둘째 자리에 쓰인 횟수가 서로 다를 수 있다. 반면에 위치 비고려 음절 빈도를 계산할 때는 어간 내 위치에 상관없이 일괄적으로 출현빈도를 집계했다. 아래의 <표 1>은 실험 1에 사용된 음절의 조건별 빈도 정보를 정리한 것이다.25)

<표 1> 실험 1: 조건별 음절빈도 정보

품사	위치고려 여부	위치	출현형/유형	음절 빈도	
				전체	상위 100개
일반명사	비고려	해당 없음	출현형	12,206,805	7,497,968
			유형	1,608	100
	고려	첫째	출현형	5,442,010	3,576,963
			유형	1,395	100
		둘째	출현형	4,739,055	3,000,212
			유형	1,322	100

24) 가령 '저어서'라는 용언 어절에서 문법적 의미의 어간은 '젓-'이지만, 실제로 '젓'이라는 음절은 나타나지 않는다. 마찬가지로 '시계'에 목적격 조사 '-ㄹ'이 붙은 일반명사 어절 '시곌' 또한 문법적 의미의 어간은 '시계'이지만, '계'는 실제로 나타나지 않는다. 본 연구에서는 음절의 출현형을 집계할 때 추상적 형태소가 아닌 실제 말뭉치에 출현한 음절의 빈도를 추출하는 것을 원칙으로 했다. 따라서 문법적 의미의 어간 원형이 활용/곡용을 할 때도 그대로 보존되는 사례(예를 들면, '젓다가' 또는 '시계를')만을 바탕으로 어간 음절(예를 들면, '젓', '시', '계')의 출현형을 집계했다.

25) 저빈도 음절 간 무선조합으로는 실제 단어를 생성할 가능성이 매우 낮은 까닭에, 실험 1, 2, 3 공히 조건별 출현형 빈도 상위 100개 음절만을 사용했다.

용언	비고려	해당 없음	출현형	4,650,857	3,425,277
			유형	1,432	100
	고려	첫째	출현형	2,665,510	2,007,813
			유형	1,188	100
		둘째	출현형	1,215,941	895,997
			유형	1,158	100

말뭉치 자료와 더불어 실험 1의 주요 재료 중 하나가 바로 무선조합 음절연쇄가 실제 단어인지 여부를 확인하는 데 쓰인 사전이다. 실험 1에서는 국립국어원에서 편찬된 50만 표제어 규모의 표준국어대사전을 대조용 사전으로 사용했다. 두 개의 음절을 무선 조합하므로 대조용 표제어 목록 수록 단어 또한 어간이 2음절인 표제어(n = 147,669)로 한정했다.[26] 이렇게 하여 무선조합 음절연쇄의 실제 단어 여부 확인을 위해 총 네 가지 대조용 표제어 목록이 구성되었다.

사실, 어떤 표제어 목록과 대조하느냐에 따라 실제 단어 생성률[27]은 양적·질적으로 큰 차이를 나타낼 수 있다. 다음 면의 <표 2>에서 보듯 2음절 체언 어간의 수는 2음절 용언 어간의 18.68배에 달하는 탓에, 표제어 목록 대조 시 체언 음절 연쇄는 용언 음절 연쇄보다 실제 단어로 조회될 확률이 압도적으로 높다. 이에 품사 간 어간 유형 빈도 불균형을 보정하기 위해서, 사전에 수록된 전체 표제어(표제어 모집단)뿐만 아니라 사전에서 무선 추출된 일부 표제어(표제어 표본)와도 대조했다. 그리고 체언 또는 용언 음절끼리 조합했을 때 과연 해당 음절들이 유래한 품사의 단어를 얼마나 만들어내는지를 확인하기 위해서 전체 체언 또는 용언 표제어(품사별 표제어 모집단)뿐만 아니라 품사별 표제어 모집단으로부터 무선 추출된 일부 표제어(품사별 표제어 표본)와도 대조했다.

26) 용언의 경우, 표제어에서 어말어미 '-다'를 제외한 나머지 부분을 어간으로 간주했다. 예를 들면, 형용사 표제어 '예쁘다'는 '-다'를 제외한 '예쁘-'가 어간이 된다. 이하 본고에서 지칭하는 표제어란 모두 2음절 표제어를 가리킴을 일러둔다.

27) 1회의 무선조합을 통해 생성된 음절연쇄 목록(n = 100) 중 대조용 사전에 수록된 단어와 일치하는 음절연쇄가 차지하는 비율을 백분율로 나타낸 값을 가리킨다. 가령, 1회의 무선조합을 통해 생성된 100개의 음절연쇄 중 대조용 사전 수록 단어와 일치하는 음절연쇄의 수가 46개였다면, 해당 무선조합 사례에서 실제 단어 생성률은 46 × 100 / 100 = 46퍼센트이다.

실험 1에 사용된 네 가지 대조용 표제어 목록을 소개하면 다음과 같다.

(1) 전체 표제어 모집단과 대조: 사전에 수록된 전체 표제어 목록과 대조하여 해당 음절연쇄가 이들 표제어와 일치하는지 여부를 확인한다.
(2) 전체 표제어 표본과 대조: 사전에 수록된 전체 표제어 중에서 무선 추출된 1,000개 표본과 대조하여 해당 음절연쇄가 이들 표제어와 일치하는지 여부를 확인한다.
(3) 품사별 표제어 모집단과 대조: 체언 + 체언 음절연쇄는 사전에 수록된 전체 체언 표제어 목록과, 용언 + 용언 음절연쇄는 전체 용언 표제어 목록과, 체언 + 용언 또는 용언 + 체언 음절연쇄는 전체 체언 + 용언 표제어 목록과 대조하여 해당 음절연쇄가 이들 표제어와 일치하는지 여부를 확인한다.
(4) 품사별 표제어 표본과 대조: 체언 + 체언 음절연쇄는 사전에서 무선 추출된 체언 표제어 1,000개 표본과, 용언 + 용언 음절연쇄는 사전에서 무선 추출된 용언 표제어 1,000개 표본과, 체언 + 용언 또는 용언 + 체언 음절연쇄는 사전에서 무선 추출된 체언 500개 + 용언 표제어 500개 표본과 대조하여 해당 음절연쇄가 이들 표제어와 일치하는지 여부를 확인한다.

아래의 <표 2>는 실제 단어 생성률 측정을 위해 참조한 네 가지 대조용 표제어 목록에 수록된 표제어 개수를 정리한 것이다.

<표 2> 실험 1, 2 3: 실제 단어 여부 검증에 사용된 대조용 표제어 목록 기술통계

표제어 모집단/표본	품사	2음절 표제어 수
모집단	전체	147,669
	체언 + 용언	135,176
	체언	128,308
	용언	6,868
표본	전체	1,000
	체언 + 용언	1,000[28]
	체언	1,000
	용언	1,000

28) 체언 500개와 용언 500개를 합한 수치이다.

2.1.2. 자료처리 도구

본 연구자는 음절 추출, 음절 간 무선조합, 통계분석 그리고 분석결과의 시각화를 위해 공용 프로그래밍 언어인 R(R Core Team 2018)을 사용했다. 말뭉치에서 품사별 · 위치별로 음절을 추출하고 추출된 음절들을 무선 조합하는 작업에는 R에 내장된 텍스트 마이닝 기능과 tidyverse(Wickham 2019) 패키지가 쓰였다. 아울러 기술통계 분석 절차를 위해 pastecs(Grosjean & Ibanez 2014), 추론통계 분석 절차를 위해 R 내장 통계분석 기능과 sjstats(Ludecke 2020) 및 broom(Robinson, 2020)이 사용되었다.

2.1.3. 통계분석 방법

실험 1에서는 음절이 유래한 품사와 음절의 어간 내 위치가 음절 간 무선조합 시 실제 단어 생성률에 미치는 영향을 통계적으로 검증하기 위해 다요인분산분석을 사용했으며, 유의수준 α는 .05로 설정했다. 아울러 후속검정 방법으로는 Bonferroni 기법을 적용했다. 아래의 <표 3>은 실험 1 결과의 통계분석에 투입된 독립변수와 종속변수를 정리한 것이다. 실험 1의 독립변수는 품사혼합 여부[29], 품사조합 순서[30], 위치고려 여부[31] 세 가지이며, 종속변수는 평균 실제 단어 생성률이다.

<표 3> 실험 1: 독립변수와 종속변수

변수	내용	음절조합 조건
독립변수	품사혼합	• 단일 품사 간 음절 무선조합

29) 음절 간 무선조합 시 체언과 용언 음절을 혼합했는지, 아니면 단일 품사 음절끼리 조합했는지 여부를 가리킨다.

30) 음절 간 무선조합 시 체언 + 용언, 용언 + 체언, 체언 + 체언, 용언 + 용언 중 어느 순서로 음절을 조합했는지를 가리킨다.

31) 어간 내 첫째 자리 출현 빈도 상위 100개 음절과 둘째 자리 출현 빈도 상위 100개 음절끼리 순서를 지켜 무선 조합했는지(어간 내 음절 위치 고려), 아니면 어간 내 위치에 관계없이 어간 내 출현 빈도 상위 100개 음절끼리 무선 조합했는지(어간 내 음절 위치 비고려) 여부를 가리킨다.

	여부	• 이종 품사 간 음절 무선조합
	품사조합 순서	• 체언 + 체언 음절 순서로 무선조합(단일 품사 조합) • 체언 + 용언 음절 순서로 무선조합(이종 품사 조합) • 용언 + 체언 음절 순서로 무선조합(이종 품사 조합) • 용언 + 용언 음절 순서로 무선조합(단일 품사 조합)
	위치고려 여부	• 위치 고려: 어간 내 첫째 자리 출현 빈도 상위 100개 음절 + 어간 내 둘째 자리 출현 빈도 상위 100개 음절 간 무선조합 • 위치 비고려: 어간 내 위치무관 출현 빈도 상위 100개 음절 + 어간 내 위치무관 출현 빈도 상위 100개 음절 간 무선조합
종속변수	실제 단어 생성률	무선조합 음절연쇄 중에서 대조용 사전에 수록된 실제 단어와 일치하는 사례의 비율

2.1.4. 연구절차

실험 1의 절차는 크게 네 단계로 나뉜다. 첫 번째는 말뭉치로부터 음절을 추출하여 출현형 빈도를 산출하는 단계이다. 먼저 프로그래밍 언어 R을 기반으로 세종 형태의미 분석 말뭉치를 분석하여 체언과 용언 어간에 등장하는 모든 한국어 음절의 출현형 빈도(위치 비고려 빈도)를 산출했다. 이어서 체언과 용언 어간의 첫째 음절과 둘째 음절만 선별하여 각각의 출현형 빈도(위치고려 빈도)를 집계했다. 이로써 품사별 · 위치별로 총 여섯 개의 음절목록을 추출했다.

두 번째, 목록 내 상위빈도 음절 100개로 구성된 실험용 음절 세트 두 개를 무선 조합하여 두 글자로 이루어진 음절연쇄 100개를 생성한다. 본 단계에서 중요한 것이 두 가지가 있다. 하나는 무선 음절조합에 사용되는 음절이고, 다른 하나는 이들 음절을 조합하는 방식이다. 아래의 <표 4>는 실험 1의 무선 음절조합 절차에 사용된 음절목록을 조건별로 정리한 것이다. 위치고려 여부(고려 vs. 비고려)에 따라 품사조합 순서(체언 + 용언 vs. 용언 + 체언 vs. 체언 + 체언 vs. 용언 + 용언)를 달리하면 총 8개의 음절조합 조건이 만들어진다. <표 4>에서 보듯, 무선조합 시 첫째 자리와 둘째 자리에 투입되는 음절의 특성은 조건에 따라 달라진다.

<표 4> 실험 1: 무선 음절조합에 투입되는 조건별 음절 목록

위치고려 여부	품사조합 순서	음절조합	
		위치	음절목록
위치 비고려	체언 + 체언	음절 1	체언 어간 출현 빈도 상위 100개 음절
		음절 2	체언 어간 출현 빈도 상위 100개 음절
	체언 + 용언	음절 1	체언 어간 출현 빈도 상위 100개 음절
		음절 2	용언 어간 출현 빈도 상위 100개 음절
	용언 + 체언	음절 1	용언 어간 출현 빈도 상위 100개 음절
		음절 2	체언 어간 출현 빈도 상위 100개 음절
	용언 + 용언	음절 1	용언 어간 출현 빈도 상위 100개 음절
		음절 2	용언 어간 출현 빈도 상위 100개 음절
위치 고려	체언 + 체언	음절 1	체언 어간 첫째 자리 출현 빈도 상위 100개 음절
		음절 2	체언 어간 둘째 자리 출현 빈도 상위 100개 음절
	체언 + 용언	음절 1	체언 어간 첫째 자리 출현 빈도 상위 100개 음절
		음절 2	용언 어간 둘째 자리 출현 빈도 상위 100개 음절
	용언 + 체언	음절 1	용언 어간 첫째 자리 출현 빈도 상위 100개 음절
		음절 2	체언 어간 둘째 자리 출현 빈도 상위 100개 음절
	용언 + 용언	음절 1	용언 어간 첫째 자리 출현 빈도 상위 100개 음절
		음절 2	용언 어간 둘째 자리 출현 빈도 상위 100개 음절

실험 1의 음절조합 절차는 다음과 같다. 우선, 조건을 고려하여 선택된 음절 목록 내 상위빈도 음절 100개로 구성된 음절 세트 두 개를 중복 없이 무선 조합하여[32] 100개의 음절연쇄를 생성한다. 그리고 대조용 표제어 목록과 비교하여 음절연쇄 중 실제 단어가 차지하는 비율을 계산한다. 상술한 절차를 8개 조건에 걸쳐 각각 5만 번 반복 시행한다. 그런데 체언 + 체언 음절 연쇄와 용언 + 용언 음절 연쇄의 단어 생성률을 직접적으로 비교할 경우, 체언과 용언 표제어의 수에 큰 차이가 존재한다는 현실을 간과하게 된다. 전술한 바와 같이, 사전에 수록된 2음절 체언 표제어 수는 2음절 용언 표제어의 18.68배에 달한다. 따라서 표제어 목록 대조 시 체언 음절 조합은 용언 음절 조합보다 실제 단어를

32) 중복이 허용되지 않는 무선조합의 경우, 한 번 무선조합에 참여한 음절은 두 번 다시 무선조합에 사용되지 않는다.

생성할 가능성이 매우 크다. 이를 고려하여 실험 1에서는 용언 + 용언 음절의 무선조합 시 체언 대 용언 표제어 수 비율을 반영하는 방식을 채택했다. 즉 용언 음절 100개 + 용언 음절 100개를 무선 조합할 때 음절의 중복사용을 허용함으로써[33] 총 1,860개(100개 × 용언 대비 체언 표제어의 비율 18.6배)의 음절연쇄를 생성한 것이다. 이는 비중복 무선조합 방식에 비해 훨씬 다양한 음절연쇄를 만들어낼 수 있으므로 대조용 표제어 목록과 비교 시 실제 단어와 일치할 확률을 좀 더 높일 수 있다.

세 번째는 무선 조합된 음절연쇄를 표제어 목록과 대조하여 실제 단어와 일치하는 음절연쇄의 비율을 측정하는 단계이다. 이때 사용되는 대조용 표제어 목록은 총 네 가지로 구분된다. 이들 네 가지 대조용 표제어 목록을 조건별로 정리하면 아래의 <표 5>와 같다.

<표 5> 실험 1: 조건별 대조용 표제어 목록

위치고려 여부	품사조합 순서	대조용 표제어 목록			
		전체 모집단	전체 표본	품사별 모집단	품사별 표본
위치 비고려	체언 + 체언	모든 품사 표제어	모든 품사 1,000개	모든 체언 표제어	체언 1,000개
	체언 + 용언	모든 품사 표제어	모든 품사 1,000개	모든 체언 + 용언 표제어	체언 500개 + 용언 500개
	용언 + 체언	모든 품사 표제어	모든 품사 1,000개	모든 체언 + 용언 표제어	체언 500개 + 용언 500개
	용언 + 용언	모든 품사 표제어	모든 품사 1,000개	모든 용언 표제어	용언 1,000개
위치 고려	체언 + 체언	모든 품사 표제어	모든 품사 1,000개	모든 체언 표제어	체언 1,000개
	체언 + 용언	모든 품사 표제어	모든 품사 1,000개	모든 체언 + 용언 표제어	체언 500개 + 용언 500개
	용언 + 체언	모든 품사 표제어	모든 품사 1,000개	모든 체언 + 용언 표제어	체언 500개 + 용언 500개
	용언 + 용언	모든 품사 표제어	모든 품사 1,000개	모든 용언 표제어	용언 1,000개

아래는 대조용 표제어 목록별로 표제어 - 음절연쇄 간 대조절차를 설명한 것이다.

33) 중복이 허용되는 무선조합의 경우, 한 번 무선조합에 참여한 음절도 여러 번 무선조합에 재사용될 수 있다.

(1) 전체 표제어 모집단과 대조: 무선 조합된 100개의 음절연쇄를 사전에 수록된 전체 표제어 목록(n = 147,669)과 대조한 후 이들 표제어와 일치하는 음절연쇄의 수(단어 생성률)를 집계한다. 8개 조건에 걸쳐 단어 생성률 100개짜리 세트를 총 500번 무선 산출한다.

(2) 전체 표제어 표본과 대조: 무선 조합된 100개의 음절연쇄를 사전 수록 전체 표제어 목록에서 무선 추출된 1,000개 표본[34]과 대조한 후 이들 표제어와 일치하는 음절연쇄의 수를 집계한다. 8개 조건에 걸쳐 단어 생성률 100개짜리 세트를 총 500번 무선 산출한다.

(3) 품사별 표제어 모집단과 대조

(가) 체언 + 체언 음절 연쇄: 무선 조합된 100개의 음절연쇄를 사전에 수록된 전체 체언 표제어 목록(n = 128,308)과 대조한 후 이들 표제어와 일치하는 음절연쇄의 수를 집계한다. 단어 생성률 100개짜리 세트를 총 500번 무선 산출한다.

(나) 용언 + 용언 음절 연쇄: 무선 조합된 1,860개의 음절연쇄를 사전에 수록된 전체 용언 표제어 목록(n = 6,868)과 대조한 후 이들 표제어와 일치하는 음절연쇄의 수를 집계한다. 단어 생성률 100개짜리 세트를 총 500번 무선 산출한다.

(다) 체언 + 용언 또는 용언 + 체언 음절 연쇄: 무선 조합된 100개의 음절연쇄를 사전에 수록된 전체 체언 + 용언 표제어 목록(n = 135,176)과 대조한 후 이들 표제어와 일치하는 음절연쇄의 수를 집계한다. 단어 생성률 100개짜리 세트를 총 500번 무선 산출한다.

(4) 품사별 표제어 표본과 대조

(가) 체언 + 체언 음절 연쇄: 무선 조합된 100개의 음절연쇄를 사전에서 무선 추출된 체언 표제어 표본(n = 1,000)과 대조한 후 이들 표제어와 일치하는 음절연쇄의 수를 집계한다. 단어 생성률 100개짜리 세트를 총 500번 무선 산출한다.

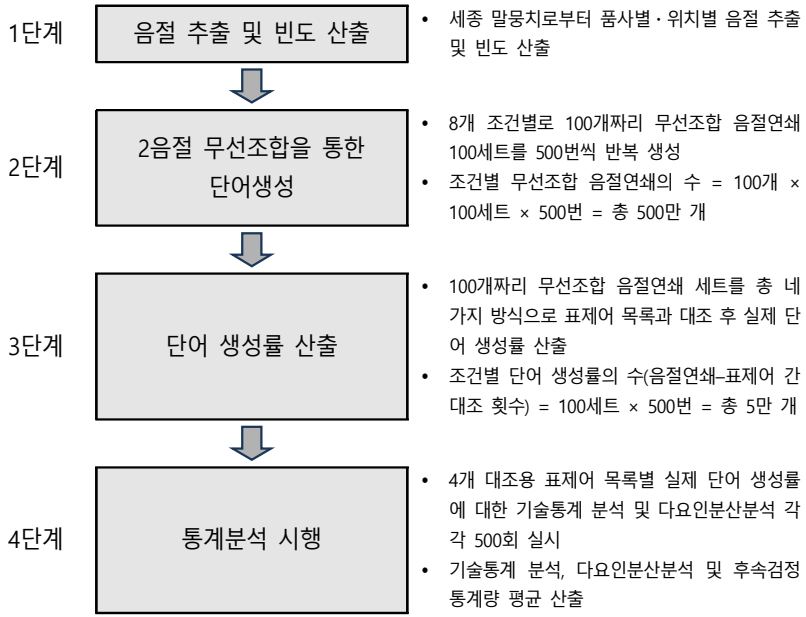
(나) 용언 + 용언 음절 연쇄: 무선 조합된 1,860개의 음절연쇄를 사전에서 무선 추출된 용언 표제어 표본(n = 1,000)과 대조한 후 이들 표제어와 일치하는 음절연쇄의 수를 집계한다. 단어 생성률 100개짜리 세트를 총 500번 무선 산출한다.

(다) 체언 + 용언 또는 용언 + 체언 음절 연쇄: 무선 조합된 100개의 음절연쇄를

34) 단, 표제어 표본 목록은 무선조합 음절연쇄 1세트(n = 100)와 대조할 때마다 표제어 모집단 목록에서 새로이 무선 추출되므로, 음절연쇄와의 대조 작업 시 완벽하게 동일한 표제어 표본 목록이 2회 이상 반복 사용될 가능성은 매우 희박하다.

사전에서 무선 추출된 체언(n = 500) + 용언(n = 500) 표제어 통합 표본과 대조한 후 이들 표제어와 일치하는 음절연쇄의 수를 집계한다. 단어 생성률 100개짜리 세트를 총 500번 무선 산출한다.

마지막으로, 8개 조건별로 수집된 단어 생성률 100개짜리 세트 500개에 대해 기술통계 분석, 다요인분산분석 및 후속검정을 시행한 후 평균을 계산한다. 이때 무선조합 음절연쇄를 표제어 목록과 대조하는 방식이 네 가지이므로, 네 유형의 자료에 대해 각각 500번의 통계분석을 실시한다. 이어서 500번의 기술통계 분석, 다요인분산분석 및 후속검정을 통해 얻은 통계량에 대해 평균을 산출한다. 상술한 4단계 실험절차를 도식화하면 아래의 <그림 1>과 같다.



<그림 1> 실험 1: 4단계 실험절차

2.2. 결과 및 논의

실험 1, 2, 3에서는 대조용 표제어 목록에 따라 무선 음절조합의 단어 생성률을 네 가지 방식 — 전체 표제어 모집단 대조, 전체 표제어 표본 대조, 품사별 표제어 모집단 대조, 그리고 품사별 표제어 표본 대조 — 으로 총 500번 산출했다. 그리고 음절 요인이 네 버전의 단어 생성률에 미치는 영향을 검증하기 위해 버전별로 500회의 기술통계 분석, 다요인분산분석 및 후속검정을 실시한 후 통계량 평균을 계산했다. 아래의 <표 6>은 네 버전의 무선 음절조합 단어 생성률에 대한 기술통계량 평균을 조건별로 정리한 것이다.

<표 6> 실험 1: 조건별 무선 단어 생성률 기술통계 평균[35)]

대조용 목록	통계량	위치 비고려				위치고려			
		단일 품사		이종 품사		단일 품사		이종 품사	
		NN	VV	NV	VN	NN	VV	NV	VN
HP	*M*	80.89	41.91	55.49	57.24	77.38	46.61	67.74	51.09
	SE	0.01	0.02	0.01	0.01	0.01	0.02	0.02	0.01
	SD	3.22	3.48	3.29	3.15	3.32	3.42	3.50	3.14
HS	*M*	2.14	1.08	1.48	1.49	2.02	1.20	1.77	1.34
	SE	0.01	0.005	0.01	0.01	0.01	0.005	0.01	0.01
	SD	1.43	1.02	1.20	1.21	1.39	1.07	1.30	1.14
HPC	*M*	79.23	6.02	54.17	55.83	75.80	6.21	66.01	49.96
	SE	0.01	0.002	0.01	0.01	0.01	0.002	0.02	0.01
	SD	3.27	0.50	3.32	3.17	3.33	0.51	3.55	3.12
HSC	*M*	1.11	1.05	1.42	1.24	1.05	1.08	1.62	1.13
	SE	0.005	0.001	0.01	0.005	0.005	0.001	0.01	0.005
	SD	1.05	0.23	1.18	1.09	1.02	0.23	1.24	1.04

35) 첫째, 대조용 표제어 목록 유형의 HP는 전체 표제어 모집단(Headword Population: 표준국어대사전에 등장하는 모든 표제어), HS는 전체 표제어 표본(Headword Sample: 표준국어대사전에 등장하는 모든 표제어 중에서 무작위로 뽑은 1,000개의 표제어), HPC는 품사별 표제어 모집단(Headword Population for the relevant word Class: 표준국어대사전에 등장하는 모든 체언, 용언 또는 체언 + 용언 표제어), 그리고 HSC는 품사별 표제어 표본(Headword Sample for the relevant word Class: 표준국어대사전에 등장하는 모든 체언, 용언 또는 체언 + 용언 표제어 중에서 각각 무작위로 뽑은 1,000개의 표제어)을 가리킨다. 둘째, 통계량의 *M*은 평균(mean), *SE*는 평균의 표준오차(standard error) 그리고 *SD*는 표준편차(standard deviation)를 가리킨다. 셋째, NN(noun + noun)은 체언 + 체언, VV(verb + verb)는 용언 + 용언, NV(noun

아래의 <표 7>은 음절요인(품사조합 순서, 품사혼합 여부, 위치고려 여부)이 무선 음절조합 단어 생성률에 미치는 영향을 검증한 다요인분산분석의 통계량 평균을 버전별로 구분하여 정리한 것이다. <표 7>에 따르면 HSC를 제외한 세 버전에서 품사조합 순서의 주효과가 확인되었다. 품사혼합 여부는 HS를 제외한 세 버전에서 단어 생성률에 통계적으로 유의미한 영향을 미치는 것으로 나타났다. 위치고려 여부는 전체(HP) 또는 개별(HPC) 품사 모집단 표제어와 대조한 경우에만 단어 생성률에 통계적으로 유의미한 영향을 미치는 것으로 검증되었다. 이는 무선조합 음절쌍을 충분히 큰 표제어 목록과 대조할 때에만 위치고려 여부의 효과가 제한적으로 나타남을 의미한다. 그리고 HP와 HPC 버전에서 위치고려 여부는 품사혼합 여부 및 품사조합 순서와 통계적으로 유의미한 상호작용을 나타냈다. 이는 위치고려 여부에 따라 품사조합 조건들의 생산성이 서로 다른 패턴을 나타냄을 시사한다.

<표 7> 실험 1: 음절요인이 무선 단어 생성률에 미치는 영향 분산분석 결과 평균[36)]

대조용 목록	변수	자유도	제곱합	F	p	ω^2
HP	품사조합 순서	2	127166.12	5794.98	< .0001	0.85
	품사혼합 여부	1	2908.67	265.10	< .0001	0.02
	위치고려 여부	1	674.84	61.49	< .0001	0.004
	품사혼합 × 위치고려	1	311.81	28.42	.0001	0.002
	품사순서 × 위치고려	2	10177.64	463.78	< .0001	0.07
	오차	792	8712.86			
HS	품사조합 순서	2	98.18	32.83	< .0001	0.07
	품사혼합 여부	1	4.23	2.85	.341	0.002
	위치고려 여부	1	3.08	2.06	.386	0.001
	품사혼합 × 위치고려	1	2.90	1.93	.389	0.001
	품사순서 × 위치고려	2	11.47	3.85	.169	0.01
	오차	792	1185.05			
HPC	품사조합 순서	2	514921.22	31429.38	< .0001	0.90

+ verb)는 체언 + 용언 그리고 VN(verb + noun)은 용언 + 체언 음절 조합을 가리킨다.

36) ω^2(omega squared)는 F-검정 시 주로 사용되는 효과 크기(effect size) 유형이다. ω^2의 범위는 -1에서 +1 사이이며, F 값이 1보다 작은 경우 0보다 작은 값으로 표현된다. ω^2가 크면 클수록 검정결과가 통계적으로 유의미한지 여부와 관계없이 해당 독립변수의 설명력이 높은 것으로 간주할 수 있다(Field 2010).

	품사혼합 여부	1	43074.93	5258.42	< .0001	0.08
	위치고려 여부	1	100.83	12.30	.015	0.0002
	품사혼합 × 위치고려	1	1066.23	130.18	< .0001	0.002
	품사순서 × 위치고려	2	8173.33	498.98	< .0001	0.01
	오차	792	6507.02			
HSC	품사조합 순서	2	15.23	8.22	.057	0.02
	품사혼합 여부	1	18.05	19.41	.017	0.02
	위치고려 여부	1	1.82	1.97	.392	0.001
	품사혼합 × 위치고려	1	1.94	2.12	.399	0.001
	품사순서 × 위치고려	2	5.91	3.21	.237	0.01
	오차	792	731.74			

위치고려 변수가 단어 생성률에 미치는 영향이 품사혼합 여부 및 품사조합 순서에 따라 어떻게 달라지는지를 확인하기 위해 500회의 후속검정을 실시하고 통계량 평균을 산출했다. 그 결과, 대조용 표제어 목록이 HP일 때 위치고려 NN 조합은 위치 비고려 NN 조합보다 통계적으로 유의미하게 더 많은 단어를 생성해낸 것으로 나타났다($t(99) = -7.66$, $p < .0001$, $r = 0.61$).[37] 위치고려 VN 조합과 위치 비고려 VN 조합을 비교했을 때도 마찬가지였다($t(99) = -13.89$, $p < .0001$, $r = 0.81$). 반면에 위치고려 VV 조합은 위치 비고려 VV 조합보다 통계적으로 유의미하게 더 낮은 단어 성생률을 보였다($t(99) = 9.71$, $p < .0001$, $r = 0.69$). 위치고려 NV 조합과 위치 비고려 NV 조합 간 비교에서도 유사한 패턴이 관찰되었다($t(99) = 25.73$, $p < .0001$, $r = 0.93$). 따라서 음절이 유래한 단어 내 출현위치를 고려하여 음절쌍을 무선 조합했을 때 위치에 대한 고려 없이 무선 조합했을 경우보다 더 많은 단어를 만들어낼 것이라는 가정은 부분적으로 지지되었다.

이어서 품사혼합 여부 및 품사조합 순서 변수가 단어 생성률에 미치는 영향이 위치고려 여부에 따라 어떻게 달라지는지를 확인하기 위해 500회의 후속검정을 실시한 후 통계량 평균을 계산했다. 위치를 고려하여 음절을 조합한 경우, 이종 품사 조합과 단일 품사 조합 조건의 단어 생성률에 통계적으로 유의미한

37) 지면의 제한으로 본고에서는 실험 1, 2, 3 모두 HP 대조 조건의 후속검정 통계량만을 소개한다. 4개 표제어 대조 조건별 후속검정 통계량 일체는 제1저자의 github 저장소(https://github.com/cognitivepsychology/cognitive_psychology)를 참조하라.

차이가 있었다(NV/VN vs. NN: $t(99) = -20.13/-58.24$, $p < .0001$, $r = 0.89/0.99$; NV/VN vs. VV: $t(99) = 43.54/9.71$, $p < .0001$, $r = 0.97/0.69$). 이때 NN 음절 조합은 두 가지 이종 품사 음절 조합(NV/VN)보다 높은 단어 생성률을 나타낸 반면, VV 음절 조합은 두 가지 이종 품사 음절 조합보다 낮은 단어 생성률을 나타냈다.

아울러 위치를 고려하지 않고 음절을 조합한 경우 또한 이종 품사 조합과 단일 품사 조합 조건 간에 통계적으로 유의미한 단어 생성률 차이가 관찰되었다(NV/VN vs. NN: $t(99) = -55.56/-52.90$, $p < .0001$, $r = 0.98/0.98$; NV/VN vs. VV: $t(99) = 28.52/32.89$, $p < .0001$, $r = 0.94/0.96$). 이때 앞서 위치를 고려하여 음절을 조합했을 때와 마찬가지로 NN 음절 조합은 두 가지 이종 품사 음절 조합보다 생산성이 높았지만, VV 음절 조합은 두 가지 이종 품사 음절 조합보다 생산성이 낮았다. 이렇듯 위치고려 여부에 관계없이 NN 음절 조합이 이종 품사 음절 조합보다 더 많은 단어를 만들어냈다는 사실은 체언 음절의 높은 생산성을 잘 보여준다.

3. 실험 2

어종이 고유어냐 한자어냐에 따라 단어를 이루는 음절의 언어적 · 통계적 패턴이 일정한 차이를 나타낸다(권인한 1997; 남성현 · 김선회 2018; 안소진 2009). 아울러 모어화자의 심성어휘집에 고유어와 한자어가 분리되어 표상되어 있을 가능성을 시사하는 증거 또한 적지 않다(예를 들면, 이광오 2003; Kim & Na 2000). 따라서 실험 2는 두 개의 고빈도 음절을 무작위로 조합했을 때 음절이 유래한 어종이 단어 생성률에 미치는 영향을 통계적으로 검증하는 데 목적을 둔다. 이때 체언이 용언에 비해 상대적으로 한자어 비율이 높고 고유어 비율이 낮은 것을 고려하여, 품사 변수 효과의 혼입을 방지하기 위해 단일 품사(체언 또는 용언)에서 추출한 음절끼리 무선 조합하는 것을 원칙으로 한다. 또한 위치 변수가 단어 생성률에 미치는 영향을 통제하기 위해 개별 품사 표제어 내 첫째 자리 고빈도 음절과 개별 품사 표제어 내 둘째 자리 고빈도 음절을 무선 조합한다. 이로써 동일한 어종으로부터 유래한 음절쌍을 무선 조합했을

때, 서로 다른 어종으로부터 유래한 음절쌍을 무선 조합했을 경우보다 실제 한국어 단어를 더 많이 생성할 것이라는 가정을 통계적으로 검증하고자 한다.

3.1. 연구방법

3.1.1. 연구자료

실험 2에서는 품사별 · 어종별 한국어 음절 정보 추출을 위해 표준국어대사전을 사용했다. 이때 무선 조합된 2음절 연쇄가 실제 단어인지 여부를 확인하기 위해 표준국어대사전 수록 2음절 단어들과 대조하므로, 표제어 출현형 빈도 상위 100개 음절을 실험재료로 채택했다. 실험 2에 투입될 음절의 품사로는 실험 1과 마찬가지로 체언(체언)과 용언(동사와 형용사)이, 어종으로는 한글과 한자가 채택되었다. 체언/용언 음절은 한글 음절과 한자 유래 음절로 나뉘며, 한글/한자 음절은 다시 첫째 음절과 둘째 음절로 구분된다. 아래의 <표 8>은 실험 2에 사용된 음절의 조건별 빈도 정보를 정리한 것이다. 단, 실험 2에 사용된 대조용 표제어 목록은 실험 1과 동일하다.

<표 8> 실험 2: 조건별 음절빈도 정보

품사	어종	위치	출현형/유형	음절 빈도	
				전체	상위 100개
일반 명사	한글	첫째	출현형	76,329	40,195
			유형	1,680	100
		둘째	출현형	81,268	44,476
			유형	1,588	100
	한자	첫째	출현형	205,295	125,315
			유형	788	100
		둘째	출현형	196,675	120,500
			유형	600	100
용언	한글	첫째	출현형	38,458	18,171
			유형	1,510	100

		둘째	출현형	38,344	22,229
			유형	1,378	100
	한자	첫째	출현형	46,663	27,359
			유형	540	100
		둘째	출현형	45,539	26,153
			유형	466	100

3.1.2. 자료처리 도구

실험 2에 사용된 자료처리 도구는 실험 1과 동일하다.

3.1.3. 통계분석 방법

실험 2에서는 품사 변수를 통제했을 때 어종 변수가 음절 간 무선조합 시 실제 단어 생성률에 미치는 영향을 통계적으로 검증하기 위해 다요인분산분석을 사용했으며, 유의수준 α는 .05로 설정했다. 후속검정 방법으로는 Bonferroni 기법을 사용했다. 아래의 <표 9>는 실험 2 결과의 통계분석에 투입된 독립변수와 종속변수를 정리한 것이다. 실험 2의 독립변수는 어종혼합 여부[38], 어종조합 순서[39], 품사[40] 세 가지이며, 종속변수는 평균 실제 단어 생성률이다.

<표 9> 실험 2: 독립변수와 종속변수

변수	내용	음절조합 조건
독립변수	어종혼합 여부	• 단일 어종 간 음절 무선조합 • 이종 어종 간 음절 무선조합
	어종조합 순서	• 한글 + 한글 음절 순서로 무선조합(단일 어종 조합) • 한글 + 한자 음절 순서로 무선조합(이종 어종 조합)

38) 음절 간 무선조합 시 한글과 한자 음절을 혼합했는지, 아니면 동일 어종 음절끼리 조합했는지 여부를 가리킨다.

39) 음절 간 무선조합 시 한글 + 한글, 한글 + 한자, 한자 + 한글, 한자 + 한자 중 어느 순서로 음절을 조합했는지를 가리킨다.

40) 체언 음절끼리 조합했는지, 아니면 용언 음절끼리 조합했는지 여부를 가리킨다.

		• 한자 + 한글 음절 순서로 무선조합(이종 어종 조합) • 한자 + 한자 음절 순서로 무선조합(단일 어종 조합)
	품사	• 체언: 체언 음절 간 무선조합 • 용언: 용언 음절 간 무선조합
종속변수	실제 단어 생성률	무선조합 음절연쇄 중에서 대조용 사전에 수록된 실제 단어와 일치하는 사례의 비율

3.1.4. 연구절차

실험 2의 절차는 실험 1과 같이 네 단계로 구분된다. 첫 번째는 말뭉치로부터 음절을 추출하여 출현형 빈도를 산출하는 단계이다. 먼저 프로그래밍 언어 R을 바탕으로 표준국어대사전을 분석, 체언과 용언 표제어에 포함된 한글 또는 한자 음절[41] 중 첫째 또는 둘째 자리에 쓰인 음절의 출현형 빈도를 산출했다. 이로써 품사별 · 어종별 · 위치별로 총 8개의 음절목록을 추출했다.

두 번째, 목록 내 상위빈도 음절 100개로 구성된 실험용 음절 세트 두 개를 일정한 조건에 따라 무선 조합하여 두 글자로 이루어진 음절연쇄 100개를 생성한다. 아래의 <표 10>은 실험 2의 무선 음절조합 절차에 사용된 음절목록을 조건별로 정리한 것이다. 품사(체언 + 체언 vs. 용언 + 용언)에 따라 어종조합 순서(한글 + 한글 vs. 한글 + 한자 vs. 한자 + 한글 vs. 한자 + 한자)를 달리하면 총 8개의 음절조합 조건이 만들어진다. 실험 2의 음절조합 절차는 실험 1과 동일하다.

<표 10> 실험 2: 무선 음절조합에 투입되는 조건별 음절 목록

품사	어종조합 순서	음절조합	
		위치	음절목록
체언	한글 + 한글	음절 1	한글 유래 첫째 자리 출현 빈도 상위 100개 음절
		음절 2	한글 유래 둘째 자리 출현 빈도 상위 100개 음절
	한글 + 한자	음절 1	한글 유래 첫째 자리 출현 빈도 상위 100개 음절

41) 가령 “취(取)하 - ”라는 용언의 경우, “취”는 한자 음절로 그리고 “하”는 한글 음절로 분석되었다.

		음절 2 한자 유래 둘째 자리 출현 빈도 상위 100개 음절
	한자 + 한글	음절 1 한자 유래 첫째 자리 출현 빈도 상위 100개 음절
		음절 2 한글 유래 둘째 자리 출현 빈도 상위 100개 음절
	한자 + 한자	음절 1 한자 유래 첫째 자리 출현 빈도 상위 100개 음절
		음절 2 한자 유래 둘째 자리 출현 빈도 상위 100개 음절
용언	한글 + 한글	음절 1 한글 유래 첫째 자리 출현 빈도 상위 100개 음절
		음절 2 한글 유래 둘째 자리 출현 빈도 상위 100개 음절
	한글 + 한자	음절 1 한글 유래 첫째 자리 출현 빈도 상위 100개 음절
		음절 2 한자 유래 둘째 자리 출현 빈도 상위 100개 음절
	한자 + 한글	음절 1 한자 유래 첫째 자리 출현 빈도 상위 100개 음절
		음절 2 한글 유래 둘째 자리 출현 빈도 상위 100개 음절
	한자 + 한자	음절 1 한자 유래 첫째 자리 출현 빈도 상위 100개 음절
		음절 2 한자 유래 둘째 자리 출현 빈도 상위 100개 음절

세 번째는 무선 조합된 음절연쇄를 표제어 목록과 대조하여 실제 단어와 일치하는지 여부를 확인하는 단계이다. 이때 사용되는 대조용 표제어 목록을 조건별로 정리하면 아래의 <표 11>과 같다.

<표 11> 실험 2: 조건별 대조용 표제어 목록

품사	어종조합 순서	대조용 표제어 목록			
		전체 모집단	전체 표본	품사별 모집단	품사별 표본
체언	한글 + 한글	모든 품사 표제어	모든 품사 1,000개	모든 체언 표제어	체언 1,000개 표제어
	한글 + 한자	모든 품사 표제어	모든 품사 1,000개	모든 체언 표제어	체언 1,000개 표제어
	한자 + 한글	모든 품사 표제어	모든 품사 1,000개	모든 체언 표제어	체언 1,000개 표제어
	한자 + 한자	모든 품사 표제어	모든 품사 1,000개	모든 체언 표제어	체언 1,000개 표제어
용언	한글 + 한글	모든 품사 표제어	모든 품사 1,000개	모든 용언 표제어	용언 1,000개 표제어
	한글 + 한자	모든 품사 표제어	모든 품사 1,000개	모든 용언 표제어	용언 1,000개 표제어
	한자 + 한글	모든 품사 표제어	모든 품사 1,000개	모든 용언 표제어	용언 1,000개 표제어
	한자 + 한자	모든 품사 표제어	모든 품사 1,000개	모든 용언 표제어	용언 1,000개 표제어

아래는 대조용 표제어 목록별로 표제어 - 음절연쇄 간 대조절차를 설명한 것이다.

(1) 전체 표제어 모집단과 대조: 실험 1과 동일하다.
(2) 전체 표제어 표본과 대조: 실험 1과 동일하다.
(3) 품사별 표제어 모집단과 대조
  (가) 체언 + 체언 음절 연쇄: 실험 1과 동일하다.
  (나) 용언 + 용언 음절 연쇄: 실험 1과 동일하다.
(4) 품사별 표제어 표본과 대조
  (가) 체언 + 체언 음절 연쇄: 실험 1과 동일하다.
  (나) 용언 + 용언 음절 연쇄: 실험 1과 동일하다.

마지막으로, 8개 조건별로 수집된 단어 생성률 100개짜리 세트 500개에 대해 기술통계 분석, 다요인분산분석 및 후속검정을 시행한다. 해당 통계분석 절차는 실험 1과 동일하다.

3.2. 결과 및 논의

실험 1, 2, 3에서는 대조용 표제어 목록에 따라 무선 음절조합의 단어 생성률을 네 가지 방식으로 각각 500번 산출했다. 그리고 음절 요인이 네 버전의 단어 생성률에 미치는 영향을 검증하기 위해 버전별로 500회의 기술통계 분석, 다요인분산분석 및 후속검정을 실시한 후 통계량 평균을 계산했다. 아래의 <표 12>는 네 버전의 무선 음절조합 단어 생성률에 대한 기술통계량 평균을 조건별로 구분하여 정리한 것이다.

<표 12> 실험 2: 조건별 무선 단어 생성률 기술통계 평균[42)]

대조용 목록	통계량	체언				용언			
		단일 어종		이종 어종		단일 어종		이종 어종	
		KK	CC	KC	CK	KK	CC	KC	CK
HP	*M*	51.91	91.54	63.93	65.08	40.06	86.57	52.26	51.38
	SE	0.02	0.01	0.02	0.02	0.02	0.01	0.02	0.02
	SD	4.03	2.63	3.46	3.56	4.01	3.11	3.41	3.48

HS	*M*	1.09	2.56	1.57	1.59	0.78	2.30	1.24	1.21
	SE	0.005	0.007	0.005	0.01	0.004	0.01	0.005	0.005
	SD	1.03	1.56	1.23	1.24	0.87	1.50	1.10	1.08
HPC	*M*	47.63	90.70	61.96	62.38	7.57	2.48	4.29	5.15
	SE	0.02	0.01	0.02	0.02	0.002	0.001	0.002	0.002
	SD	4.01	2.75	3.45	3.52	0.55	0.33	0.43	0.47
HSC	*M*	0.66	1.25	0.86	0.87	1.32	0.43	0.75	0.90
	SE	0.004	0.005	0.004	0.004	0.001	0.001	0.001	0.001
	SD	0.80	1.11	0.92	0.92	0.26	0.15	0.20	0.22

아래의 <표 13>은 음절요인(어종조합 순서, 어종혼합 여부, 음절 유래 품사)이 무선 음절조합 단어 생성률에 미치는 영향을 검증한 다요인분산분석의 통계량 평균을 버전별로 구분하여 정리한 것이다. <표 13>에 따르면 HSC를 제외한 세 버전에서 어종조합 순서, 어종혼합 여부 그리고 품사의 주효과가 확인되었다. 이는 무선조합 음절연쇄를 HS와 같은 소규모 표제어 표본과 대조할 때에도 어종혼합 여부 및 어종조합 순서의 효과가 여전히 유효함을 의미한다. 따라서 어종혼합 여부 및 어종조합 순서는 단어 생성률에 충분히 강력한 영향력을 행사하고 있다고 진단할 수 있다. 그러나 HP와 HPC 버전에서 품사는 어종혼합 여부 및 어종조합 순서와 통계적으로 유의미한 상호작용을 나타냈다. 아울러 HSC 버전에서도 품사와 어종조합 순서 간 상호작용은 통계적으로 유의미했다. 이는 음절이 유래한 품사에 따라 어종조합 조건들의 생산성이 각기 다른 양상을 나타냄을 의미한다.

<표 13> 실험 2: 음절요인이 무선 단어 생성률에 미치는 영향 분산분석 결과 평균

대조용 목록	변수	자유도	제곱합	*F*	*p*	ω^2
HP	어종조합 순서	2	185568.96	7639.75	< .0001	0.78
	어종혼합 여부	1	17517.23	1442.29	< .0001	0.07
	품사	1	22255.95	1832.55	< .0001	0.09
	어종혼합 × 품사	1	925.83	76.21	< .0001	0.004

42) KK(Korean + Korean)는 한글 + 한글, CC(Chinese + Chinese)는 한자 + 한자, KC(Korean + Chinese)는 한글 + 한자 그리고 CK(Chinese + Korean)는 한자 + 한글 음절 조합을 가리킨다.

	어종순서 × 품사	2	1311.43	53.99	< .0001	0.01
	오차	792	9643.10			
HS	어종조합 순서	2	227.66	77.23	< .0001	0.15
	어종혼합 여부	1	18.69	12.69	.049	0.01
	품사	1	23.14	15.71	.031	0.01
	어종혼합 × 품사	1	2.99	2.02	.386	0.001
	어종순서 × 품사	2	5.44	1.84	.357	0.002
	오차	792	1170.34			
HPC	어종조합 순서	2	36103.30	2983.36	< .0001	0.04
	어종혼합 여부	1	2668.04	440.95	< .0001	0.003
	품사	1	739181.34	122159.40	< .0001	0.88
	어종혼합 × 품사	1	2242.20	370.59	< .0001	0.003
	어종순서 × 품사	2	57990.32	4791.89	< .0001	0.07
	오차	792	4818.19			
HSC	어종조합 순서	2	5.15	5.62	.120	0.01
	어종혼합 여부	1	2.32	5.07	.237	0.004
	품사	1	1.71	3.53	.292	0.003
	어종혼합 × 품사	1	1.29	2.82	.343	0.002
	어종순서 × 품사	2	57.08	62.00	< .0001	0.13
	오차	792	367.06			

품사 변수가 단어 생성률에 미치는 영향이 어종혼합 여부 및 어종조합 순서에 따라 어떻게 달라지는지를 살펴보기 위해 500회의 후속검정을 실시하고 통계량 평균을 산출했다. 그 결과, 대조용 표제어 목록이 HP 버전일 경우 체언 KK 조합은 용언 KK 조합보다 통계적으로 유의미하게 더 많은 단어를 생성해 낸 것으로 나타났다($t(99) = -20.98$, $p < .0001$, $r = 0.90$). 체언 KC 조합과 용언 KC 조합을 비교했을 때도 마찬가지였다($t(99) = -24.19$, $p < .0001$, $r = 0.92$). 체언 CC 조합 또한 용언 CC 조합보다 통계적으로 유의미하게 더 높은 단어 성생률을 나타냈다($t(99) = -12.27$, $p < .0001$, $r = 0.77$). 그리고 체언 CK 조합과 용언 CK 조합 간 비교에서도 유사한 패턴이 관찰되었다($t(99) = -27.68$, $p < .0001$, $r = 0.94$).

이어서 어종혼합 여부 및 어종조합 순서 변수가 단어 생성률에 미치는 영향이 품사에 따라 어떻게 달라지는지를 확인하기 위해 500회의 후속검정을 실시한 후 통계량 평균을 계산했다. 체언 음절끼리 무선 조합한 경우, 이종 어종

조합과 단일 어종 조합의 단어 생성률 간에 통계적으로 유의미한 차이가 있었다(KC/CK vs. KK: $t(99) = 22.86/24.69$, $p < .0001$, $r = 0.92/0.93$; KC/CK vs. CC: $t(99) = 64.05/60.22$, $p < .0001$, $r = 0.99/0.99$). 이때 KK 조합은 두 가지 이종 어종 조합(KC/CK)보다 낮은 단어 생성률을 나타낸 반면, CC 조합은 두 가지 이종 어종 조합보다 높은 단어 생성률을 나타냈다.

아울러 용언 음절끼리 무선 조합한 경우 또한 이종 어종 조합과 단일 어종 조합 조건 간에 통계적으로 유의미한 단어 생성률 차이가 관찰되었다(KC/CK vs. KK: $t(99) = 23.32/21.48$, $p < .0001$, $r = 0.92/0.91$; KC/CK vs. CC: $t(99) = 75.03/76.26$, $p < .0001$, $r = 0.99/0.99$). 이때 앞서 체언 음절끼리 무선 조합한 경우와 마찬가지로 KK 조합은 두 가지 이종 어종 조합보다 생산성이 낮았지만, CC 조합은 두 가지 이종 어종 조합보다 높은 생산성을 보였다. 이렇듯 음절이 유래한 품사에 관계없이, CC 조합만 이종 어종 조합보다 단어 생성률이 통계적으로 유의미하게 높았다. 따라서 단일 어종 간 음절 조합이 이종 어종 간 음절 조합보다 더 많은 실제 단어를 만들어낼 것이라는 가정은 부분적으로 지지되었다. 그리고 CC 조합이 KK 조합보다 실제 단어를 더 많이 만들어낼 것이라는 가정은 완벽하게 지지되었다. 이렇듯 음절이 유래한 품사에 관계없이 단어 내 한자 음절의 비율이 높을수록 단어 생성률이 높아진다는 사실은 한자 음절의 높은 생산성을 단적으로 보여준다.

4. 실험 3

한국어 표제어 중 가장 큰 비중을 차지하는 품사인 체언과 용언은 기능, 형태, 의미 면에서 명확히 구분된다. 뿐만 아니라 체언은 용언에 비해 한자어의 점유율이 훨씬 높은 반면, 용언은 체언에 비해 고유어의 점유율이 훨씬 높다. 그런 까닭에 품사에 따라 단어를 구성하는 음절이 언어적 · 분포적으로 차이를 나타낼 가능성을 배제할 수 없다. 따라서 실험 3은 고빈도 음절을 두 개씩 무작위로 조합했을 때 음절이 유래한 품사가 실제 단어 생성률에 미치는 영향을 통계적으로 검증하는 데 목적을 둔다. 이때 품사와 어종 간 밀접한 관련성을 고려하여 어종 변수 효과의 혼입을 방지하기 위해 단일 어종(고유어 또는 한자

어)에서 추출한 음절끼리 무선 조합하는 것을 원칙으로 한다. 또한 위치 변수가 단어 생성률에 미치는 영향을 통제하기 위해 개별 어종 표제어 내 첫째 자리 고빈도 음절과 개별 어종 표제어 내 둘째 자리 고빈도 음절을 무선 조합한다. 이로써 동일한 품사로부터 유래한 음절쌍을 무선 조합했을 때, 서로 다른 품사로부터 유래한 음절쌍을 무선 조합했을 경우보다 실제 한국어 단어를 더 많이 만들어낼 것이라는 가정을 통계적으로 검증하고자 한다.

4.1. 연구방법

4.1.1. 연구자료

실험 3에 사용된 음절은 실험 2에 사용된 것과 동일하며, 실험 3에 사용된 대조용 표제어 목록 역시 실험 1, 2와 동일하다.

4.1.2. 자료처리 도구

실험 3에 사용된 자료처리 도구는 실험 1, 2와 동일하다.

4.1.3. 통계분석 방법

실험 3에서는 어종 변수 효과를 통제했을 때 품사 변수가 음절 간 무선조합 시 실제 단어 생성률에 미치는 영향을 통계적으로 검증하기 위해 다요인분산분석을 사용했으며, 유의수준 α는 .05로 설정했다. 후속검정 방법으로는 Bonferroni 기법을 사용했다. 아래의 <표 14>는 실험 3 결과의 통계분석에 투입된 독립변수와 종속변수를 정리한 것이다. 실험 3의 독립변수는 품사혼합 여부, 품사조합 순서, 어종[43] 세 가지이며, 종속변수는 평균 실제 단어 생성률이다.

43) 한글 유래 음절끼리 조합했는지, 아니면 한자 유래 음절끼리 조합했는지 여부를 가리킨다.

<표 14> 실험 3: 독립변수와 종속변수

변수	내용	음절조합 조건
독립변수	품사혼합 여부	• 단일 품사 간 음절 무선조합 • 이종 품사 간 음절 무선조합
	품사조합 순서	• 체언 + 체언 음절 순서로 무선조합(단일 품사 조합) • 용언 + 용언 음절 순서로 무선조합(단일 품사 조합) • 체언 + 용언 음절 순서로 무선조합(이종 품사 조합) • 용언 + 체언 음절 순서로 무선조합(이종 품사 조합)
	어종	• 한글: 한글 유래 음절 간 무선조합 • 한자: 한자 유래 음절 간 무선조합
종속변수	실제 단어 생성률	무선조합 음절연쇄 중에서 대조용 사전에 수록된 실제 단어와 일치하는 사례의 비율

4.1.4. 연구절차

실험 3의 절차는 실험 1, 2와 같이 네 단계로 나뉜다. 첫 번째는 말뭉치로부터 음절을 추출하여 출현형 빈도를 산출하는 단계이다. 먼저 프로그래밍 언어 R을 기반으로 표준국어대사전을 분석, 체언과 용언 표제어에 포함된 한글 또는 한자 음절 중 첫째 또는 둘째 자리에 쓰인 음절의 출현형 빈도를 산출했다. 이로써 품사별 · 어종별 · 위치별로 총 8개의 음절목록을 추출했다.

두 번째, 목록 내 상위빈도 음절 100개로 구성된 실험용 음절 세트 두 개를 일정한 조건에 따라 무선 조합하여 두 글자로 이루어진 음절연쇄 100개를 생성한다. 아래의 <표 15>는 실험 3의 무선 음절조합 절차에 사용된 음절목록을 조건별로 정리한 것이다. 어종(한글 + 한글 vs. 한자 + 한자)에 따라 품사조합 순서(체언 + 체언 vs. 체언 + 용언 vs. 용언 + 체언 vs. 용언 + 용언)를 달리하면 총 8개의 음절조합 조건이 만들어진다. 실험 3의 음절조합 절차는 실험 1, 2와 동일하다.

<표 15> 실험 3: 무선 음절조합에 투입되는 조건별 음절 목록

어종	품사조합 순서	음절조합	
		위치	음절목록
한글	체언 + 체언	음절 1	한글 유래 첫째 자리 출현 빈도 상위 100개 음절
		음절 2	한글 유래 둘째 자리 출현 빈도 상위 100개 음절
	체언 + 용언	음절 1	한글 유래 첫째 자리 출현 빈도 상위 100개 음절
		음절 2	한글 유래 둘째 자리 출현 빈도 상위 100개 음절
	용언 + 체언	음절 1	한글 유래 첫째 자리 출현 빈도 상위 100개 음절
		음절 2	한글 유래 둘째 자리 출현 빈도 상위 100개 음절
	용언 + 용언	음절 1	한글 유래 첫째 자리 출현 빈도 상위 100개 음절
		음절 2	한글 유래 둘째 자리 출현 빈도 상위 100개 음절
한자	체언 + 체언	음절 1	한자 유래 첫째 자리 출현 빈도 상위 100개 음절
		음절 2	한자 유래 둘째 자리 출현 빈도 상위 100개 음절
	체언 + 용언	음절 1	한자 유래 첫째 자리 출현 빈도 상위 100개 음절
		음절 2	한자 유래 둘째 자리 출현 빈도 상위 100개 음절
	용언 + 체언	음절 1	한자 유래 첫째 자리 출현 빈도 상위 100개 음절
		음절 2	한자 유래 둘째 자리 출현 빈도 상위 100개 음절
	용언 + 용언	음절 1	한자 유래 첫째 자리 출현 빈도 상위 100개 음절
		음절 2	한자 유래 둘째 자리 출현 빈도 상위 100개 음절

세 번째는 무선 조합된 음절연쇄를 표제어 목록과 대조하여 실제 단어와 일치하는지 여부를 확인하는 단계이다. 이때 사용되는 대조용 표제어 목록을 조건별로 정리하면 아래의 <표 16>과 같다.

<표 16> 실험 3: 조건별 대조용 표제어 목록

어종	품사조합 순서	대조용 표제어 목록			
		전체 모집단	전체 표본	품사별 모집단	품사별 표본
한글	체언 + 용언	모든 품사 표제어	모든 품사 1,000개	모든 체언 표제어	체언 1,000개
	용언 + 체언	모든 품사 표제어	모든 품사 1,000개	모든 체언 + 용언 표제어	체언 500개 + 용언 500개
	체언 + 체언	모든 품사 표제어	모든 품사 1,000개	모든 체언 + 용언 표제어	체언 500개 + 용언 500개
	용언 + 용언	모든 품사 표제어	모든 품사 1,000개	모든 용언 표제어	용언 1,000개

한자	체언 + 용언	모든 품사 표제어	모든 품사 1,000개	모든 체언 표제어	체언 1,000개
	용언 + 체언	모든 품사 표제어	모든 품사 1,000개	모든 체언 + 용언 표제어	체언 500개 + 용언 500개
	체언 + 체언	모든 품사 표제어	모든 품사 1,000개	모든 체언 + 용언 표제어	체언 500개 + 용언 500개
	용언 + 용언	모든 품사 표제어	모든 품사 1,000개	모든 용언 표제어	용언 1,000개

아래는 대조용 표제어 목록별로 표제어 - 음절연쇄 간 대조절차를 설명한 것이다.

(1) 전체 표제어 모집단과 대조: 실험 1, 2와 동일하다.
(2) 전체 표제어 표본과 대조: 실험 1, 2와 동일하다.
(3) 품사별 표제어 모집단과 대조
  (가) 체언 + 체언 음절 연쇄: 실험 1, 2와 동일하다.
  (나) 용언 + 용언 음절 연쇄: 실험 1, 2와 동일하다.
(4) 품사별 표제어 표본과 대조
  (가) 체언 + 체언 음절 연쇄: 실험 1, 2와 동일하다.
  (나) 용언 + 용언 음절 연쇄: 실험 1, 2와 동일하다.

마지막으로, 8개 조건별로 수집된 단어 생성률 100개짜리 세트 500개에 대해 기술통계 분석, 다요인분산분석 및 후속검정을 시행한다. 해당 통계분석 절차는 실험 1, 2와 동일하다.

4.2. 결과 및 논의

실험 1, 2, 3에서는 대조용 표제어 목록에 따라 무선 음절조합의 단어 생성률을 네 가지 방식으로 각각 500번 산출했다. 그리고 음절 요인이 네 버전의 단어 생성률에 미치는 영향을 검증하기 위해 버전별로 500회의 기술통계 분석, 다요인분산분석 및 후속검정을 실시한 후 통계량 평균을 계산했다. 아래의 <표 17>은 네 버전의 무선 음절조합 단어 생성률에 대한 기술통계량 평균을 조건별로 구분하여 정리한 것이다.

<표 17> 실험 3: 조건별 무선 단어 생성률 기술통계 평균

대조용 목록	통계량	KK				CC			
		단일 품사		이종 품사		단일 품사		이종 품사	
		NN	VV	NV	VN	NN	VV	NV	VN
HP	*M*	51.92	40.04	42.60	45.57	91.54	86.56	88.24	89.12
	SE	0.02	0.02	0.02	0.02	0.01	0.01	0.01	0.01
	SD	4.03	4.02	3.90	3.95	2.64	3.10	2.93	2.89
HS	*M*	1.08	0.79	0.90	0.93	2.54	2.30	2.40	2.42
	SE	0.00	0.004	0.004	0.00	0.01	0.01	0.01	0.01
	SD	1.03	0.88	0.93	0.96	1.56	1.50	1.51	1.53
HPC	*M*	47.66	7.57	38.60	42.44	90.69	2.48	87.27	88.47
	SE	0.02	0.003	0.017	0.02	0.01	0.00	0.01	0.01
	SD	4.00	0.56	3.74	3.86	2.74	0.33	3.02	2.96
HSC	*M*	0.66	1.33	1.18	1.22	1.26	0.43	1.60	1.73
	SE	0.004	0.001	0.005	0.005	0.01	0.001	0.006	0.01
	SD	0.80	0.26	1.06	1.08	1.12	0.15	1.25	1.30

아래의 <표 18>은 음절요인(품사조합 순서, 품사혼합 여부, 음절 유래 어종)이 무선 음절조합 단어 생성률에 미치는 영향을 검증한 다요인분산분석의 통계량 평균을 버전별로 구분하여 정리한 것이다. <표 18>에 따르면 HP와 HPC 버전에서 품사조합 순서의 주효과가 나타났다. 품사혼합 여부는 HS를 제외한 세 버전에서 단어 생성률에 통계적으로 유의미한 영향을 미치는 것으로 확인되었다. 이는 무선조합 음절쌍을 HSC와 같은 소규모 표제어 표본과 대조할 때에도 품사혼합 여부의 효과가 여전히 유효함을 의미한다. 따라서 품사혼합 여부는 단어 생성률에 충분히 강력한 영향을 미치고 있다고 간주할 수 있다. 아울러 음절이 유래한 어종 또한 HSC를 제외한 세 버전에서 단어 생성률에 통계적으로 유의미한 영향을 미치는 것으로 검증되었다. 그러나 품사혼합 여부 및 품사조합 순서는 HS를 제외한 세 버전에서 어종과 통계적으로 유의미한 상호작용을 나타냈다. 이는 음절이 유래한 어종에 따라 품사조합 조건들의 생산성이 각기 다른 패턴을 나타냄을 가리킨다.

<표 18> 실험 3: 음절요인이 무선 단어 생성률에 미치는 영향 분산분석 결과 평균

대조용 목록	변수	자유도	제곱합	F	p	ω^2
HP	품사조합 순서	2	7499.68	311.02	< .0001	0.02
	품사혼합 여부	1	269.10	22.32	.001	0.001
	어종	1	384296.91	31870.45	< .0001	0.95
	품사혼합 × 어종	1	129.95	10.79	.028	0.0003
	품사순서 × 어종	2	1325.57	54.96	< .0001	0.003
	오차	792	9577.06			
HS	품사조합 순서	2	12.97	4.05	.156	0.01
	품사혼합 여부	1	2.88	1.80	.409	0.001
	어종	1	445.84	278.33	< .0001	0.25
	품사혼합 × 어종	1	2.96	1.84	.388	0.001
	품사순서 × 어종	2	5.89	1.84	.356	0.002
	오차	792	1271.29			
HPC	품사조합 순서	2	412214.79	23402.72	< .0001	0.47
	품사혼합 여부	1	146799.16	16668.05	< .0001	0.17
	어종	1	219922.62	24971.20	< .0001	0.25
	품사혼합 × 어종	1	40279.75	4573.78	< .0001	0.05
	품사순서 × 어종	2	58080.69	3297.29	< .0001	0.07
	오차	792	7000.79			
HSC	품사조합 순서	2	5.12	2.75	.265	0.004
	품사혼합 여부	1	54.64	58.65	< .0001	0.06
	어종	1	6.76	7.25	.126	0.01
	품사혼합 × 어종	1	20.23	21.85	.010	0.02
	품사순서 × 어종	2	60.30	32.54	< .0001	0.07
	오차	792	736.64			

어종 변수가 단어 생성률에 미치는 영향이 품사혼합 여부 및 품사조합 순서에 따라 어떻게 달라지는지를 살펴보기 위해 500회의 후속검정을 실시하고 통계량 평균을 산출했다. 그 결과, 대조용 표제어 목록이 HP 버전일 경우 한자 NN 조합은 한글 NN 조합보다 통계적으로 유의미하게 더 많은 단어를 생성해 낸 것으로 나타났다($t(99) = 82.94$, $p < .0001$, $r = 0.99$). 한자 NV 조합과 한글 NV 조합을 비교했을 때도 마찬가지였다($t(99) = 94.35$, $p < .0001$, $r = 0.99$). 한자 VV 조합 또한 한글 VV 조합보다 통계적으로 유의미하게 더 높은 단어 성생률을 나타냈다($t(99) = 92.24$, $p < .0001$, $r = 0.99$). 그리고 한자 VN 조합과 한글

VN 조합 간 비교에서도 유사한 패턴이 관찰되었다($t(99) = 89.74$, $p < .0001$, $r = 0.99$).

이어서 품사혼합 여부 및 품사조합 순서 변수가 단어 생성률에 미치는 영향이 어종에 따라 어떻게 달라지는지를 확인하기 위해 500회의 후속검정을 실시한 후 통계량 평균을 계산했다. 한글 음절끼리 무선 조합한 경우, 이종 품사 조합과 단일 품사 조합 조건의 단어 생성률에 통계적으로 유의미한 차이가 있었다(NV/VN vs. NN: $t(99) = -16.65/-11.36$, $p < .0001$, $r = 0.86/0.75$; NV/VN vs. VV: $t(99) = -4.60/-9.88$, $p = .024/< .0001$, $r = 0.41/0.70$). 이때 NN 조합은 두 가지 이종 품사 조합(NV/VN)보다 높은 단어 생성률을 나타낸 반면, VV 조합은 두 가지 이종 품사 조합보다 낮은 단어 생성률을 나타냈다.

아울러 한자 유래 음절끼리 무선 조합한 경우 또한 품사혼합 조합과 단일 품사 조합 조건 간에 통계적으로 유의미한 단어 생성률 차이가 관찰되었다(NV/VN vs. NN: $t(99) = -8.42/-6.24$, $p < .0001$, $r = 0.64/0.53$; NV/VN vs. VV: $t(99) = -3/96/-6.09$, $p = .084/.001$, $r = 0.37/0.52$). 이때 앞서 한글 음절끼리 무선 조합한 경우와 마찬가지로, NN 조합은 두 가지 이종 품사 조합(NV/VN)보다 통계적으로 유의미하게 생산성이 더 높았다. 그리고 VV 조합의 생산성은 NV/VN 조합보다 통계적으로 (거의) 유의미하게 낮았다. 어종 변수를 통제하고도 NN 조합이 이종 품사 조합보다 훨씬 단어 생성률이 높았다는 사실은 체언 음절의 높은 생산성을 시사한다. 그러나 단일 어종끼리 무선 조합한 경우 NN 음절쌍은 이종 품사 음절쌍보다 높은 단어 생성률을 나타냈지만, VV 음절쌍은 반대의 양상을 나타냈다. 따라서 단일 품사 조합보다 이종 품사 조합이 더 많은 단어를 만들어낼 것이라는 가정은 부분적으로 지지되었다. 그리고 NN 조합이 VV 조합보다 실제 단어를 더 많이 만들어낼 것이라는 가정은 완벽하게 지지되었다.

5. 전체 논의

5.1. 연구문제 1: 음절의 출현위치가 단어 생성률에 미치는 영향

본 연구에서는 두 개의 고빈도 음절 목록을 무작위로 조합하여 2음절 연쇄를 생성했을 때 음절이 유래한 단어 내 출현위치가 실제 한국어 단어가 만들어질 확률에 어떤 영향을 미치는지 실험 1을 통해 살펴보았다. 이때 음절이 유래한 단어 내 출현위치를 고려하여 음절쌍을 무선 조합했을 때, 출현위치를 고려하지 않고 음절쌍을 무선 조합했을 경우보다 실제 단어가 더 많이 만들어질 것으로 가정했다. 실험 결과, 위치고려 여부는 전체 사전 표제어 모집단(HP) 및 품사별 표제어 모집단(HPC)과 대조한 경우 단어 생성률에 통계적으로 유의미한 영향을 미치는 것으로 드러났다. 하지만 위치고려 여부는 품사혼합 여부 및 품사조합 순서와 통계적으로 유의미한 상호작용을 나타냈다. HP와 대조한 경우, 위치고려 NN/VN 조합은 위치 비고려 NN/VN 조합보다 통계적으로 유의미하게 더 많은 단어를 생성해냈다. 반면에 위치고려 VV/NV 조합의 경우, 위치 비고려 VV/NV 조합보다 통계적으로 유의미하게 더 적은 단어를 생성해냈다.

만약 모든 음절이 첫째 자리와 둘째 자리에 동등한 확률로 사용된다면, 위치고려 조건과 위치 비고려 조건에 사용되는 고빈도 음절 목록과 이들 음절의 빈도분포에 차이가 없을 것이다. 따라서 위치고려 조건과 위치 비고려 조건의 단어 생성률 또한 차이를 나타내지 않을 것이다. 그러나 실험 1의 결과는 그렇지 않았다. 적어도 체언 음절끼리 위치를 고려하여 무선 조합했을 때에는 위치를 고려하지 않고 무선 조합한 경우에 비해 실제 단어가 훨씬 더 많이 생성되었기 때문이다. 상술한 결과는 체언의 첫째 자리에 자주 쓰이는 음절과 둘째 자리에 자주 쓰이는 음절 간에 일정한 차이가 존재할 가능성을 암시한다.

실제로 이은하 · 남기춘(2020)에 따르면, 첫째 자리 유형 빈도 상위 100개 음절 목록 가운데 둘째 자리 유형빈도 상위 100개 음절 목록에 들지 않은 음절이 체언 표제어는 28개 그리고 용언 표제어는 35개나 된다. 예를 들면, 체언 표제어 첫째 자리에 자주 쓰인 ‘오’, ‘초’, ‘한’, ‘저’, ‘예’ 같은 음절은 상대적으로 둘째 자리에 잘 나오지 않았다. 반면에 ‘리’, ‘레’, ‘라’처럼 두음법칙상 첫째

자리에 분포의 제약이 있는 음절 또는 '음', '식', '적' 같은 생산성 높은 명사화 접미사 음절은 체언 표제어의 첫째 자리보다 둘째 자리에 더 자주 쓰였다. 게다가 체언 표제어 내 첫째 자리와 둘째 자리에 공통으로 쓰이는 음절들 또한 유형빈도 순위의 상관이 중간 수준(r = .67, p < .0001)에 그쳤다. 이는 체언 조어에 참여하는 첫 음절과 둘째 음절이 언어적으로 구별될 뿐만 아니라 빈도분포에서도 차이가 있음을 보여준다.

상술한 증거들은 음절이 유래한 단어 내 위치가 한국어 단어의 어휘성에 일정한 영향을 미칠 가능성을 짐작케 한다. 만약 한국어 모어화자가 표기음절의 위치별 분포에 대한 암시적 지식을 지니고 있다면, 한국어 (비)단어의 어휘성을 판단할 때 이를 활용할 것으로 가정할 수 있다. 이를 확인하기 위해 본 연구자는 한국어 모어화자를 대상으로 한 어휘판단과제 실험을 통해 음절의 단어 내 출현위치 변수가 반응시간에 미치는 영향을 살펴본 바 있다. 그 결과 단어 내 첫째 자리와 둘째 자리에 자주 나타나는 음절[44]을 차례로 무선 조합한 비단어에 대한 반응시간이, 단어 내 출현위치 구분 없이 고빈도 음절을 무선 조합한 비단어에 대한 반응시간보다 통계적으로 유의미하게 길었다(저자 미출간: $F(1, 128.47) = 6.45$, $p = .012$, $\omega^2 = 0.23$). 전자(위치고려 조건 비단어)의 반응시간이 후자(위치 비고려 조건 비단어)보다 길었다는 것은 전자가 후자보다 실제 단어처럼 보인 탓에 비단어로 판정하는 데 더 오랜 시간이 걸렸음을 의미한다. 이는 모어화자가 표기음절의 위치별 확률분포에 대한 암시적 지식을 지니고 있으며, 시각적으로 제시된 단어의 어휘성을 판단할 때 해당 지식을 활용할 가능성을 시사한다.

그러나 음절의 단어 내 출현위치 변수가 체언 음절끼리 무선 조합한 경우에만 연구가설을 지지하는 결과가 나왔다는 사실은 다시 한 번 생각해볼 필요가 있다. 실험 1의 결과에 따르면 위치고려 NN 조합은 위치 비고려 NN 조합보다 단어 생성률이 높았지만, 위치고려 VV 조합은 위치 비고려 VV 조합보다 단어 생성률이 낮았다. 이렇듯 품사별로 위치고려 변수가 단어 생성률에 미치는 영향이 서로 다르게 나타난 이유로는 품사별 어종 비율의 차이를 고려할 수 있다. 용언은 체언에 비해 고유어와 혼종어가 차지하는 비중이 매우 크다. 또한 고유

44) 세종 형태의미 분석 말뭉치에 100만 어절당 1,000회 이상 출현하는 고빈도 음절에 한한다.

어와 혼종어를 구성하는 한글 음절은 한자 음절에 비해 생산성이 크게 떨어진다(권인한 1997; 배성봉 · 이광오 · 박혜원 2012; 안소진 2009; 이광오 2003).

또한 위치 비고려 조합에서는 음절연쇄의 첫째 자리와 둘째 자리에 동일한 음절 세트가 사용되는 반면, 위치고려 조합에서는 음절연쇄의 첫째 자리와 둘째 자리에 서로 다른 음절 세트가 사용된다. 후자의 경우, 두 음절 세트에 포함된 한글 음절이 합쳐지면서 무선조합에 참여하는 한글 음절의 비중이 증폭되는 결과로 이어졌을 가능성이 높다. 실제로 위치 비고려 VV 조합용 첫 음절 세트(예를 들면, '따', '읽', '얻', '떠', '듣')와 둘째 음절 세트(예를 들면, '히', '끼', '렵', '랑', '꾸')에는 위치고려 VV 조합용 음절 세트에 없는 한글 음절이 다수 포함되어 있었다(이은하 2020).

이렇듯 비록 음절의 위치 변수가 단어 생성률에 미치는 영향은 통계적으로 유의미했지만, 음절이 유래한 품사 및 어종의 효과와 따로 떼어 논의하기 어렵다. 게다가 위치 변수의 효과 크기($\omega^2 = 0.004$) 또한 품사혼합 여부의 효과 크기($\omega^2 = 0.02$)에 비하면 5분의 1 수준에 지나지 않는다. 뿐만 아니라 대조용 목록의 규모가 작은 경우(예를 들면, HS나 HSC), 위치 변수의 주효과는 통계적으로 유의미하지 않았다. 반면에 품사혼합 여부의 주효과는 대조용 목록에 관계없이 항상 통계적으로 유의미했다. 이는 음절이 유래한 품사가 음절의 단어 내 위치보다 단어 생성률에 더 큰 영향을 미칠 가능성을 의미한다. 이에 따라 이어지는 논의에서는 무선 단어생성 시 음절이 유래한 어종 및 품사가 음절연쇄의 어휘성에 미치는 영향에 대해 심층적으로 고찰하고자 한다.

5.2. 연구문제 2: 음절이 유래한 어종이 단어 생성률에 미치는 영향

본 연구에서는 두 개의 고빈도 음절 목록을 무작위로 조합하여 2음절 연쇄를 생성했을 때 음절이 유래한 단어의 어종이 실제 한국어 단어가 만들어질 확률에 어떤 영향을 미치는지 실험 2를 통해 살펴보았다. 이때 동일한 어종에서 추출한 음절쌍을 무선 조합했을 때, 서로 다른 어종에서 추출한 음절쌍을 무선 조합했을 경우보다 실제 단어가 더 많이 생성될 것으로 가정되었다. 아울러 한자 음절끼리 무선 조합했을 때, 한글 음절끼리 무선 조합했을 경우보다 실제

단어가 더 많이 생성될 것으로 예상되었다. 실험 결과, 음절의 위치 변수와 품사 변수를 통제했을 때 어종 변수 — 음절쌍 무선조합 시 어종혼합 여부 및 어종조합 순서 — 는 HP, HPC 및 HS과 대조한 경우 단어 생성률에 통계적으로 유의미한 영향을 미치는 것으로 나타났다. HP와 대조한 경우, 체언 CC 조합은 체언 KC/CK 조합보다 단어 생성률이 높았던 반면, 체언 KK 조합은 체언 KC/CK 조합보다 단어 생성률이 낮았다. 이러한 경향은 용언 음절을 무선 조합한 경우에도 마찬가지였다.

실험 2의 결과는 단일 어종 음절 조합이 이종 어종 조합보다 더 많은 실제 단어를 만들어낼 것이라는 가정을 부분적으로 지지한다. 왜냐하면 한자 음절끼리 무선 조합했을 때에만 이종 어종끼리 무선 조합했을 경우보다 단어 생성률이 더 높았기 때문이다. 그러나 무선조합 음절연쇄를 소규모 표제어 표본(HS)과 대조했을 때에도 어종 변수의 주효과가 나타났다는 사실은 해당 변수가 단어 생성률에 미치는 영향이 충분히 강력함을 시사한다. 이러한 결과는 단어를 구성하는 음절의 언어적 특성과 확률적 분포가 어종에 따라 차이를 나타낸다는 기존 연구결과와도 일맥상통한다(예를 들면, 권인한 1997; 김유범 2016; 남성현 · 김선회 2018; 박나영 2015; 신지영 2009; 안소진 2009). 만약 한자어와 고유어에 주로 쓰이는 음절 간에 언어적 · 분포적 차이가 없다면, CC 조합과 KK 조합의 실제 단어 생성률은 유사했을 것이다. 하지만 CC 조합이 KK 조합보다 월등한 단어 생성률을 보였다는 사실은 한자어 생성에 주로 쓰이는 음절과 고유어 생성에 주로 쓰이는 음절 간에 언어적으로나 분포적으로나 차이가 존재함을 의미한다.

실제로 실험 2에 사용된 어종별 · 위치별 상위빈도 100개 음절 목록을 살펴보면, 체언 첫째 자리 한자 - 한글 음절 목록 간 중복음절 개수는 40개 그리고 체언 둘째 자리 한자 - 한글 음절 목록 간 중복음절 개수는 36개뿐이었다. 용언 음절의 경우, 첫째 자리 한자 - 한글 음절 목록 간 중복음절 개수는 34개 그리고 둘째 자리 한자 - 한글 음절 목록 간 중복음절 개수는 27개에 불과했다(이은하 2020). 이는 고유어 생성에 주로 사용되는 음절과 한자어 생성에 주로 사용되는 음절 간에 분포적 차이가 실재함을 보여주는 증거이다. 하지만 단일 어종끼리 결합된 KK 음절쌍이 이종 어종끼리 결합된 KC/CK 음절쌍보다 훨씬 낮은 단어

생성률을 나타낸 것은 흥미로운 결과이다. 개별 음절이 어종별로 고유한 언어적 · 분포적 특성을 지닌다면, 한글이든 한자든 상관없이 단일 어종 음절쌍이 실제 단어가 될 확률이 이종 어종 음절쌍이 실제 단어가 될 확률보다 커야 할 것이기 때문이다. 게다가 표준국어대사전에 등재된 전체 혼종어의 비율(20.46퍼센트)은 고유어(20.89퍼센트)보다도 작다(국립국어원 2020).

그럼에도 불구하고 단어 내 한자 음절의 비율이 높을수록 단어 생성률이 높은 이유는 한자 음절 자체의 높은 생산성으로밖에 설명하기 어렵다. 즉 음절 무선조합 시 단어 생성률에 결정적 영향을 미치는 요인은 단어를 구성하는 음절들이 유래한 어종 간 동질성보다 음절들이 유래한 어종 그 자체라는 것이다. 실제로 체언 음절의 경우, 실험에 사용된 상위빈도 한자 음절 100개 목록의 누적 표제어 출현 빈도를 한글 음절 100개 목록의 누적 표제어 출현 빈도와 비교하면 첫 음절이 무려 3.12배 그리고 둘째 음절이 2.71배에 달한다(이은하 2020). 이는 고빈도 한자 음절이 표제어 생성에 얼마나 생산적으로 참여하는지를 단적으로 보여주는 예이다.

또한 어종이 확인된 표준국어대사전 표제어 가운데 한자어가 차지하는 비율은 고유어의 2.54배 — 2음절 표제어는 3.14배 — 에 달한다(국립국어원 2020). 이는 자연히 한자 음절의 높은 생산성으로 이어지며, KK 음절쌍보다 CC 음절쌍에서 더 많은 실제 단어가 생성된 이유를 설명해준다. 아울러 사전 표제어 중 2음절 고유어의 수(n = 31,487)이 2음절 혼종어(n = 2,536)를 압도함에도 KK 조합보다 이종 어종 조합에서 단어 생성률이 더 높았던 것 또한 마찬가지 이유일 것이다. 개별 한자 음절은 하나의 형태소로서 다른 한자 또는 한글 음절과 활발하게 결합하지만, 한글 음절은 소수의 접사나 어근을 제외하고는 하나의 형태소로서 생산적으로 조어에 참여하는 모습을 찾아보기 어렵다. 이종 어종 조합이 그나마 KK 조합보다 생산성이 높았던 것도 생산성 높은 전형적 한자 음절과 한자 · 한글 공용 음절의 조합이 만들어낸 결과일 것이다. 이러한 양상은 음절이 유래한 어종이 한국어 단어 형성 시 음절 간 조합 확률을 결정하는 데 중요한 영향을 미칠 가능성을 시사한다.

뿐만 아니라 어종 변수와 품사 변수 간 상호작용이 HP, HS 및 HPC 버전에 걸쳐 두루 통계적으로 유의미했다는 점도 주목할 필요가 있다. 이는 무선 단어

생성에 사용된 음절이 어느 품사에서 왔는지에 따라 어종 변수가 단어 생성률에 미치는 영향의 양상이 달라질 수 있음을 의미한다. 하지만 음절이 유래한 품사의 종류에 관계없이 단어 생성률의 크기가 CC, KC/CK 그리고 KK 조합의 순서를 나타낸 것으로 미루어볼 때, 어종이 어휘성 결정에 미치는 영향은 품사에 의해 온전히 조절되는 것으로 보이지 않는다. 그럼에도 불구하고 체언의 69.12퍼센트가 한자어로 이루어진 반면 용언은 한자어가 존재하지 않는다는 사실을 고려한다면, 음절이 유래한 품사가 단어 생성률에 미치는 영향을 어종과 따로 떼어 살펴볼 필요가 있다. 이에 따라 이어지는 절에서는 무선 단어생성 시 음절이 유래한 품사가 어휘성 결정에 미치는 영향에 대해 심도 있게 검토하고자 한다.

5.4. 연구문제 3: 음절이 유래한 품사가 단어 생성률에 미치는 영향

본 연구에서는 두 개의 음절을 무작위로 조합하여 2음절 연쇄를 생성했을 때 음절이 유래한 단어의 품사가 실제 한국어 단어가 만들어질 확률에 어떤 영향을 미치는지 실험 3을 통해 살펴보았다. 이때 동일한 품사에서 추출한 음절쌍을 무선 조합했을 때, 서로 다른 품사에서 추출한 음절쌍을 무선 조합했을 경우보다 실제 단어가 더 많이 생성될 것으로 가정되었다. 아울러 체언 음절끼리 무선 조합했을 때, 용언 음절끼리 무선 조합했을 경우보다 실제 단어가 더 많이 생성될 것으로 예상되었다. 실험 결과, 음절의 위치 변수와 어종 변수를 통제했을 때 품사혼합 여부는 HP, HPC 및 HSC와 대조한 경우 단어 생성률에 통계적으로 유의미한 영향을 미치는 것으로 나타났다. 그리고 품사조합 순서의 주효과는 HP 및 HPC와 대조한 경우에만 통계적으로 유의미했다. HP와 대조한 경우, 한글 NN 조합은 한글 NV/VN 조합보다 단어 생성률이 높았던 반면, 한글 VV 조합은 한글 NV/VN 조합보다 단어 생성률이 낮았다. 아울러 한자 NN 조합 또한 한자 NV/VN 조합보다 단어 생성률이 높았던 반면, 한자 VV 조합은 한자 NV/VN 조합보다 단어 생성률이 낮았다.

실험 3의 결과는 같은 품사 음절끼리 무선 조합했을 때 다른 품사 음절끼리 무선 조합했을 경우보다 실제 단어가 더 많이 만들어질 것이라는 가정을 부분

적으로 지지한다. NN 조합과 반대로 VV 조합에서는 이종 품사 조합보다 실제 단어가 더 적게 생성되었기 때문이다. 그럼에도 불구하고 무선조합 음절연쇄를 소규모 표제어 표본(HSC)과 대조했을 때에도 품사혼합 여부의 주효과가 나타났다는 것은 품사 변수가 단어 생성률에 강력한 영향을 미치고 있음을 암시한다. 이는 체언이 용언에 비해 표제어 수 — 2음절 표제어의 경우 18.68배(음절수 구분 없는 경우 3.68배) — 가 더 많은 데 가장 큰 이유가 있을 것이다(국립국어원 2020). 뿐만 아니라 어종에 관계없이 체언과 용언의 세종 말뭉치 출현 빈도 상위 100개 음절 비교 시, 용언에만 주로 나타나는 음절이 첫째 음절(47개)과 둘째 음절(50개) 모두 절반에 달한다(이은하 2020).

이렇듯 음절연쇄를 구성하는 개별 음절의 품사적 동질성이 단어 생성률에 영향을 미친다는 사실은 체언과 용언 표제어 구성 음절의 음운론적 · 분포적 차이를 보고한 연구와도 궤를 같이한다(예를 들면, 이은하 · 남기춘 2020). 뿐만 아니라 무선조합 음절쌍 내에서 체언 유래 음절이 차지하는 비율이 높을수록 단어 생성률도 높았다. NN 조합은 NV/VN 조합보다 그리고 NV/VN 조합은 VV 조합보다 실제 단어를 더 많이 만들어냈기 때문이다. 만약 체언과 용언 표제어에 자주 사용되는 음절들이 대동소이하다면, NN 조합과 VV 조합 간에 단어 생성률의 차이는 나타나지 않았을 것이다. 이로 미루어볼 때, 체언 생성에 자주 사용되는 음절과 용언 생성에 자주 사용되는 음절 간에 언어적 · 분포적 차이가 실재하는 것으로 추론할 수 있다.

그러나 여기서 간과해선 안 될 사실이, 체언과 용언은 어종분포 면에서 큰 차이가 있다는 점이다. 여러 차례 지적했듯이, 품사와 어종은 분리하기 어려운 상호적 관계를 맺고 있다. 실험 3의 결과 또한 이를 뒷받침한다. 음절이 유래한 어종에 따라 품사혼합 여부의 효과 크기가 달랐기 때문이다. 품사혼합 여부의 효과 크기는 한글 음절끼리 무선 조합했을 때(NV vs. NN: r = 0.86; VN vs. NN: r = 0.75)보다 한자 음절끼리 무선 조합했을 때(NV vs. NN: r = 0.64; VN vs. NN: r = 0.53) 훨씬 작았다. 뿐만 아니라 한글 조건에서는 NV 및 VN 조합 모두 VV 조합보다 단어 생성률이 통계적으로 유의미하게 높았지만, 한자 조건에서는 VN 조합과 VV 조합 간 단어 생성률 차이가 통계적으로 유의미하지 않았다.

이와 같은 양상은 무선조합 음절연쇄의 어휘성을 결정하는 데 품사보다 어종이 더 큰 영향력을 행사할 가능성을 시사한다. 실험에 사용된 한자 음절의 체언 - 용언 간 중복도(첫 음절 75개, 둘째 음절 74개)는 한글 음절의 체언 - 용언 간 중복도(첫 음절 52개, 둘째 음절 52개)보다 훨씬 높다(이은하 2020). 품사에 따라 자주 쓰이는 한자 음절의 차이가 크지 않은 것은, 체언 내 한자어 비율(69.12퍼센트)이 높은 데다 용언 또한 혼종어 비율(64.65퍼센트)이 작지 않기 때문일 것이다(국립국어원 2020). 반면에 실험에 사용된 한글 음절의 경우, 용언 표제어에만 주로 나타나는 음절(예를 들면, '내', '잘', '되', '절', '꺼')의 비율이 절반(48개)에 가깝다(이은하 2020). 이는 용언 표제어 가운데 고유어가 차지하는 상당한 지분(35.35퍼센트)과 밀접한 관련이 있다(국립국어원 2020). 따라서 생산성이 높은 고빈도 한자 음절로 구성된 무선조합 음절연쇄는 개별 음절이 어느 품사에서 유래했건 실제 단어로 조회될 가능성이 한글 음절로 구성된 음절연쇄보다 높다. 결론적으로, 음절이 유래한 품사와 어종은 밀접한 상호관계를 맺고 있으며, 품사보다 어종이 한국어 단어 형성 시 음절 간 조합 확률에 더 큰 영향을 미치는 것으로 추론할 수 있다.

6. 결론

6.1. 주요 결과 요약

본 연구는 말뭉치 분석에 바탕을 둔 음절 무선조합 단어생성 실험을 통해 한국어의 어휘성에 영향을 미치는 표기음절 변수를 탐색함으로써, 모어화자가 지닌 한국어 음절 분포 및 음절 조합 패턴에 대한 직관의 실체를 규명하는 것을 목적으로 했다. 이에 대규모 말뭉치와 프로그래밍 기법에 바탕을 둔 세 가지 실험을 통해서, 음절이 유래한 단어의 품사, 음절이 유래한 단어의 어종 그리고 음절의 단어 내 출현위치에 따라 음절들을 무선 조합했을 때 실제 단어가 생성될 확률이 어떤 양상을 나타내는가를 살펴보았다.

세 가지 음절 무선조합 실험의 주요 결과는 다음과 같다. 첫째, 표제어 내 위치를 고려하여 추출한 음절끼리 무선 조합된 표본은 표제어 내 위치를 고려

하지 않고 추출한 음절끼리 무선 조합된 표본보다 실제 단어 생성률이 더 높았다. 둘째, 체언과 용언 간 어종비율의 차이를 고려하여 품사 변수의 효과를 배제하고도 어종 변수의 효과가 유효한지 여부를 확인한 결과, 단일 어종 음절끼리 무선 조합했을 때 이종 어종 음절 조합보다 더 많은 실제 단어를 만들어냈다. 아울러 한자 음절끼리 조합했을 때 한글 음절 조합보다 실제 단어를 더 많이 생성했다. 셋째, 체언과 용언 간 어종비율의 차이를 고려하여 어종 변수의 효과를 배제하고도 품사 변수의 효과가 유효한지 여부를 확인한 결과, 단일 품사 음절끼리 조합했을 때 이종 품사 음절 간의 조합보다 실제 단어를 더 많이 만들어냈다. 또한 체언 음절끼리 조합했을 때 용언 음절 조합보다 실제 단어를 더 많이 생성해냈다.

이러한 결과는 음절이 유래한 품사와 어종 그리고 음절의 출현위치가 한국어 단어 형성 시 음절 간 조합 확률에 유의미한 영향을 미치는 증거로 간주할 수 있다. 음절 간 확률적 조합과 단어형성 기제 간의 밀접한 관계는 음절이 하위어휘 단위로서 어휘처리 과정뿐만 아니라 어휘표상의 조직에도 중요한 역할을 담당할 가능성을 시사한다.

6.2. 연구의 의의 및 한계점

본 연구는 계량언어학적 방법론을 바탕으로 한국어 모어화자가 지닌 한국어 음절 분포 및 음절 조합 패턴에 대한 암시적 지식의 실체를 규명하고자 한 최초의 시도로서 의의가 있다. 아울러 실험 참여자를 필요로 하지 않으므로 시공간과 예산의 제약 없이 다양한 가설을 검증해볼 수 있다는 측면에서 인간 대상 실험의 대안으로서 높은 시의성을 지닌다.

그럼에도 불구하고 몇 가지 한계점을 지적하지 않을 수 없다. 첫째, 실험 1을 통해 확인된 위치고려 여부 변수의 효과에 의문의 여지가 있다. 위치를 고려하지 않고 음절을 추출하여 무선 조합하는 경우, 첫째 자리와 둘째 자리에 동일한 음절 세트가 투입된다. 반면에 위치를 고려하여 추출된 음절을 무선 조합하는 경우, 첫째 자리와 둘째 자리에 서로 다른 음절 세트가 사용된다. 만약 음절이 단어 내 위치의 구분 없이 동등한 조건으로 조어에 참여한다면, 단어 무선생성

시 훨씬 다양한 음절이 사용되는 위치고려 조건에서 실제 단어가 생성될 확률 — 표제어 목록 대조 시 실제 단어와 일치하는 표제어가 더 많이 조회될 확률 — 이 훨씬 높을 것으로 예상할 수 있다. 즉 음절이 유래한 단어 내 위치고려 여부보다 무선조합에 참여한 음절의 다양성이 단어 생성률에 더 큰 영향을 미쳤을 수 있다는 것이다.

둘째, 체언은 용언보다 표제어의 수가 현저히 더 많다. 체언 유래 음절로 조합된 음절연쇄는 용언 유래 음절로 조합된 음절연쇄에 비해 실제 단어로 조회될 확률 또한 자연히 더 높을 수밖에 없다. 즉 개별 품사 유래 음절이 지닌 고유한 특성보다 각 품사의 표제어 규모가 실제 단어 생성률에 실질적 영향을 미쳤을 가능성을 배제할 수 없다는 것이다. 이러한 한계를 보완하고자 본 연구에서는 대조용 목록의 표제어 수를 제한하거나(예를 들면, HS나 HSC의 사용), 용언 음절 무선조합 시 음절의 중복사용을 허용하는 방법을 채택했다. 하지만 이들 방법 또한 체언과 용언 표제어 간 양적 격차를 완벽하게 극복하기는 어렵다. 후속연구에서는 대조용 목록의 체언과 용언 표제어 수를 같은 비율로 맞추거나, 단어생성에 참여하는 자극의 품사를 체언으로 한정하는 등의 다양한 대안을 강구할 필요가 있다.

셋째, 두 분 심사위원의 지적대로 본 연구에 대조용 사전으로 사용된 표준국어대사전이 한국어 모어언중의 실제 언어행태를 핍진하게 반영하고 있는지 의문스럽다. 특히 해당 사전의 한자어 표제어 점유율은 50퍼센트 이상인데, 과연 한국어 모어언중이 일상생활에서 사용하는 어휘의 유형 빈도 또한 유사한 패턴을 따를지 재고의 여지가 있다. 따라서 기존 사전에는 실리지 않았지만 실제로 사용되고 있는 다양한 신어와 전문용어까지 두루 수록된 국립국어원의 우리말샘 같은 다양한 어휘자원을 활용하는 것도 좋은 대안이 될 수 있다.

마지막으로, 본 연구에서는 대규모 말뭉치에 기반을 둔 계량언어학적 방법론을 사용하여 한국어 단어의 어휘성 결정에 영향을 미치는 음절 변수를 탐색하고자 했다. 그러나 이는 어디까지나 모어화자가 생산한 언어사용 자료를 통한 귀납적 추론에 불과하므로, 본 연구를 통해 확인된 음절 변수가 실제로 모어화자의 어휘성 판단에 영향을 미치는지를 실증적으로 확인할 필요가 있다. 현재 후속연구 차원에서 한국어 모어화자 대상 어휘판단과제 실험연구가 진행

중이다. 이를 통해 음절이 어휘처리 과정과 어휘표상 조직에서 어떤 역할을 수행하는지에 대한 의미 있는 단서를 얻을 수 있기를 기대한다.

참고문헌

강범모 · 김흥규(2009), 한국어 사용 빈도, 서울: 한국문화사.

국립국어원(2008), 표준국어대사전, 서울: 국립국어원.

국립국어원(2020), 표준국어대사전 통계, 월드와이드웹: https://stdict.korean.go.kr/statistic/dicStat.do#static_menu3_1에서 2020년 4월 20일에 검색했음.

권유안(2012), "첫 음절 토큰 빈도와 타입 빈도가 단어 및 유사 단어 어휘 판단 시간에 미치는 영향", 한국심리학회지: 인지 및 생물 24(4), 315-328.

권유안 · 남기춘(2011), "한글 음절 이웃 효과에서 한자어 형태소의 영향 :표기 및 음운 이웃과 한자어 이웃과의 관련성 중심으로", 한국심리학회지: 인지 및 생물 23(3), 한국심리학회, 301-319.

권유안 · 이윤형(2014), "한국어 시각 단어 처리 관련 변인을 반영하는 사건관련 뇌파 파형", *Journal of the Korean Data Analysis Society* 16(3), 한국자료분석학회, 1527-1539.

권유안 · 이윤형(2015), "시각 단어 재인시 나타나는 음절 빈도효과의 원인 규명: 사건관련 뇌전위 연구", 언어과학 22(4), 한국언어과학회, 1-17.

권인한(1997), "현대국어 한자어의 음운론적 고찰", 국어학 29, 국어학회, 243-260.

김미란 · 최재웅 · 홍정하(2014), "한국어 초성-중성 결합의 분포적 특성 및 모음의 군집분석 연구", 음성음운형태론연구 20(1), 한국음운론학회, 23-49.

김유범(2016), "현대국어 한자음의 음운론과 형태론", 우리말연구 44, 우리말학회, 5-26.

김제홍 · 이창환 · 남기춘(2018), "한국어 명사 어절 재인에서 나타나는 음절교환 효과", 한국심리학회지: 인지 및 생물 30(3), 한국심리학회, 261-268.

남성현 · 김선회(2018), "한국어 자음-모음 연쇄의 어휘계층 간 비교", 언어 43(3), 한국언어학회, 485-506.

박나영(2014), "한국어 명사의 음소배열제약에 대한 기계학습", 음성음운형태론연구 20(3), 한국음운론학회, 297-322.

박나영(2015), "고유어와 한자어의 비교 음소배열제약", 한국어학 67, 한국어학회, 147-192.

배성봉 · 이광오 · 박혜원(2012), "한자어 인지와 학습에서 의미투명성의 효과", 교육심리연구 26(2), 한국교육심리학회, 607-620.

신지영(2009), “한국 한자음의 빈도 관련 정보 및 음절 구조 제약”, 말소리와 음성과학 1(2), 한국음성학회, 129-140.

신지영(2010), “한국어 사전 표제어 발음의 음소 및 음절 빈도”, *Communication Sciences and Disorders* 15(1), 한국언어청각임상학회, 94-106.

신하선 · 남기춘(2019), “한글 시각 어절 재인에서 첫 음절빈도와 음절 이웃 빈도 효과: 사건관련 뇌전위(ERP) 연구”, 한국인지및생물심리학회 2019년 학술대회 발표논문, 평창.

안소진(2009), “한자어 구성 음절의 특징에 대하여: 고빈도 2음절 한자어를 대상으로”, 형태론 11, 형태론, 43-59.

이광오(2003), “단어인지 수행은 어종에 따라 다를까?”, 한국심리학회지: 인지 및 생물 15(4), 한국심리학회, 479-498.

이광오 · 배성봉(2009a), “한국어 고유어의 인지에서 형태소 처리”, 한국심리학회지: 인지 및 생물 21(3), 한국심리학회, 233-247.

이광오 · 배성봉(2009b), “한국어 음절의 표기빈도와 형태소빈도가 단어인지에 미치는 효과”, 인지과학 20(3), 한국인지과학회, 309-333.

이광오 · 정진갑 · 배성봉(2007), “표기 체계와 시각적 단어 인지 :한자어의 인지에서 형태소의 표상과 처리”, 한국심리학회지: 인지 및 생물 19(4), 한국심리학회, 313-327.

이용은(2016), “분절음 연쇄의 분포 정보와 한국어 음절하위 구성소”, 언어연구 33(스페셜), 경희대학교 언어정보연구소, 187-208.

이은하(2020), 위치별, 품사별, 어종별 표제어 출현 빈도 상위 100개 음절 목록, 월드와이드웹: https://github.com/cognitivepsychology/cognitive_psychology에서 2020년 5월 20일에 검색했음.

이은하 · 남기춘(2020), “세종 말뭉치에 나타난 한국어 음절의 빈도와 분포”, 언어과학연구 92, 언어과학회, 79-130.

이창환 · 김제홍(2018), “한국어 다음절 단어재인에 있어서 글자교환 효과”, 언어과학연구 86, 언어과학회, 339-352.

태진이 · 남예은 · 이윤형 · 김태훈(2015), “한국어 시각단어재인에서 음절과 음절체의 역할”, 언어과학연구 73, 언어과학회, 205-224.

Álvarez, C. J., Carreiras, M., & de Vega, M.(2000), “Syllables and Morphemes: Contrasting Frequency Effects in Spanish”, *Psicológica* 21(2), 341-374.

Álvarez, C. J., Carreiras, M., & Taft, M.(2001), “Frequency and Neighborhood Effects on Lexical Access: Lexical Similarity or Orthographic Redundancy?”, *Journal of Experimental Psychology: Learning, Memory, and Cognition* 27(2), 545-555.

Andrews, S.(1992), “Frequency and Neighborhood Effects on Lexical Access: Lexical Similarity or Orthographic Redundancy?”, *Journal of Experimental*

Psychology: Learning, Memory, and Cognition 18(2), 234-254.

Balota, D. A., & Chumbley, J. I.(1984), "Are Lexical Decisions a Good Measure of Lexical Access? the Role of Word Frequency in the Neglected Decision Stage", *Journal of Experimental Psychology: Human Perception and Performance* 10(3), 340-357.

Carlisle, J. F., & Stone, C. A.(2005), "Exploring the Role of Morphemes in Word Reading", *Reading Research Quarterly* 40(4), 428-449.

Chetail, F., Colin, C., & Content, A.(2012), "Electrophysiological Markers of Syllable Frequency during Written Word Recognition in French", *Neuropsychologia* 50(14), 3429-3439.

Conrad, M., & Jacobs, A.(2004), "Replicating Syllable Frequency Effects in Spanish in German: One More Challenge to Computational Models of Visual Word Recognition", *Language and Cognitive Processes* 19(3), 369-390.

Conrad, M., Carreiras, M., Tamm, S., & Jacobs, A. M.(2009), "Syllables and Bigrams: Orthographic Redundancy and Syllabic Units Affect Visual Word Recognition at Different Processing Levels", *Journal of Experimental Psychology: Human Perception and Performance* 35(2), 461-479.

Conrad, M., Stenneken, P., & Jacobs, A. M.(2006), "Associated or Dissociated Effects of Syllable Frequency in Lexical Decision and Naming", *Psychonomic Bulletin & Review* 13(2), 339-345.

Doignon-Camus, N., Bonnefond, A., Touzalin-Chretien, P., & Dufour, A.(2009), "Early Perception of Written Syllables in French: An Event-Related Potential Study", *Brain and Language* 111(1), 55-60.

Field, A.(2010), *Discovering Statistics using SPSS* (3rd ed.), London: Sage.

Forster, K. I.(1989), "Basic Issues in Lexical Processing", in William Marslen-Wilson ed., *Lexical Representation and Process*, MIT Press, MA: Cambridge, 75-107.

Grosjean, P., & Ibanez, F.(2014), *pastecs: Package for Analysis of Space-Time Ecological Series (R Package Version 1.3-18)*, Retrieved April 20, 2020, from the Word Wide Web: https://CRAN.R-project.org/package=pastecs/

Harm, M. W., & Seidenberg, M. S.(2004), "Computing the Meanings of Words in Reading: Cooperative Division of Labor between Visual and Phonological Processes", *Psychological Review* 111(3), 662-720.

Kim, H., & Na, D. L.(2000), "Dissociation of Pure Korean Words and Chinese-Derivative Words in Phonological Dysgraphia", *Brain and Language* 74(1), 134-137.

Ludecke, D.(2020), *sjstats: Statistical Functions for Regression Models (R*

Package Version 0.17.9), Retrieved April 20, 2020, from the Word Wide Web: https://CRAN.R-project.org/package=sjstats/

Macizo, P., & Van Petten, C.(2007), "Syllable Frequency in Lexical Decision and Naming of English Words", *Reading and Writing* 20(4), 295-331.

McClelland, J. L., & Rumelhart, D. E.(1981), "An Interactive Activation Model of Context Effects in Letter Perception: I. an Account of Basic Findings", *Psychological Review* 88(5), 375-407.

Newman, M. E. J.(2005), "Power Laws, Pareto Distributions and Zipf's Law", *Contemporary Physics* 46(5), 323-351.

Perea, M., & Carreiras, M.(1998), "Effects of Syllable Frequency and Syllable Neighborhood Frequency in Visual Word Recognition", *Journal of Experimental Psychology: Human Perception and Performance* 24(1), 134-144.

Plaut, D. C.(1997), "Structure and Function in the Lexical System: Insights from Distributed Models of Word Reading and Lexical Decision", *Language and Cognitive Processes* 12(5-6), 765-806.

R Core Team(2018), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing.

Robinson, D., & Hayes, A.(2020), *broom: Convert Statistical Analysis Objects into Tidy Tibbles (R Package Version 0.5.5)*, Retrieved April 20, 2020, from the Word Wide Web: https://CRAN.R-project.org/package=broom/

Rumelhart, D. E., & McClelland, J. L.(1982), "An Interactive Activation Model of Context Effects in Letter Perception: II. the Contextual Enhancement Effect and some Tests and Extensions of the Model", *Psychological Review* 89(1), 60-94.

Schiller, N. O(2004), "The Onset Effect in Word Naming", *Journal of Memory and Language* 50(4), 477-490.

Schiller, N. O.(1999), "Masked Syllable Priming of English Nouns", *Brain and Language* 68, 300-305.

Seidenberg, M. S.(1987), "Sublexical Structures in Visual Word Recognition: Access Units or Orthographic Redundancy?", in Max Coltheart ed., *Attention and Performance 12: The Psychology of Reading*, Hillsdale, NJ: Lawrence Erlbaum Associates, 245-263.

Seidenberg, M. S., & McClelland, J. L.(1989), "A Distributed, Developmental Model of Word Recognition and Naming", *Psychological Review* 96(4), 523-568.

Spoehr, K. T., & Smith, E. E.(1973), "The Role of Syllables in Perceptual Processing", *Cognitive Psychology* 5(1), 71-89.

Taft, M.(1979), “Lexical Access Via an Orthographic Code: The Basic Orthographic Syllabic Structure (BOSS)”, *Journal of Memory and Language* 18(1), 21-39.

Taft, M.(1987), “Morphographic Processing: The BOSS Re-Emerges”, in Max Coltheart ed., *Attention and Performance XII: The Psychology of Reading*, Hove, UK: Lawrence Erlbaum Associates, 265-280.

Taft, M., & Forster, K. I.(1975), “Lexical Storage and Retrieval of Prefixed Words”, *Journal of Verbal Learning and Verbal Behavior* 14(6), 638-647.

Taft, M., & Forster, K. I.(1976), “Lexical Storage and Retrieval of Polymorphemic and Polysyllabic Words”, *Journal of Memory and Language* 15(6), 607-620.

Wickham et al.(2019), “Welcome to the tidyverse”, *Journal of Open Source Software* 4(43), 1686.

Zipf, G. K.(1935), *The Psychobiology of Language*, Boston: Houghton Mifflin.

이은하(제1 저자)
고려대학교 지혜과학연구센터(연구교수)
서울 성북구 안암로 145 구 법학관 208호
02841
전화 번호 : 02-3290-2548
전자 우편 : seekermonth@korea.ac.kr

남기춘(교신 저자)
고려대학교 문과대학 심리학과(교수) · 고려대학교 지혜과학연구센터(소장)
서울 성북구 안암로 145 구 법학관 405호
02841
전화 번호 : 02-3290-2068
전자 우편 : kichun@korea.ac.kr

원고 접수일 : 2020. 05. 20.
원고 수정일 : 2020. 07. 24.
게재 확정일 : 2020. 09. 18.