韓国語の話しことばと書きことばにおける音素，音節，音節結合の出現頻度

李在鎬（情報通信研究機構）
玉岡賀津雄（麗澤大学）
林炫情（山口県立大学）

1. 背景と目的

ある言語の音素，音節，音節結合(bi-syllable)の出現頻度は，その言語の母語話者およびその言語を外国語として学ぶ学習者の音韻処理や音韻習得のメカニズムを解明するための基礎資料として重要である。たとえば，日本語の音韻処理の研究では，重音節(CV+特殊音)と軽音節(CV)の音韻処理の速度(命名潜時)を測定した結果，CV と CVN の両ユニットがほぼ同じ速度で発音に達すること，さらに重音節が音韻産出のユニットになっている(Tamaoka & Terao, 2004)ことを示した。しかし，この実験では，モーラ(軽音節)およびモーラ結合の頻度が考慮されていなかったため，Tamaoka & Makioka（2008 in press）では，さらに，Tamaoka & Makioka (2004)データ[1]を使って，これらの頻度を統制した類似の実験を行い，先行研究の結果をより信頼性の高いものにしている。このように，音韻的なユニットの頻度は，心理言語学的なアプローチによる音韻処理や音韻習得研究に欠かせない基礎データである。

一方，韓国語における音素，音節，音節結合の出現頻度に関する計量的調査には，유재원(1993)や강범모(1995, 1997)などがある。しかし，これらの先行研究は工学的な利用に特化したものであることや，聖書等の書きことばのみを対象にしていることから，韓国語母語話者および韓国語を外国語として学ぶ学習者の音韻処理や音韻習得のメカニズムを解明するための基礎データとして用いるには偏りがある。そこで，本研究では「21 世紀世宗企画による Malmungchi（以下，世宗コーパスと記す）」の書きことばと話しことばのデータ集より 892,884 字の文字列を無作為に抽出し，音素，音節，音節結合のそれぞれの出現頻度を算出した。そして，これらの比較的小規模の話しことばと書きことばの頻度データの相関をみることで，音素，音節，音節結合の頻度データの普遍性を検討する。その結果を踏まえて，韓国語における音韻処理や音韻習得の今後の研究のために，音素，音節，音節結合の頻度と千分率(1/1000)を提供することが本研究の目的である。

2. 韓国語の音韻構造

韓国語の音節構造の特徴は以下のようにまとめることができる(野間 2007:273)。

1. 母音のみが核（nucleus）を形成しうる。
2. 子音が単独で音節を形成することはなく，常に音節副音(nonsyllabic)となる。

基本構造は(1)のように捉えることができる(Sohn, 1994:445)。

(1) ./(C)(s)V(:)(C)/.

[1] Tamaoka & Makioka (2004)では，天野・近藤（2000）の朝日新聞の語彙コーパスをもとに，日本語の音素，モーラ，音節を単位とした頻度を算出している。

(1)における「.」は音節境界,「C」は子音,「s」は半母音,「V」は母音,「:」は母音長[2], ( )はオプションであることを表す。語頭の C を初声(initial)と呼び, 語中の V を中声(medial), 語末の C を終声(final)と呼ぶ。(1)が示す通り, 韓国語では核となる V を除くいずれの要素も必須要素ではない。また, 音節言語としての特徴をもつ韓国語は, 開音節(open syllable)と閉音節(closed syllable)を共に有する。ただし, 初声と終声には, 一つの子音のみが立ち, 子音連続(cluster)はない。韓国語の音節の基本パターンは表 1 のように表すことができる(野間, 2007:273)。また, 表1のそれぞれの組み合わせを頻度別に示したのが表2である。

表1 韓国語の音節構造

初声 initial	中声 medial		終声 Final
(C)	(s)	V	(C)
子音 consonant	半母音 semivowel	単母音 monophthong	子音 consonant
onset		音節主音	Coda
H	W	a	L

表2 音節のタイプ頻度[3]

		音節タイプ			
		開音節		閉音節	
音素の位置	音素の数	φV	CV	φVC	CVC
初声(C)	19	21	399	588	11,172
中声(V)	21				
終声(C)	28				

3. データと調査方法

ハングル字母は韓国語の音素や音韻構造を直接反映しているため, これを基本単位として用いた(cf. 강범모 1997, 2003)。データ収集は以下の手順で行った。

第 1 に, 世宗コーパスの書きことばと話しことばのデータ収録一覧表からそれぞれ 10 ファイルずつランダムサンプリングで選択した。第 2 に, 正規表現を使用し, 英数字や記号などを取り除き, ハングルのみを残した。第 1 と第 2 の結果, 書きことばコーパスから 629,039 字(延べ数), 話しことばコーパスから263,845 字(延べ数), 合計 892,884 字(延べ数)のハングルデータを抽出することができた。第 3 に, このハングルデータを使って, 一文字単位で改行を挿入するとともに, Perl 5.10 で子母を分割するスクリプト処理を行い, 各々の音節を音素単位に分割させた。そのあと, 第 4 に, 書きことばデータと話しことばデータそれぞれについて, 音素, 音節, 音節結合の各レベルの頻度を計算した。第 5 に, 書きことばと話しことばの頻度の相関関係を調べるため, ピアソンの相関係数を算出した。そして, 両コーパスの音韻的な単位の使用頻度の一貫性を証明した上で, 第6に, 千分率を計算した(第 4 の手続きは相関係数を算出するためのもの)。

[2] 韓国語の母音長, つまり長母音は第 2 音節以下では短母音に変わってしまうため, 音長が弁別力をもっているのは第1音節においてのみとなる。また, 近年のソウル方言のとりわけ若い世代では第1音節でさえ, 音調の区別が不明瞭になる傾向をみせており, 音長の音素としての地位が揺れている(李翊燮・李相億・蔡琬, 2004; 趙義成・呉文淑 2004)。そのことから, 本研究では長母音か短母音かの区別は取り上げない。

[3] 終声として可能な文字は 27 であるが, 開音節の終声(φ)も数に含めたため 28 になる。よって CVC として理論的に可能な組み合わせは 19×21×28 になることから, 11,172 という値になっている。

4. 書きことばと話しことばの相関関係と頻度データの信頼性

相関係数は，初声が 0.998，中声が 0.995，終声が 0.995 と，完全な相関である 1.00 に極めて近い値となった。また，音節レベルでは φV が 0.987，CV が 0.985，φVC が 0.987，CVC が 0.978 と，こちらもきわめて高い相関係数であった。さらに，2 種類の音節結合の場合も，母音のみの音節結合が 0.990，全体の音節結合が 0.886 と，音素，音節同様にきわめて高い相関係数が得られた。つまり，たとえ比較的小さなコーパスであっても，書きことばと話しことばで，音韻的な単位の頻度が非常に類似していることを示しており，頻度情報がこの規模のコーパスで十分に信頼に足ることを示している。したがって，以下では，音素，音節，音節結合の出現頻度別にその結果を報告する。なお，表ではこれらの頻度を研究指標として使い易いように千分率で示した。

5. 調査結果

5.1 韓国語音素の出現頻度

韓国語の音素の出現頻度は表3に示したとおりである。表3から分かるように，音素レベルでは，初声は「ㅇ，ㄱ，ㄷ」，中声は「ㅏ，ㅣ，ㅡ」，終声は「ㄴ，ㄹ，ㅇ」の順で高い頻度を示した。本研究の結果は，강범모(1997) が 김홍규，강범모(1996)のデータを用いて行った調査の結果とおおむね一致する。しかし，初声の「ㅋ，ㅆ，ㅉ，ㅃ」，中声の「ㅙ，ㅒ，ㅞ」はその出現頻度がきわめて低かった。また，終声の特徴としては，ㅂを境にして，よく使われる音素とあまり使われない音素で二極化していること，また겹받침(二文字の終声)も出現頻度としてはかなり低いことが分かった。とりわけ終声の「ㅥ，ㄻ，ㅋ，ㄿ，ㅀ」の出現頻度が極めて低く，限定された単語でのみ用いられていることがうかがえる。

表３ 音素のコーパス別出現頻度

位置	音素	出現頻度 書きことば	話しことば	全体	千分率
初声	ㅇ	149424	61509	210933	236.2
	ㄱ	85226	36211	121437	136.0
	ㄷ	56432	24403	80835	90.5
	ㅅ	54801	21071	75872	85.0
	ㅈ	51957	23020	74977	84.0
	ㅎ	46927	19882	66809	74.8
	ㄴ	43246	18212	61458	68.8
	ㄹ	44839	16600	61439	68.8
	ㅁ	32220	12956	45176	50.6
	ㅂ	23342	11126	34468	38.6
	ㅊ	11564	6248	17812	19.9
	ㅌ	6641	3074	9715	10.9
	ㅍ	6495	2423	8918	10.0
	ㄲ	5364	1950	7314	8.2
	ㅋ	3951	1855	5806	6.5
	ㅆ	2835	1295	4130	4.6
	ㅉ	1886	973	2859	3.2
	ㅃ	1061	572	1633	1.8
	合計	828	465	1293	1.4
		629039	263845	892884	
中声	ㅏ	134209	57614	191823	214.8
	ㅣ	97909	40399	138308	154.9
	ㅡ	83163	30271	113434	127.0
	ㅓ	68777	27185	95962	107.5
	ㅗ	59083	25676	84759	94.9
	ㅜ	39788	19666	59454	66.6
	ㅔ	28949	11966	40915	45.8
	ㅐ	28513	12898	41411	46.4
	ㅋ	27681	12430	40111	44.9
	ㅢ	14744	4818	19562	21.9
	ㅘ	11818	5074	16892	18.9
	ㅚ	7405	3508	10913	12.2
	ㅛ	8330	2829	11159	12.5
	ㅑ	4311	2378	6689	7.5
	ㅝ	3240	2129	5369	6.0
	ㅖ	3353	1432	4785	5.4
	ㅠ	3127	1337	4464	5.0
	ㅟ	3360	1236	4596	5.1
	ㅙ	820	716	1536	1.7
	ㅒ	301	199	500	0.6
	ㅞ	158	84	242	0.3
	合計	629039	263845	892884	
終声	ㄴ	95908	38680	134588	332.8
	ㄹ	53016	22028	75044	185.5
	ㅇ	38524	19825	58349	144.3
	ㄱ	32158	12493	44651	110.4
	ㅁ	16795	7781	24576	60.8
	ㅆ	13977	6468	20445	50.5
	ㅅ	9982	3271	13253	32.8
	ㅂ	8516	4750	13266	32.8
	ㄵ	2287	994	3281	8.1
	ㅎ	2300	857	3157	7.8
	ㅄ	1842	921	2763	6.8
	ㅌ	1698	724	2422	6.0
	ㄷ	1287	599	1886	4.7
	ㅈ	1036	540	1576	3.9
	ㅍ	976	469	1445	3.6
	ㄺ	609	231	840	2.1
	ㅊ	562	237	799	2.0
	ㄲ	432	189	621	1.5
	ㄽ	352	84	436	1.1
	ㄾ	247	129	376	0.9
	ㄶ	147	71	218	0.5
	ㄼ	137	48	185	0.5
	ㄳ	74	6	80	0.2
	ㄻ	53	10	63	0.2
	ㅋ	54	7	61	0.2
	ㄿ	35	3	38	0.1
	ㅀ	43	4	47	0.1
	合計	283047	121419	404466	

5.2 韓国語音節の出現頻度

本研究でその使用が確認された音節のタイプ頻度は表 4 に示したとおりである。表 2 で示した理論上可能な音節のタイプ頻度と本研究で明らかになった実際使用されるタイプ頻度とでは大きな隔たりがあることが明らかになった。つまり，理論上では組み合わせが可能であっても実際は使用されない文字が多く存在していることが推測される。上位 10 までの音節の出現頻度は表 5 のとおりである。φV は「이, 의, 에」, CV は「다, 고, 가」, φVC は「을, 은, 있」, CVC は「는, 한, 것」の順で出現頻度が高くなっている。

表4 コーパス別出現頻度

	コーパス別の出現頻度		
音節	書きことば	話しことば	全体
φV	21	21	21
CV	285	274	294
CVC	1828	1374	1948
φVC	157	139	164

表 5 上位 10 位までの音節のコーパス別出現頻度[4]

区分		上位 10 音節	出現頻度			
			書きことば	話しことば	全体	千分率
開音節	φV	ㅇㅣ	24626	8960	33586	273.0
		ㅇㅢ	14434	4248	18682	151.8
		ㅇㅔ	12835	4463	17298	140.6
		ㅇㅓ	6803	3111	9914	80.6
		ㅇㅏ	5242	2625	7867	63.9
		ㅇㅡ	4780	2112	6892	56.0
		ㅇㅜ	2920	1336	4256	34.6
		ㅇㅛ	2579	1289	3868	31.4
		ㅇㅕ	2743	993	3736	30.4
		ㅇㅑ	1789	1177	2966	24.1
	CV	ㄷㅏ	18067	7421	25488	52.2
		ㄱㅗ	10724	4579	15303	31.3
		ㄱㅏ	10684	4525	15209	31.1
		ㅎㅏ	11263	3820	15083	30.9
		ㅈㅣ	9459	4069	13528	27.7
		ㅅㅓ	8807	2996	11803	24.2
		ㄹㅗ	8953	2817	11770	24.1
		ㄱㅡ	8759	2439	11198	22.9
		ㄱㅣ	7036	2545	9581	19.6
		ㄴㅏ	6343	2632	8975	18.4
閉音節	φVC	ㅇㅡㄹ	11351	4406	15757	74.7
		ㅇㅡㄴ	7205	2967	10172	48.3
		ㅇㅣㅆ	6321	2020	8341	39.6
		ㅇㅣㄴ	4890	1760	6650	31.5
		ㅇㅣㄹ	3000	1693	4693	22.3
		ㅇㅓㅆ	2084	692	2776	13.2
		ㅇㅓㅂ	1749	900	2649	12.6
		ㅇㅝㄴ	1543	1017	2560	12.1
		ㅇㅏㄴ	1167	962	2129	10.1
		ㅇㅣㅂ	1313	672	1985	9.4
	CVC	ㄴㅡㄴ	18626	5934	24560	27.9
		ㅎㅏㄴ	8361	3557	11918	13.5
		ㄱㅓㅅ	6481	1732	8213	9.3
		ㄹㅡㄹ	5754	2202	7956	9.0
		ㄷㅡㄹ	5048	1920	6968	7.9

[4] CV および CVC では φV および φVC と重複した事例は表示していないが，千分率を計算する際の母集団には含まれている。

	ㅈㅓㄱ	5104	1111	6215	7.0
	ㅈㅓㅇ	3056	2118	5174	5.9
	ㅁㅕㄴ	3788	1044	4832	5.5
	ㅁㅏㄴ	3333	1309	4642	5.3
	ㅁㅜㄴ	3434	1032	4466	5.1

5.3 韓国語音節結語の出現頻度

まず，母音同士の音節結合のタイプ頻度は21の二乗なので441となるが，実際は435パターンのみが確認された。「ㅖ+ㅏ，ㅖ+ㅐ，ㅖ+ㅓ，ㅠ+ㅖ，ㅚ+ㅖ，ㅝ+ㅖ,」の6つのパターンは見られなかった。また，音節全体の結合の場合，理論的には11,172の二乗として124,813,584のパターンが可能である。しかし，実際の使用においては，81,791 パターンのみが確認され，理論と使用とではかなりの隔たりがあることが分かった。

本研究で確認された母音のみの結合頻度とCVとCVC，φVCの音節結合頻度を上位10まで示したのが表6である。

表６　音節結合の出現頻度

音節結合	上位10 音節結合	出現頻度		
		書きことば	話しことば	全体
母音のみ	ㅏ+ㅏ	28648	11868	40516
	ㅣ+ㅏ	27184	11426	38610
	ㅏ+ㅡ	23791	8138	31929
	ㅏ+ㅣ	22314	9468	31782
	ㅡ+ㅏ	15226	5768	20994
	ㅗ+ㅏ	13980	5889	19869
	ㅏ+ㅗ	12822	6262	19084
	ㅓ+ㅣ	13831	5201	19032
	ㅣ+ㅡ	13916	5084	19000
	ㅡ+ㅓ	13863	4494	18357
全体	ㅇㅡ+ㄹㅗ	3683	1341	5024
	ㅇㅐ+ㅅㅓ	3547	1082	4629
	ㅎㅏ+ㄴㅡㄴ	2780	645	3425
	ㄴㅣ+ㄷㅏ	2309	1063	3372
	ㄱㅓㅅ+ㅇㅣ	2513	681	3194
	ㄴㅡㄴ+ㄱㅓㅅ	2568	454	3022
	ㅇㅣ+ㄷㅏ	2247	669	2916
	ㄱㅗ+ㅇㅣㅆ	1892	750	2642
	ㄷㅏ+ㄱㅡ	1993	521	2514
	ㅎㅏ+ㄱㅗ	1848	626	2474

6. 結論

本研究では，世宗コーパスを使って，韓国語の音素，音節，音節結合の出現頻度を調べた。その結果，ランダムに抽出した書きことばと話しことばのコーパスにおいて，音素，音節，音節結合の全てにおいて相関係数が極めて高いことが明らかになった。世宗コーパスから抽出した書きことばと話しことばのデータは，書かれた文字と話された内容を文字化したという違いばかりでなく，トピックそのものが異なっている。それにも関わらず音素，音節，音節結合の各頻度において，書きことばと話しことばの相関がきわめて高かったことは，両コーパスの音韻的な単位における頻度の類似

性が高いことを示しており，さらに，本研究で扱った規模のコーパスのサイズで信頼できる指標となりうることを意味している。したがって，本研究で提供する頻度および千分率を指標として，音韻的処理や習得の研究が展開できることになる。

〈引用文献〉

天野成昭・近藤公久 (2000)『日本語の語彙特性―朝日新聞の語彙文字頻度調査・第7巻』, 三省堂.

趙義成・呉文淑 (2004)「朝鮮語」『言語情報学研究報告』4, 27-49.

김홍규, 강범모 (1996) 「고려대학교 한국어 말모둠 1: 설계 및 구성」『한국어학』 3, .233-258. (Kim, Honggyu &Kang, Boemmo(1996)「高麗大学韓国語コーパス 1:設計と構成」『韓国語学』3)

강범모 (1995)「한글/한자 전자 텍스트의 로마자화 및 역방향 변환 프로그램과 한국어 데이터베이스」,『한국어 데이터베이스의 설계 및 응용을 위한 기초 연구』, 민음사, 223 - 270. (Kang Boemmo (1995)「ハングル/漢字電子テキストのローマ字化および逆方向変換プログラムと韓国語のデータベース」『韓国語のデータベースの設計および応用のための基礎研究』ミヌムサ)

강범모 (1997)「기계적 처리의 관점에서 본 국어의 로마자 표기법 개정안 97」,『국어의 로마자 표기법 개정 공청회 자료집』, 문화체육부, 35-45. (Kang Boemmo (1997)「機械的処理の観点から見た国語のローマ字表記法の改正案 97」『国語のローマ字表記法改訂公聴会資料集』文化体育部): (http://ikc. korea.ac.kr/~bmkang/rom97.PDF)

강범모 (2003)『언어, 컴퓨터, 코퍼스 언어학』, 교려대학교 출판부. (Kang Boemmo (2003)『言語, コンピューター, コーパス言語学』高麗大学出版部)

李翊燮・李相億・蔡琬 (2004)『韓国語概論』, 大修館書店.

野間秀樹(編著) (2007)『韓国語教育論講座』, 221-350, くろしお出版.

Sohn, Ho-min (1994) *Korean*, London& New York: Routledge.

Tamaoka, K. & Makioka, S.（2004）. Frequency of Occurrence for Units of Phonemes, Morae and Syllables Appearing in a Lexical Corpus of a Japanese Newspaper. *Behavior Research Methods, Instruments & Computers,* 36(3), 531-547.

Tamaoka, K. & Makioka, S. (2008, in press). Japanese mental syllabary and effects of mora, syllable, bi-mora and word frequencies on Japanese speech production. *Language and Speech*.

Tamaoka, K., & Terao, Y.（2004）. Mora or syllable? – Which unit do Japanese use in naming visually-presented stimuli? *Applied Psycholinguistics,* 25, 1-27.

유재원 (1993)「음운통계에 대한 연구」,『운율단위 음운논 및 음운 통계에 관한 연구』한국전자통신 연구소 연구보고서, 77-95. (Yu Jaewon(1993)「音韻統計に関する研究」『音律単位音韻論および音韻統計に関する研究』韓国電子通信研究所研究報告書)

〈言語資源および分析ツール〉

21 世紀世宗企画による Malmungchi: http: //www.sejong.or.kr

Perl 5.10: http://www.activestate.com/