２字漢字語の音韻類似性・音韻的距離に関する日韓中越データベースのオンライン検索エンジンの構築

于劭贇（名古屋大学院生）・金志宣（同）・玉岡賀津雄（名古屋大学）
＜共同研究者＞Hoang Thi Lan Phuong（名古屋大学院生）・ 張婧禕（名古屋大学）

1. はじめに

中国語を起源とした漢字表記の語彙は，周辺の日本語，韓国語，ベトナム語にも広く借用されている。日本では，幕末以降に，西欧の概念を表すための翻訳借用として新しい漢字語が造られ，広くアジア圏に波及した。韓国語では，漢字表記が非常に限定された場面でしか使われないものの，漢字語由来の語彙は全体の 60%に上る（Sohn, 2001）。ベトナム語では，アルファベット表記（クオック・グー）が使われているが，漢字由来の漢越語は語彙全体のおよそ 3 分の 1 を占めている（DeFrancis, 1977）。そのため，韓国語，中国語，またはベトナム語を母語とする日本語学習者は，日本語を学習する際に，母語の漢字語の知識を活用することができると予想される。しかし，日韓中越の間で漢字語の発音は多様である。L1 と L2 の間の音韻的な類似性のばらつきは，日本語学習者の漢字語の習得に強く影響すると予想される。こうした音韻的な研究を可能にするために，日韓中越 4 言語で共通した漢字語の言語間の音韻的な類似性に関する基礎データが不可欠である。

2. データベース

日本語では 2 字漢字語が非常に頻繁に使われており，これらの語彙が日本語の国語辞書の見出し語の約 70%を占めると報告されている（Yokosawa & Umeda, 1988）。本研究では，旧・『日本語能力試験出題基準』（2007，改訂版）4～2 級の 2 字漢字語 2,058 語（朴・熊・玉岡, 2014）を対象とする。日韓中越のそれぞれの言語において，対象となる 2 字漢字語のローマ字による音素表記の正規化を試みた（日本語の場合は音読みのある漢字語に限定した）。正規化された音素表記に基づき，日中，日韓，日越，中韓，中越，韓越の計 6 つのペアで，2 字漢字語の言語間の「音韻的距離」および「音素類似性」の指標を計算した。現在，本検索エンジンのデータベースは，日中で共通した 2 字漢字語 1,491 語，日韓で共通した 1,491 語，日越で共通した 1,475 語，中韓で共通した 1,509 語，中越で共通した 1,487 語, 韓越で共通した 1,487 語の音韻的距離および音素類似性の情報を収録している。

2-1. 音韻的距離と音素類似性の計算方法

2 言語間の音韻類似性を示す客観的基準として，①音韻的距離と②音素類似性を計算した。以下，音韻類似性という表現は，この 2 つの指標を示す。

2-1-1. 音韻的距離

英語などのアルファベット表記の言語を中心とした研究では，客観的な音韻類似性の指標として，一般化レーベンシュタイン距離（generalized Levenshtein distance）がよく使われている（Miwa, Dijkstra, Bolger, & Baayen, 2014; Schepens, Dijkstra, & Grootjen, 2011; Gooskens, Heeringa, 2004）。早川・于・初・玉岡（2017）は，日本語のヘボン式ローマ字表記と中国語のピンイン表記で，日中両言語間の一般化レーベンシュタイン距離を計算し，被験者による主観的な音韻類似性の判定結果と比較した。その結果，日中両言語の音素表記の対応付けが行われていないにも関わらず，客観的な音韻類似性の指標と主観的な音韻類似性の間に，中程度の相関が確認された（$r = -0.49, p < .01$）。本研究では，訓令式のローマ字表記に長母音を区別して（例えば，「公園」は/koo en/），より正規化された音素表記に基づき，日韓中越 4 言語の全てのペアについて，R の cba パッケージの sdists 関数（Buchta & Hahsler, 2017）で一般化レーベンシュタイン距離を計算し，言語間の「音韻的距離」とした。

表 1　音韻的距離の算出の例

日本語	中国語	編集操作	コスト
D	d	−	0
-	i	挿入	1
E	a	置換	2
N	n	−	0
-	h	挿入	1
W	-	削除	1
-	u	挿入	1
A	a	−	0
		合計（音韻的距離）	6

一般化レーベンシュタイン距離の計算は，1 つの文字配列をもう 1 つの文字配列に変形する際の挿入，削除，置換といった編集操作に重みを付けて，編集操作のコストが最小になるように 2 つの文字配列を整列させ，この最適整列（optimal alignment）のもとで編集コストを求める。本研究では編集操作の重み付けに，sdists のデフォルト値を使用した（挿入：1，削除：1，置換：2）。例えば，sdists で求められた日本語の「denwa（電話）」と中国語の「dianhua（电话）」の最適整列および音韻的距離は表 1 に示すとおりである。

2-1-2. 音素類似性

一般化レーベンシュタイン距離に基づく音韻的距離の指標は，2 つの文字配列の相違点を数値化しているが，文字列の長さに大きく影響されるという欠点がある。極端な例を挙げると，2 つの文字列に共通点がなく，類似性が常にゼロの場合でも，文字列が長ければ長いほど，音韻的距離の値が大きくなる。そこで，2 つの文字配列の相違した部分ではなく，共通した部分に注目した「音素類似性」という指標を考案した。音素類似性の計算方法は，まず前述の最適整列に基づいて，2 つの文字配列の共通した文字数，つまり編集操

作が不要な文字数を求める。そして，音素類似性の値が最小 0，最大 1 になるように，共通した文字数の値を 2 倍にして，その結果を 2 つの文字列の文字数の和で標準化する（式 1）。音素類似性の実際の計算は，筆者が開発した R パッケージ phonosim（Version 0.1; Yu, 2016）で行われた。

$$\frac{\text{文字列 A と B で共通した文字数}\times 2}{\text{文字列 A の文字数}+\text{文字列 B の文字数}} \quad \text{（式 1）}$$

図 1 は，日韓中越 4 言語の全てのペアで，本検索エンジンに収録されている 2 字漢字語の言語間の音素類似性の分布をカーネル推定したものである。図 1 から分かるように，日中，日韓，日越，中韓，中越，韓越の 2 字漢字語の音素類似性は，近似した分布に従っており，ほとんどが 0.4～0.5 の値をピークとして，高低で裾野が広がるようにほぼ対称的に分布していた。つまり，日韓中越 4 言語の全てのペアにおいて，音素類似性が中程度の 2 字漢字語が最も多く，類似性が高くまたは低くなるにつれて，語の数が少なくなっていくというパターンであることが共通している。

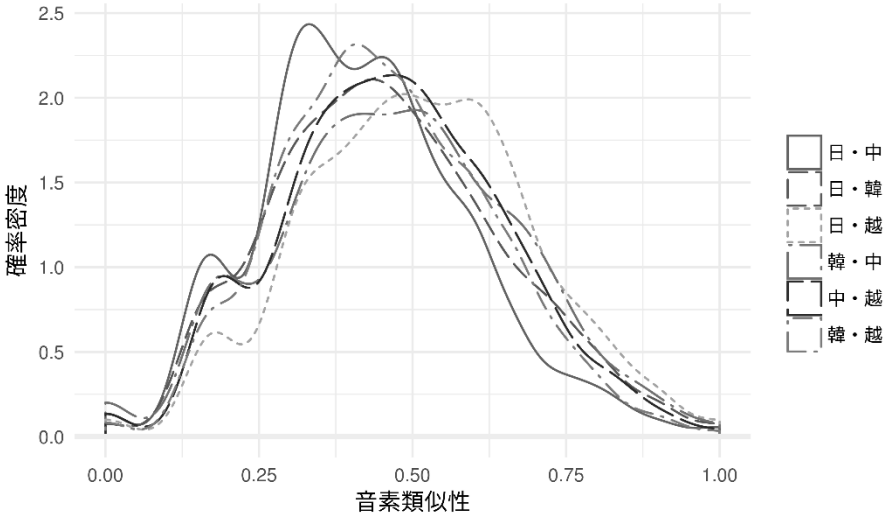

図 1　日韓中越 4 言語間における 2 字漢字語の音素類似性の分布

3. 検索エンジンの開発

本検索エンジン（http://kanjigodb.herokuapp.com/）は，既存の検索エンジン「同形二字漢字語の品詞性に関する日韓中データベース」（于・玉岡, 2015）の内容と機能を拡張したものである。「同形二字漢字語の品詞性に関する日韓中データベース」は，日韓中 3 言語の同形二字漢字語の書字，音韻，意味関係，品詞性の情報および日本語の使用頻度と難易度の情報をウェブ上で公開しており，詳細な検索オプションを提供している。

本研究では，この検索エンジンをベースとして，4 点の新しい機能を開発した。第 1 に，ベトナム語による検索を可能にし，従来の検索結果の画面にベトナム語の音韻情報を追加した（図 2）。

図 2　ベトナム語の音韻情報の追加

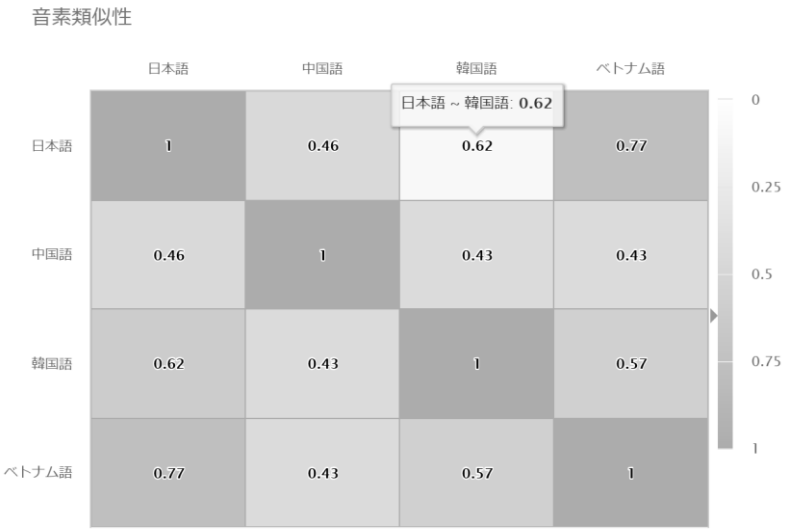

図 3　音素類似性の行列

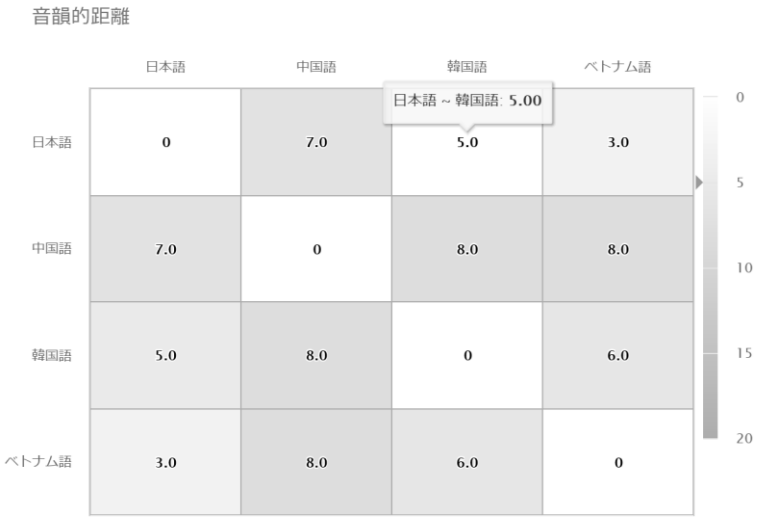

図 4　音韻的距離の行列

第 2 に，検索結果に，日韓中越 4 言語の全てのペアにおける音素類似性と音韻的距離の値を行列の形で表示するようにした（図 2，3)。これらの行列は，インタラクティブなものであり，マウスを個々のセルに移動すると，現在のセルが表している言語ペアが表示される。そのほか，行列の右端に，現在のセルの音素類似性または音韻的距離の値の全スケールにおける位置も表示される。これらの行列で，個々の漢字語について，日韓中越 4 言語間の音素類似性と音韻的距離を直感的に確認し，比較できる。

図 5　日中の音素類似性の全分布における個別の漢字語の位置（菱形のポイント）

図 6　音素類似性と音韻的距離による検索機能

第 3 に，音素類似性または音韻的距離の全体の分布における個別の漢字語の位置を示す機能を開発した（図 5)。音素類似性または音韻的距離の行列のセルの数値をクリックすると，現在のセルが表している言語ペアの 2 言語間の音素類似性または音韻的距離の全分布を表すヒストグラムが現れ，セルの数値の全分布における位置も表示される。第 4 に，音素類似性と音韻的距離による検索機能をもとの検索エンジンの「詳細検索」の画面に追加した（図 6)。日中，日韓，日越，中韓，中越，韓越の 6 つのペアで，音素類似性と音韻的距離の下限値と上限値をそれぞれ細かく指定でき，指定した範囲内の漢字語を検索することができる。

本検索エンジンは，日韓中越 4 言語の間の 2 字漢字語の音素類似性と音韻的距離の 2 つの客観的な音韻類似性の指標を，アクセスしやすく分かりやすい形で公開した。これらの 2 つの音韻類似性の指標は，日本語学習者が日本語の学習と共に形成する複数の言語のメンタルレキシコンでの音韻的な結合関係を研究するための基本的な情報となるであろう。

参考文献

（1）Sohn, H.-M. (2001). *The Korean Language*. Cambridge University Press.

（2）DeFrancis, J. (1977). *Colonialism and language policy in Viet Nam*. Mouton De Gruyter.

（3）Yokosawa, K., & Umeda, M. (1988). Processes in human Kanji-word recognition. *Proceedings of the 1988 IEEE international conference on systems, man, and cybernetics* (pp. 377-380). August 8-12, 1988, Beijing and Shenyang, China.

（4）朴善娟・熊可欣・玉岡賀津雄 (2014)「同形二字漢字語の品詞性に関する日韓中データベース」『ことばの科学』第 27 号 (特集号), 53–111.

（5）Miwa, K., Dijkstra, T., Bolger, P., & Baayen, R. H. (2014). Reading English with Japanese in mind: Effects of frequency, phonology, and meaning in different-script bilinguals. *Bilingualism: Language and Cognition*, *17*(3), 445–463.

（6）Schepens, J., Dijkstra, T., & Grootjen, F. (2011). Distributions of cognates in Europe as based on Levenshtein distance. *Bilingualism: Language and Cognition*, *15*(1), 157–166.

（7）Gooskens, C., & Heeringa, W. (2004). Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data. *Language Variation and Change*, *16*(3), 189–207.

（8）早川杏子・于劭贇・初相娟・玉岡賀津雄 (2017)「日中二字漢字語における客観的音韻類似性指標　―主観的音韻類似性指標との比較―」『関西学院大学日本語教育センター紀要』(6), 21–34.

（9）Buchta, C., & Hahsler, M. (2017). cba: Clustering for business analytics. *R Package Version 0.2-19*. https://CRAN.R-project.org/package=cba

（10）Yu, S. (2016). phonosim: An experimental R package for calculating phonological similarity. https://github.com/rongmu/phonosim

（11）于劭贇・玉岡賀津雄 (2015)「日韓中同形二字漢字語の品詞性ウェブ検索エンジン」『ことばの科学』第 29 号, 43–61.