日韓中越 4 言語における 2 字漢字語の音韻類似性に関するデータベースおよび検索エンジンの構築[1]

于 劭贇[2]
玉岡 賀津雄[3]
ホアーン ティ ラン フォン[4]

DOI: 10.18999/stul.33.75

要約:漢字文化圏の諸言語は，日中両言語を除けば書字体系が大きく異なっている。そのため，日中韓越 4 言語に数多く共通した漢字語彙の言語間類似性を定量化するにあたり，ヨーロッパ諸言語のように書字類似性を計測することは難しく，音韻類似性を計測することがより普遍性の高いアプローチになる。本研究では，旧・『日本語能力試験出題基準』の4〜2 級の2字漢字語 2,058 語から，日中，日韓，日越，中韓，中越，韓越という 6 つの言語対で共通した 2 字漢字語を抽出した。そして，日中韓越 4 言語間の漢字語彙の音韻類似性を客観的に定量化するために，一般化レーベンシュタイン距離に基づいて「音韻的距離」と，語長によるバイアスを排除した「音素類似性」という 2 つの指標を計算した。ドイツ語とオランダ語のように音韻類似性が高い同根語を多数共有する言語対はなく，漢字語の音韻類似性は，日中韓越 4 言語の 6 つの対が近似した分布をしていることが示された。日中韓越の 4 言語では，中程度の音韻的な距離の漢字語が多かった。さらに，ウェブ検索エンジン（http://kanjigodb.herokuapp.com/）を開発して，日中韓越 4 言語間の 2 字漢字語の音韻類似性の指標を直感的かつ効率的に調べることができるようにした。

キーワード:漢字語，同根語，音韻類似性，レーベンシュタイン距離，検索エンジン

1 English title: Construction of a database and search engine on the phonological similarity of two-kanji compound words in Japanese, Korean, Chinese and Vietnamese

2 Shaoyun Yu (Graduate Student, Graduate School of Humanities, Nagoya University, Japan;
E-mail: s.yu@nagoya-u.jp)
3 Katsuo Tamaoka (Professor, Graduate School of Humanities, Nagoya University, Japan;
E-mail: ktamaoka@gc4.so-net.ne.jp)
4 HOANG, Thi Lan Phuong (Graduate Student, Graduate School of Humanities, Nagoya University, Japan;
E-mail: lanp2579@gmail.com)

1. 研究背景

ヨーロッパ諸言語は，共通したインド・ヨーロッパ祖語を起源としており，古代から歴史的・文化的に深い関連性を持っている。英語，ドイツ語，オランダ語，フランス語，イタリア語，スペイン語などの諸言語では，書字，音韻的に類似している同根語（cognate）が多くみられる。実際，ヨーロッパ言語を中心としたバイリンガル研究で，同根語の認知処理が非同根語よりも速いという同根語促進効果（cognate facilitation effect）が注目されてきた（Dijkstra, 2005; Marian & Spivey, 2003; Costa, Caramazza & Sebastián- Gallés, 2000 など）。

一方，アジアでは，中国語，日本語，韓国語およびベトナム語は，言語起源における関連性が比較的薄いものの，これらの言語は同じ漢字文化圏にあったため，中国語を起源とした漢字語彙は，日本語，韓国語，およびベトナム語でも非常に広く使用されている（Yokosawa & Umeda, 1988; Sohn, 2001; DeFrancis, 1977）。ヨーロッパ諸言語は，ほとんどアルファベットに基づいた書字体系を使っているため，同根語の類似性を，書字類似性と音韻類似性という 2 つのアプローチから定量的に検討することが容易である（Schepens, Dijkstra & Grootjen, 2012; Schepens et al., 2013）。

それに対して，漢字語彙を共有する文化圏の諸言語では書字体系が大きく異なっている。韓国語とベトナム語では，漢字がほぼ使われなくなり，漢字由来の語彙の表記にも，韓国語では音節単位のハングルと呼ばれる表音文字，ベトナム語ではアルファベット表記のクオックグー（quốc ngữ）[5]と呼ばれる表音文字が使われている。中国語と日本語は，漢字を使用しているが，中国大陸における近代の漢字簡略化により，字形の異なった漢字が数多く生まれた。かくして，日中韓越 4 言語における漢字語彙の書字類似性については，中国語と日本語以外の言語では定量化そのものができず，音韻類似性のほうがより普遍的なアプローチといえよう。

同根語による促進効果は，書字体系の異なったヘブライ語と英語のバイリンガルに観察されている（Gollan, Forster & Frost, 1997）。さらに，日本語と英語のバイリンガルが英単語を認識するプロセスで，日英両言語の音韻類似性は複雑な影響を与え，全体的には音韻類似性が高いほど，反応時間が有意に速くなることが明らかにされている（Miwa, et al., 2014）。漢字由来の同根語は日中韓越 4 言語に多数存在しているが，中国語と日本

5 クオックグーは，「国語」という漢字のベトナム語での発音で，日本語の/koku go/とも発音がどことなく似ている。

語を除いたほとんどの対では言語間の書字的な類似性がないため，漢字圏のバイリンガルの漢字語彙の習得および認知処理を研究する上では，言語間の音韻類似性の影響が一層重視されるべきである。日本語の学習者に焦点を絞ると，音韻類似性は，漢字を使わない韓国語およびベトナム語の母語話者の漢字語の習得にとりわけ重要な役割を果たしていると考えられ，中国語母語話者に関しても，音韻類似性は書字類似性と共に漢字語の習得に強く影響すると考えられる。漢字圏における同根語の音韻に関連する研究を前進させるべく，本研究では，日中韓越 4 言語で共通した 2 字漢字語の音韻類似性の定量化を試み，ウェブ上で検索可能なデータベースを構築した。

http://kanjigodb.herokuapp.com/

2. 研究対象

本研究では，旧・『日本語能力試験出題基準』(2007，改訂版)の 4～2 級の 2 字漢字語 2,058 語(朴・熊・玉岡，2014)の範囲から，研究対象の漢字語を選定した。日本語では漢字 2 字によって構成される漢字語が最も多く(Yokosawa & Umeda, 1988)，初級から中級までの日本語学習者に向けた 2 字漢語は，日本語で最も一般的に使用されている漢字語を代表すると考えられる。その内，日中で共通した 2 字漢字語が 1,491 語，日韓で共通したのが 1,491 語，日越で共通したのが 1,475 語，中韓で共通したのが 1,509 語，中越で共通したのが 1,487 語，韓越で共通したのが 1,487 語であった(日本語の場合は音読みのある漢字語に限定した)。本研究では，4 言語の対で共有されている 2 字漢字語を研究対象とした。さらに，日中韓越 4 言語での漢字語の発音を比較するために，対象となる各言語の 2 字漢字語を訓令式[6]のローマ字で表記して，音素表記で対応付けた。

3. 客観的な音韻類似性の指標

音韻類似性の定量化は，主観的な評価法と客観的な評価法の 2 種類のアプローチがある。一つは，リッカート尺度などを利用した主観的な評価法で，2 言語間の音韻的な異同に対する人間の直感的な判断を反映できる利点がある。しかし，判定者の個人差で評

6 訓令式のローマ字表記は，日本の言語学および音韻論の研究でも音素表記に使用されている。長母音は，aaやooなどのように母音を2回繰り返して表記して計算に使った。

価が揺れる傾向があり，日中韓越 4 言語に存在する 6 つの言語対の音韻類似性を，すべて一定の基準で評価することは難しい。もう一つは，客観的な評価法で，2 言語間の語の音韻的な特性に基づいて，アルゴリズムで計算する方法である。これは，個々人の主観的な評価で起こりうる不安定性を排除することができる。これまで，一般化レーベンシュタイン距離（generalized Levenshtein distance）による音韻類似性の指標は，アルファベット表記の言語の研究で有効性が示されてきた（Miwa, et al., 2014; Schepens et al., 2013; Gooskens & Heeringa, 2004）。本研究は，日韓中越 4 言語のすべての対について，2 字漢字語の言語間の一般化レーベンシュタイン距離を計算し，「音韻的距離」とした。さらに，音韻的距離の指標の短所を補うために，言語間で相違した部分ではなく，共通した部分に着目した「音素類似性」という指標を考案し，日中韓越 4 言語について計算した。

3.1 音韻的距離

一般化レーベンシュタイン距離は，重み付け編集距離（operation-weight edit distance）ともいわれ，文字配列を目標の文字配列に変形するための最小の編集コストを表す（Gusfield, 1997 ; Boytsov, 2011）。これを，訓令式による音素表記が正規化された日中韓越 4 言語の 2 字漢字語のローマ字表記の文字列に応用すると，2 言語間の語が音韻的にどれほど異なるかを客観的に評価することができる。音韻的距離の値が小さいほど（最小はゼロ），2 言語間の語の音韻類似性が高いことになる。一般化レーベンシュタイン距離を計算する際は，まず変形に必要な挿入，削除，置換といった操作のコストに重みを付ける。

一般的に，挿入と削除のコストがそれぞれ 1，置換のコストが 2 とされることが多く，これは R の cba パッケージの sdists 関数のデフォルト値でもある（Buchta & Hahsler, 2017）。そして，設定された重み付けのもとで，アルゴリズムは編集コストが最小になるように 2 つの文字配列を整列させ，「最適整列（optimal alignment）」を求める。こうした最適整列で得られた編集コストが，一般化レーベンシュタイン距離である。本研究では，sdists 関数のデフォルトの重み付けで音韻的距離を計算した。例えば，日本語の「denwa（電話）」と中国語の「dianhua（电话）」の距離を求めてみる。表 1 に sdists での最適整列を示した。この最適整列で必要な編集操作のコストを合計すると，「電話」の日中間の音韻的距離が得られる。

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

3.2 音素類似性

前述の音韻的距離の指標には，2 つの短所がある。第1に，音韻的距離は言語間の相違点を定量化するものであり，類似性を測定したものではない。第 2 に，音韻的距離の値は語の音韻を表す文字列の長さに大きく影響されることである。こうした短所を説明するために，音韻的な共通点が一切ない 2 つの語の対を想定しよう。類似性からみれば，この 2 つの語の類似性は常にゼロのはずである。しかし，音韻的距離の観点からすると，一方の語が長ければ長いほど，もう一方の語をその語に変形するための編集コスト（特に挿入操作によるコスト）が高くなり，音韻的距離の値もそれに応じて大きくなってしまう。

そこで，本研究では，語の相違した部分ではなく，共通した部分を定量化する「音素類似性」という指標を考案した。音素類似性の計算は，まず音韻的距離と同じく，一般化レーベンシュタイン距離を計算するための編集操作の重み付けを指定し，2 つの文字配列の最適整列を求める。次に，最適整列のもとで，編集操作を不要とする文字の数，すなわち編集コストがゼロである文字の数を合計し，その結果を 2 つの文字配列の「共通文字数」とする。最後に，語長による影響を排除するために，共通文字数を2倍にし，その結果を2つの文字配列の長さの和で割る（式1）。これは共通文字数の値を2つの語の平均語長で標準化することに相当し，こうして得られた音素類似性の値はすべて 0 から 1 までの範囲内で変動する。本研究では，Yu（2016）が開発した R の phonosim パッケージ（Version 0.1）を使って，日中韓越 4 言語のすべての対について音素類似性を計算した。

$$\text{音素類似性} = \frac{\text{文字列 A と B の最適整列における共通文字数} \times 2}{\text{文字列 A の長さ} + \text{文字列 B の長さ}} \qquad (式1)$$

3.3 結果と考察

3.3.1 全体的な分布特徴

研究対象の 2 字漢字語の音韻類似性は，日中韓越 4 言語の各対でどのような分布上の特徴があるかを概観するために，音韻的距離と音素類似性の 2 種類の指標の結果に対してカーネル密度推定を行った。図 1 に示したように，日中韓越 4 言語間の音韻的距離の分布は，6 つの対について近似した分布をみせた。同様に，図 2 に示された音素類似性の分布も，よく似た分布を示した。

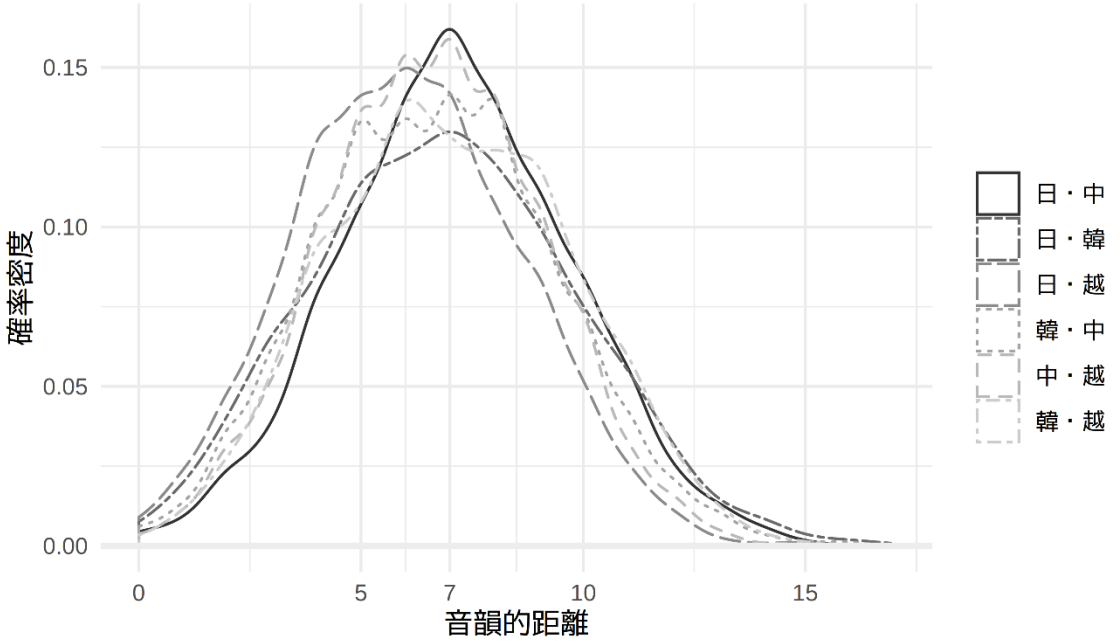

図 1　日中韓越 4 言語における 2 字漢字語の音韻的距離の分布（カーネル密度推定）

シャピロ・ウィルク検定でデータの正規性を検定したところ，すべての言語対において，音韻的距離および音素類似性の結果が正規分布に従うという帰無仮説が棄却され（すべて $p < .001$），この 2 種類の音韻類似性の指標の結果が正規分布に従うと判断することができなかった。一方，データそのものを考察すると，音韻的距離および音素類似性の結果は，いずれもデータが中央の値に集中しており，高低に裾野が広がるようにほぼ対称的に分布していることが分かる。データの歪度を計算すると，各言語対における音韻的距離および音素類似性の歪度はすべて ± 0.5 以内であり（表 2，表 3），データの分布が比較的対称的であるといえる（Bulmer, 1979）。データの尖度を計算すると，各言語対におけ

る音韻的距離および音素類似性の尖度はすべて 3 前後であり(表 2, 表 3), 正規分布の尖度に比較的近いといえる。

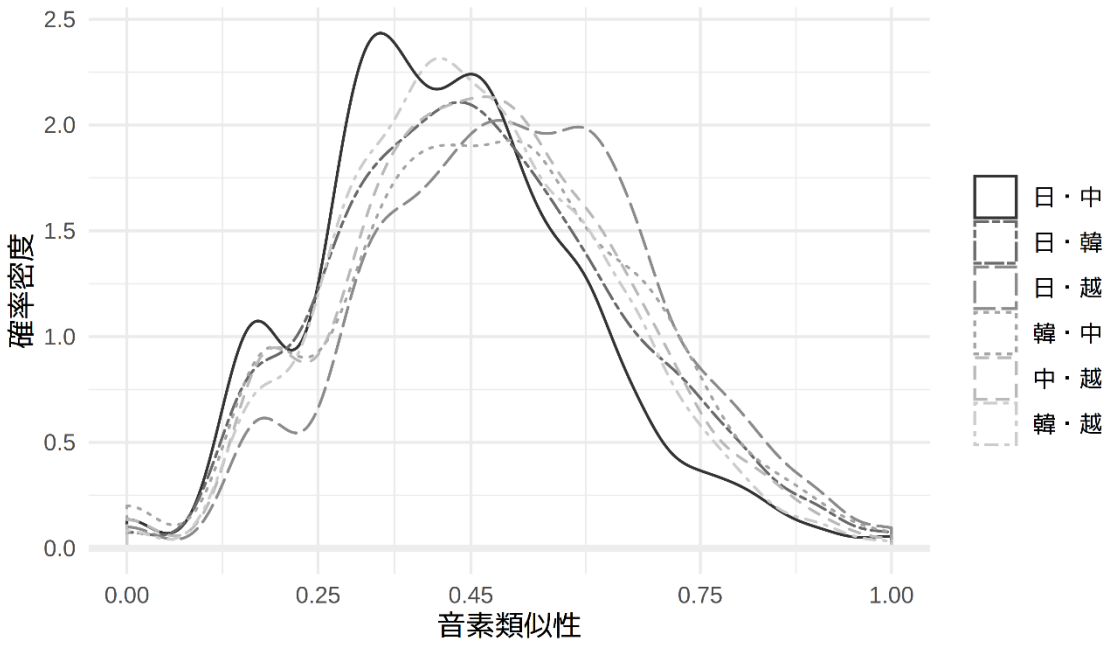

図 2 日中韓越 4 言語における 2 字漢字語の音素類似性の分布(カーネル密度推定)

表 2 日中韓越 4 言語における 2 字漢字語の音韻的距離の要約

言語対	日・中	日・韓	日・越	韓・中	中・越	韓・越
最小値	0	0	0	0	0	0
25%分位点	5	5	4	5	5	5
中央値	7	7	6	7	7	7
75%分位点	9	9	8	8	8	9
最大値	15	17	15	16	15	15
四分位範囲	4	4	4	3	3	4
歪度	0.04	0.16	0.14	0.14	0.03	0.05
尖度	2.91	2.76	2.78	2.84	2.77	2.54

表 3　日中韓越 4 言語における 2 字漢字語の音素類似性の要約

言語対	日・中	日・韓	日・越	韓・中	中・越	韓・越
最小値	0	0	0	0	0	0
25%分位点	0.31	0.33	0.38	0.33	0.33	0.33
中央値	0.40	0.44	0.50	0.47	0.46	0.44
75%分位点	0.53	0.60	0.62	0.62	0.60	0.57
最大値	1	1	1	1	1	1
四分位範囲	0.23	0.27	0.24	0.28	0.27	0.24
歪度	0.37	0.29	0.01	0.02	0.08	0.21
尖度	3.26	2.79	2.86	2.75	2.86	2.88

3.3.2　音韻的距離の結果

日中韓越 4 言語の各対の音韻的距離の分布の詳細を確認するために，データの箱ひげ図と散布図を作成した（図 3）。箱ひげ図の「箱」の中央にある縦棒は中央値に対応し，データの中心傾向を表すことができる。「箱」の両端の縦棒は，それぞれ 25% 分位点と 75% 分位点に対応し，両端の距離は四分位範囲に対応する。四分位範囲は，分布中央の 50% のデータを含んでおり，データの散布度を表すことができる。

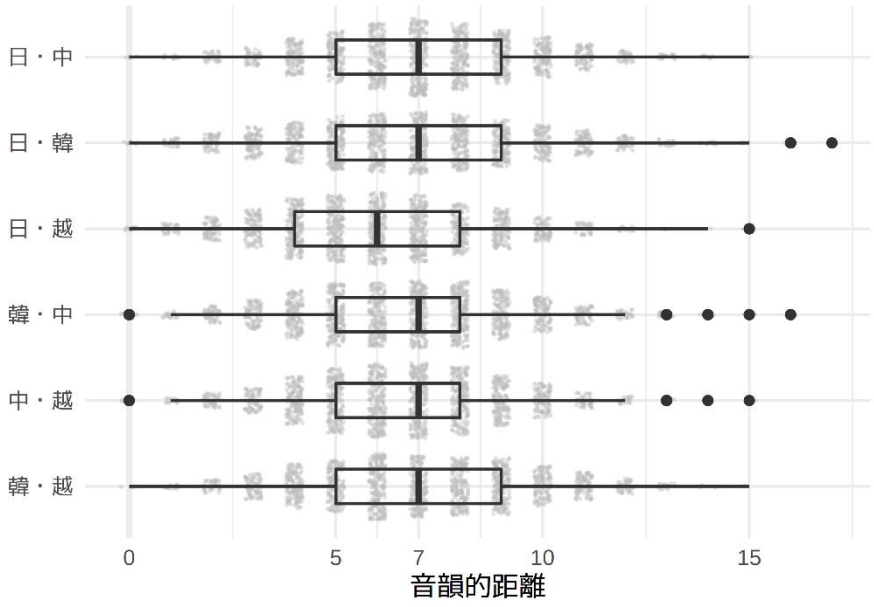

図 3　日中韓越 4 言語における 2 字漢字語の音韻的距離の分布（箱ひげ図・散布図）

各言語対の音韻的距離の中央値は，ほとんど 7 になっており，日本語とベトナム語の音韻的距離は 6 と，他と比べてやや小さい方向に寄っている可能性が示唆される。各言語対の音韻的距離の四分位範囲は 4 か 3 となっている。日中韓越 4 言語の 6 つの対がすべて近似した形の分布をしていることが，さらに確認できる。

3.3.3 音素類似性の結果

日中韓越 4 言語の各対の音素類似性の分布の詳細を確認するために，データの箱ひげ図と散布図を作成した（図 4）。また，データの分布の傾向を要約する記述統計値を表 3 にまとめた。各言語対の音素類似性の中央値は，0.45 前後となっており，日中韓越 4 言語間で，音韻類似性が中程度の 2 字漢字語が最も多いことを示している。日本語と中韓越 3 言語との音韻類似性に絞ってみると，日中の音韻類似性の中央値が 0.40 と一番低く，日韓が 0.44 と中間的な位置にあり，日越が 0.50 と一番高いという結果になっている。各言語対の音韻的距離の四分位範囲は，0.25 前後となっている。

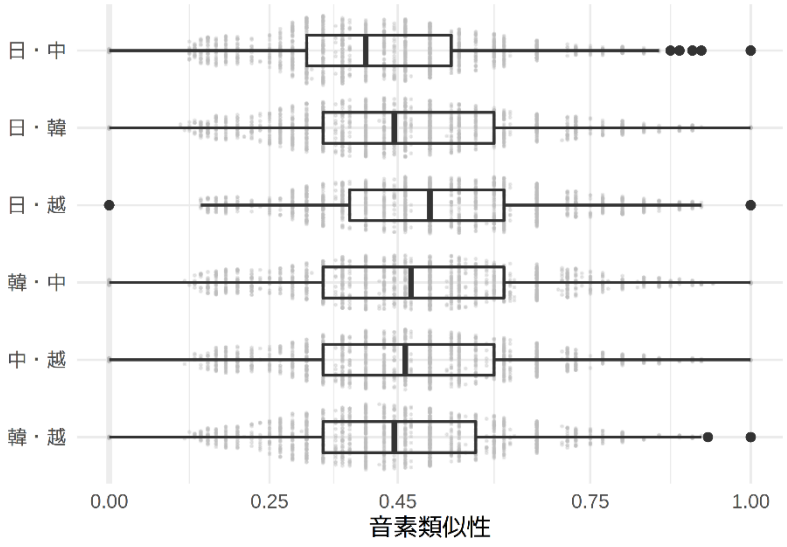

図 4 日中韓越 4 言語における 2 字漢字語の音素類似性の分布（箱ひげ図・散布図）

3.3.4 音韻的距離と音韻類似性の関係

音韻的距離と音素類似性という 2 つの指標は，いずれも最適整列に基づいて計算されているため，高い相関が予想される。そこで，日中韓越 4 言語のすべての対で，音韻的距離と音素類似性のピアソンの積率相関係数を計算した。表 4 に示したように，音韻的距離

と音素類似性の両指標は，すべての言語対において，−0.88 前後と非常に高い負の相関を示した。相関係数が負になっているのは，音韻的距離は 2 つの語の音韻的な相違点を定量化した指標であり，その値が小さいほど音韻類似性が高いためである。

表 4　日中韓越 4 言語における 2 字漢字語の音韻的距離と音素類似性の相関係数

言語対	日・中	日・韓	日・越	韓・中	中・越	韓・越
相関係数	−0.90	−0.90	−0.92	−0.86	−0.87	−0.88
95% 信頼区間	[-0.91, -0.89]	[-0.91, -0.89]	[-0.93, -0.92]	[-0.87, -0.84]	[-0.88, -0.86]	[-0.89, -0.86]

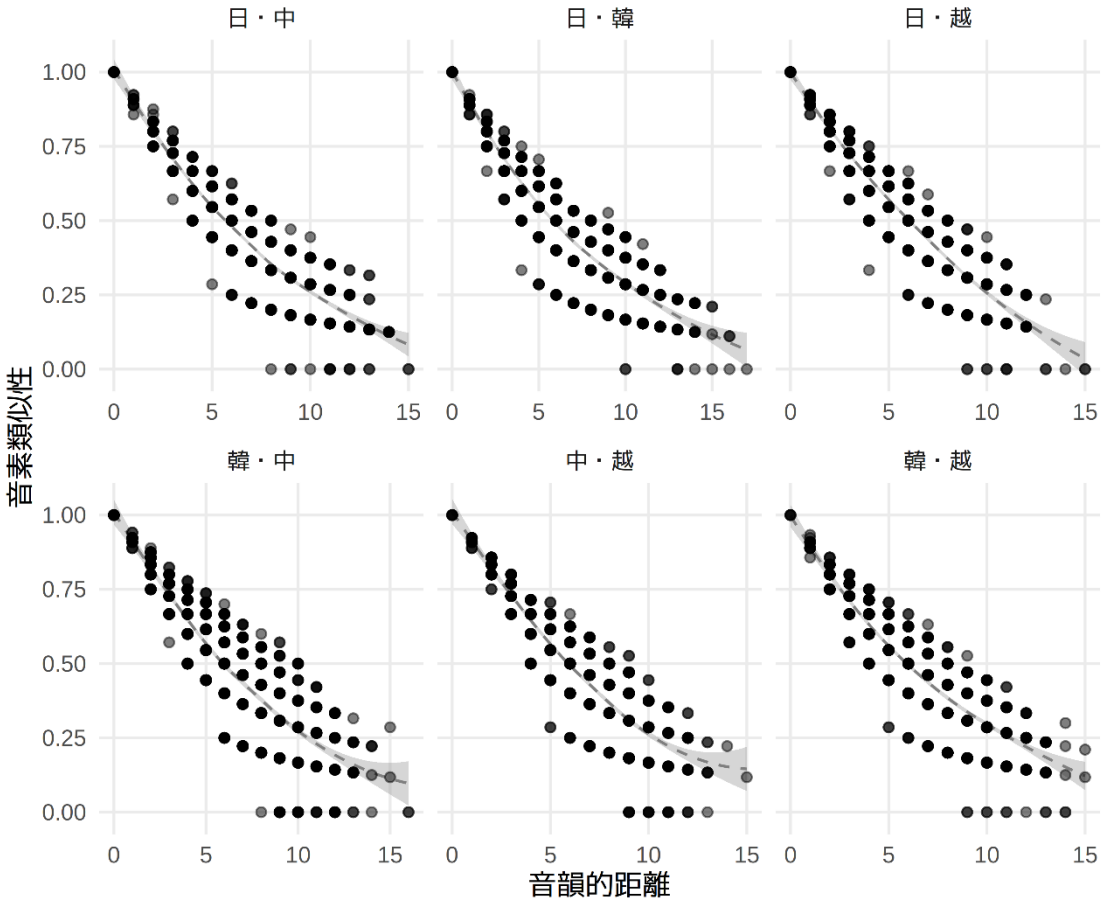

図 5　日中韓越 4 言語における 2 字漢字語の音韻的距離と音素類似性の散布図

さらに，音韻的距離と音素類似性の詳細な対応関係を調べるために，すべての言語対の両指標の散布図（図 5）を作成した。グラフの左上部ではデータポイントの集中度が高い。グラフの中下部では，データのばらつきが大きい。これは，音素類似性が高いほど両

指標の対応がよく，音素類似性が低い場合は，両指標の対応が悪いことを示している。

また，横方向の X 軸を主眼に考察すると，音韻的距離が同じである複数の漢字語の対は，縦方向の Y 軸においては音素類似性の値が多様な結果になっていることが分かる。特に，音韻的距離が 10 前後のデータをみると，音韻的距離が同じである複数の漢字語の対の中でも，Y 軸の音素類似性が0.5付近に位置するものもあれば，一番低いゼロになっているものもある。その理由は，音韻的距離の指標は，3.2 節に述べたように，語長の影響を十分に考慮しておらず，音韻類似性が大きく異なる複数の語の対であっても，それぞれの対の語長のバイアスによって同じ数値に計算される可能性があることである。一方，音素類似性の指標は，対をなす 2 つの語の平均語長で標準化を行っているため，音韻的距離のこうした短所を補うことができる。

3.3.5 考察

研究対象の 2 字漢字語の音韻類似性は，日中韓越 4 言語の 6 つの対に共通して，音韻類似性が中程度の語が最も多く，類似性が高くまた低くなるほど，語の数が少なくなっていくというパターンが観察された。こうしたパターンは，Schepens et al.（2013）のヨーロッパ 6 カ国語（英・独・蘭・仏・伊・西）を対象とした研究の結果と類似している。Schepens et al.（2013）は，ヨーロッパ 6 カ国語の翻訳同義語（translation equivalents）の書字類似性と音韻類似性を計算し，音韻類似性は書字類似性と比べた。そして，データが中央の値に集中しており，音韻類似性が高いほど語の数が急速に減っていくという傾向があることを示した。ただし，ドイツ語とオランダ語が例外であり，この 2 つの言語の翻訳同義語の音韻類似性は，中央の 0.5 から最大値の 1 までほぼ同等の数で直線的に分布していた。さらに，Schepens et al.（2013）が同根語に絞って考察している。そして，ドイツ語とオランダ語の対は，他の強い関連の言語対と比べて，音韻類似性の高い同根語を多く有していることを示した（図 6）。一方，本研究で得られた漢字圏の言語同士の音韻類似性データでは，ドイツ語とオランダ語のような非常に親密な関係にある言語対はなく，すべての言語間で漢字語の音韻はある程度離れた距離にあるようである。これは，日中韓越 4 言語の音韻体系が互いに独立していた部分が多く，また，漢字語の共有が，活発な社会・文化的な交流を介して，語彙が他の言語へと発音（音韻）を変えて借用された結果であると思われる。

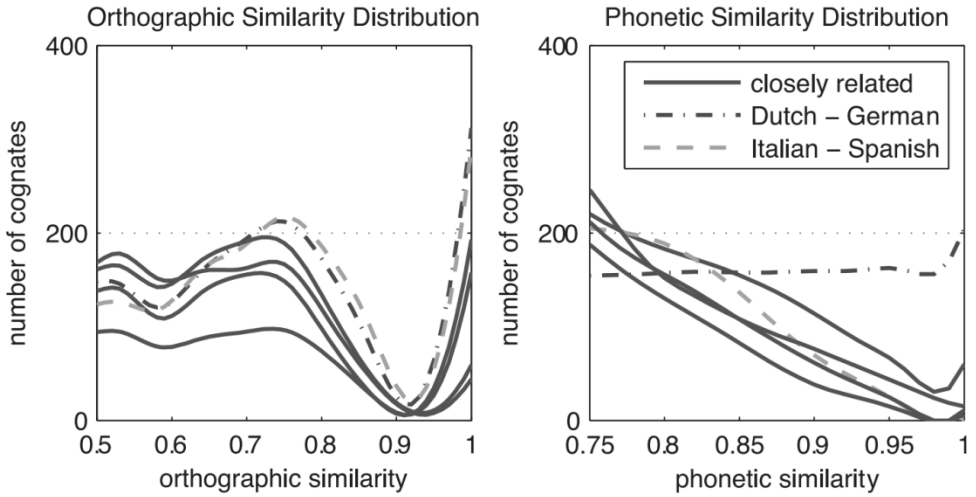

図 6 ヨーロッパ諸言語における同根語の書字・音韻類似性の分布（Schepens et al, 2013 より）

日本語と中国語，韓国語，ベトナム語の 3 言語の対の音素類似性を考察すると，日中の音素類似性が最も低い可能性が示唆された。それは，現代中国語の標準語である北京官話が，日本語の呉音・漢音の起源である古代中国語の中古音と大きなずれがあるからであろう。また，漢字圏の中で，日本語とベトナム語が地域的に最も離れているのに，日越音素類似性が，すべての言語対で最も高かったのも興味深い結果であった。

4. ウェブ検索エンジン

本研究で得られた音韻類似性の指標のデータを研究者や日本語学習者に広く利用してもらえるように，本研究ではさらに，音韻類似性のデータベースをウェブ上で検索できるようにした（http://kanjigodb.herokuapp.com/）。音韻類似性データのウェブ検索エンジンは，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）の内容および機能を拡張した形で開発された。于・玉岡（2015）の検索エンジンをベースとして，新たにベトナム語の音韻情報を追加した上で，日中韓越 4 言語の音韻的距離および音素類似性のインタラクティブな可視化を実現した。それに加えて，音韻類似性データに基づいた詳細な検索オプションも提供している。

検索エンジンは，漢字，かな，中国語のピンイン，韓国語のハングルおよびイェール式ローマ字表記，ベトナム語のクオックグーのいずれでも検索することができる。検索結果の画面の最初に，日中韓越 4 言語の音韻情報が漢字表記と共に表示される（図 7）。

図 7 日中韓越 4 言語の音韻情報

音韻情報や漢字表記などの基本情報の下に，日中韓越4言語の6つの対の音素類似性および音韻的距離の結果が，インタラクティブな「音素類似性行列」と「音韻的距離行列」としてまとめられている（図8，図9）。マウスを，行列にある個々のセルに移動すると，そのセルが表している言語対の名前と音韻類似性の指標の値がポップアップの形で強調される。また，行列の右端のスケールは，現在強調されているセルの値が，音素類似性または音韻的距離の全範囲でどの位置にあるかを標示する。利用者がこうした画面で，各言語対の音韻類似性の指標を効率よく概観し，比較できるようにした。

「音素類似性行列」または「音韻的距離行列」の特定のセルをクリックすると，そのセルが表している言語対の音素類似性または音韻的距離の全分布が，ヒストグラムの形で現れる（図 10）。そして，検索された漢字語のデータの値も菱形のポイントで横軸に標示される。利用者がこの画面で，検索された漢字語は，各言語対の全データにおいて，その音韻類似性の指標が高いかどうか，または似たような値を持つ語がどれほど存在するかを分かりやすく確認できる。

最後に，検索エンジンの「詳細検索」メニューをクリックすると，多くの高度な検索オプションが利用できるようになる。「音素類似性／音韻的距離」のセクション（図 11）では，各言語対における音素類似性または音韻的距離の下限値および上限値を指定し，その範囲内の漢字語を抽出することができる。検索オプションは併用可能なため，複数の言語対に跨って上下限値を設定し，検索語を絞り込むことも可能である。これは，特定の音韻類似性の条件を満たした語を細かく選出する場合に活用できる機能である。

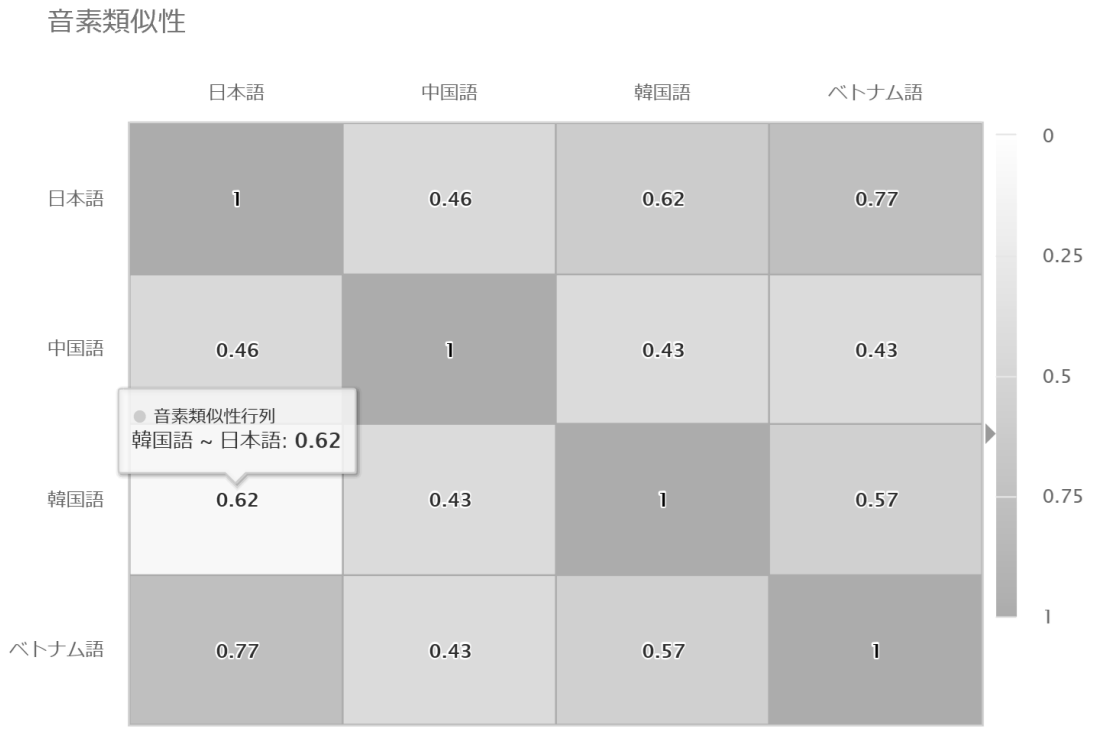

図 8　音素類似性行列

図 9　音韻的距離行列

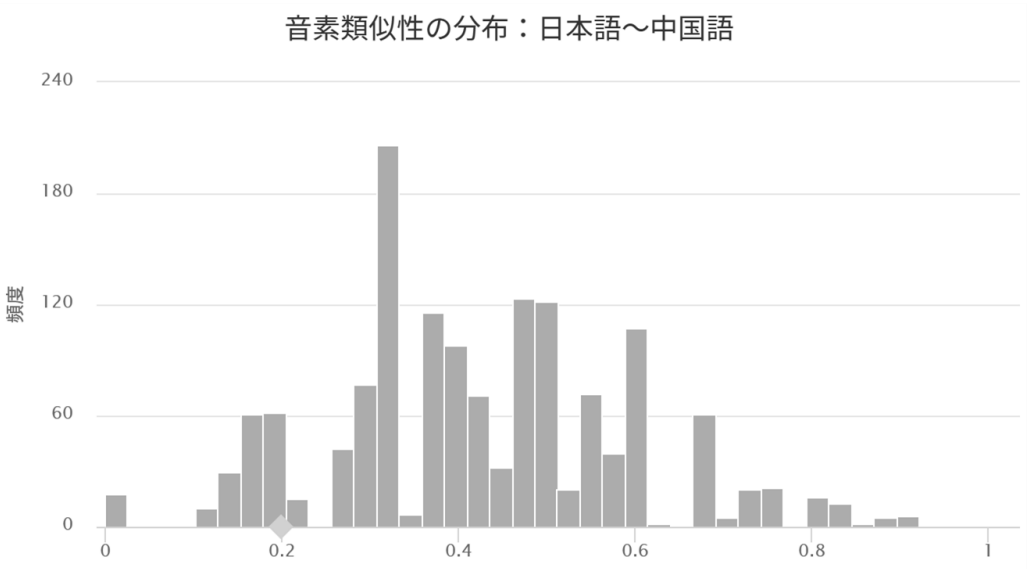

図 10　音素類似性の指標の全分布における個別の漢字語の位置

音素類似性／音韻的距離

日本語～中国語

類似性下限　類似性上限　距離下限　距離上限

日本語～韓国語

類似性下限　0.4　類似性上限　0.6　×　距離下限　距離上限

日本語～ベトナム語

類似性下限　類似性上限　距離下限　距離上限

中国語～韓国語

類似性下限　類似性上限　距離下限　距離上限

中国語～ベトナム語

リセット　検索

図 11　音素類似性と音韻的距離による検索機能

5. 総括

本研究では，音韻的距離と音素類似性という 2 つの客観的な指標を用い，日中韓越 4 言語で共通した 2 字漢字語の音韻類似性の定量化を試みた。2 つの指標は，高い相関関係を有しているが，それぞれの指標は，音韻的な相違点または共通点に重きを置いており，相補的な関係にあるといえる。

今回の研究対象である 2 字漢字語の結果を考察すると，各言語対では，いずれも音韻類似性が中程度の語が最も多く，日中韓越 4 言語の漢字語は互いに，近くもなく遠くもないという音韻的な距離が置かれていると考えられる。漢字圏のバイリンガル（2 言語併用者）は，言語間の漠然とした類似性を基に，多様に異なる漢字の発音を習得しなければならないと推察される。音韻的類似性の度合いが，第 2 言語における漢字語彙の習得および認知処理にどのように影響するかについては，さらなる研究が待たれる。本研究では，こうした研究に不可欠な基礎データを構築して，豊富な機能を備えた検索エンジンによって，アクセスしやすい形で公開した。

6. 今後の研究

漢字語の音韻類似性の定量化に関しては，今後 2 つの方向性が考えられる。一つは，日中韓越 4 言語を横断した主観的な音韻類似性のデータベースを構築することである。主観的な方法には，母語の違いや個人差が影響する。しかし，人間の音韻認知の指標を考えるならば，主観的な指標のほうが現実の 2 言語間の音韻的な関係をより良く示しているともいえそうである。また，主観的な指標は，客観的な指標を検証する上で重要であろう。

もう一つは，客観的な指標を改良することである。本研究で考案された音素類似性の指標は，語長によるバイアスを排除している点で，単純なレーベンシュタイン距離より改良されているが，音韻的な異同を二分的に捉えている点に関しては，通常のレーベンシュタイン距離を大きく超えているとはいえない。標準化レーベンシュタイン距離（Normalized Levenshtein Distance, NLD）も，本研究の音素類似性と近似している（Schepens et al., 2012）。音素類似性は，対をなす 2 つの語の平均語長で数値の標準化をしているのに対し，NLD は 2 つの語の語長の最大値を用いて標準化してる。他の方法としては，音韻的

な異同を二分的ではなく段階的に定量化するアプローチが考えられる。例えば，Kondrak（2002）は，12の音声学的特徴に基づいた段階的な音韻類似性の測定方法を提案している。

さらに，レーベンシュタイン距離に基づいたアプローチでは，Gooskens & Heeringa（2004）が，ノルウェイ方言を対象に，音声スペクトルにおける違いで一般化レーベンシュタイン距離の編集操作に段階的な重み付けをし，方言間の音韻類似性を計算している。また，Schepens et al.（2013）も，IPA（国際音声記号）に基づいた音声空間における距離によって，一般化レーベンシュタイン距離に段階的な重み付けをした。これらのアプローチは，語彙の音声的特徴をより精緻に分析した元データを基盤としている。今後，日中韓越4 言語の共通した漢字語彙についてもこうした作業が必要であろう。

［参考文献］

Boytsov, L.（2011）. Indexing methods for approximate dictionary searching: Comparative analysis. *Journal of Experimental Algorithmics*, 16, 1-1. DOI 10.1145/1963190.1963191

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

Bulmer, M. G. (1979). *Principles of Statistics*. New York: Dover.

Costa, A., Caramazza, A., & Sebastian-Galles, N. (2000). The cognate facilitation effect: implications for models of lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(5), 1283–1296.

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

Dijkstra, T. (2005). Bilingual visual word recognition and lexical access. In J. F. Kroll & A. De Groot (eds.), *Handbook of bilingualism: Psycholinguistic approaches*, pp. 178–201. Oxford: Oxford University Press.

Gollan, T. H., Forster, K. I., & Frost, R. (1997). Translation priming with different scripts: Masked priming with cognates and noncognates in Hebrew–English bilinguals. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23(5), 1122–1139.

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

189–207.

Gusfield, D.（1997）. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. New York : Cambridge University Press.

Kondrak, G. (2002). *Algorithms for Language Reconstruction* (Doctoral dissertation). University of Toronto.

Marian, V., & Spivey, M. (2003). Competing activation in bilingual language processing: Within-and between-language competition. *Bilingualism: Language and Cognition, 6*(2), 97–115.

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.

Schepens, J., Dijkstra, T., & Grootjen, F. (2012). Distributions of cognates in Europe as based on Levenshtein distance. *Bilingualism: Language and Cognition, 15*(1), 157–166.

Schepens, J., Dijkstra, T., Grootjen, F., & van Heuven, W. J. B. (2013). Cross-Language Distributions of High Frequency and Phonetically Similar Cognates. *PLoS ONE, 8*(5), e63006. https://doi.org/10.1371/journal.pone.0063006

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

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.

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

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.

朴善娴・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.

于 劭贇 － 名古屋大学大学院　人文学研究科・博士後期課程大学院生

玉岡　賀津雄 － 名古屋大学大学院　人文学研究科・教授

ホアーン　ティ　ラン　フォン － 名古屋大学大学院　人文学研究科・博士後期課程大学院生

Construction of a database and search engine on the phonological similarity of two-kanji compound words in Japanese, Korean, Chinese and Vietnamese

YU, Shaoyun (Graduate Student, Graduate School of Humanities, Nagoya University, Japan)
TAMAOKA, Katsuo (Professor, Graduate School of Humanities, Nagoya University, Japan)
HOANG, Thi Lan Phuong (Graduate Student, Graduate School of Humanities, Nagoya University, Japan)

Abstract: Although Chinese, Japanese, Korean and Vietnamese (hereafter, CJKV) share a large number of Chinese-originated cognates, the Han characters (kanji) are currently used only in Chinese and Japanese. As a result, it is difficult to quantify cognate similarities between these four Asian languages by measuring orthographic similarities, which are commonly studied when comparing between European language cognates. Instead, phonological similarity should be a more universal approach to quantify cognate similarities between the four languages of CJKV because of its independence within writing systems. Accordingly, we extracted two-kanji compound words shared by the CJKV languages from a database of 2,058 kanji words. Two objective measures of phonological similarity were computed for each language pair: (1) *Phonological Distance*, which is based on generalized Levenshtein distance, (2) *Phoneme Similarity*, which mitigates the bias of word length. All six possible language pairs followed a similar pattern: the distribution of phonological similarity of cognates was near-symmetric, centered on cognates of median similarities. While in European languages, there are language pairs such as German-Dutch that share many phonologically highly similar cognates, no such pair was found among CJKV. In order to make our calculation results accessible to the public, we developed an online search engine that features intuitive and interactive display of phonological similarity measures (http://kanjigodb.herokuapp.com/).

Keywords: two-kanji compound words, cognates, phonological similarity, Levenshtein distance, search engine