様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　元 年　 ６ 月 ２８ 日現在

機関番号：４５５０７
研究種目：基盤研究(C)（一般）
研究期間：2015～2018
課題番号：１５Ｋ００４７７

研究課題名（和文）基礎語彙を含む多次元尺度による言語系統分類自動補完のための系統樹生成手法の開発

研究課題名（英文）Development of Language Family-Trees Generation Method for Automatic Completion of Language Classifications by Multidimensional Scale Including Basic Vocabulary

研究代表者
呉　靭（Wu, Ren）

山口短期大学・情報メディア学科・准教授

研究者番号：７０７０８０１５

交付決定額（研究期間全体）：（直接経費）　　3,600,000円

研究成果の概要（和文）：本研究では、遺伝系統樹の作成手法を言語学分野に応用し、言語系統樹の自動生成法を提案した。基礎語彙に基づき、文字列類似度を用いた言語間距離の計算モデルを提案し、さらにランダムフォレストによる特徴量抽出の手法を応用して言語素性の重要度を計算することによって、言語分類に影響度が高いと思われる言語素性も考慮した言語間距離の計算モデルを提案した。言語特徴としての文法構造については、Fillmoreの格文法を用い、その表層化過程の生成規則による言語ごとの特徴の記述を試み、さらに単一化文法を適用することにより文法記述における一致の問題を解決した。これにより、文法のデータの形式化が可能となった。

研究成果の学術的意義や社会的意義
本研究は分子系統学および情報科学を言語学に応用した文理融合の学術的特色を持っており、その成果はコンピュータの機能を活かした言語特徴の横断的分析の研究、さらには言語類型論研究に寄与する。

研究成果の概要（英文）：Applying generation techniques of gene phylogenetic-tree to the linguistic research field, we proposed several methods of generating language family-trees. Firstly, a computation model was proposed to measure inter-language distances by using edit distance and Jaro-winkler distance based on basic vocabulary. Then, a model was constructed to involve those linguistic features that are considered to have high influence on language classifications by applying the method of feature extraction by random forest. Regarding grammatical structure as a language feature, the grammatical differences of languages were focused on and the characteristics of each language were described by the generation rules in the surface layering process of Fillmore’s case grammar. Finally, the concordance problem in grammar was solved by applying of unification grammar. This made it possible to formalize grammar data used in grammar analyzing.

研究分野：言語情報学

キーワード：言語系統分類　基礎語彙　系統樹説　言語間距離　文字列類似度　言語素性　波紋説　格文法

様　式　C－１９、F－１９－１、Z－１９、CK－１９（共通）

１．研究開始当初の背景

情報技術の進化に伴い、言語学分野におけるコンピュータの利用が益々広がってきている。現在のコンピュータ処理においては、「言語コード」が言語の一意的識別子として使われるのが一般的であるが、言語にコードが付けられる前に作られた言語データには言語コードが付与されていなかった。世界には数千種類の言語があるが、このようなデータでは、言語の名称が言語を識別するシンボルとして使われており、その言語の名称の決め方にも基準がなく、あいまいに付けられていることが多くあり、言語の同一性が問題となる。貴重な言語資源であるが、この問題が、世界諸言語の言語特徴を横断的に解析し言語類型論的研究を進めていくうえでの利用における大きな妨げの一つとなっていた。

そこで、我々は言語名と共に言語系統情報を活用することによってあいまいさを解消し、同一性を高精度で判定するアルゴリズムを開発した。これにより、判定作業を自動化することができ、大量の言語データの判定が可能となった。これは、言語系統分類情報が存在している場合にのみ適用できる。ところが、その系統分類情報は言語学者による手作業によって見いだされたものであるため、系統分類情報がごっそり欠落していることがしばしばあり、これが本アルゴリズム適用時の大きな障害の一つとなっている。本研究では、この問題を解決するため、系統分類情報の自動補完を可能とする新しい手法を開発する。これには、以下に述べるように、基礎語彙に基づく言語間関係の推定法を用いる。

２．研究の目的

言語の変化・変遷を系統樹モデルとして仮説を立て、分子系統学の手法を応用し、言語系統樹を自動的に生成する。このことにより、欠落されている言語系統分類情報の自動補完を目的とするアルゴリズムを開発する。発音の近さの度合いを考慮した基礎語彙の類似度の導入によって、言語間距離の計算手法を確立し、基礎語彙に基づく言語系統樹の生成手法を開発する。さらに、基礎語彙以外の言語属性も取り入れ、多次元的尺度に基づく言語系統樹の生成手法を開発する。これを基に、言語間の遠近関係をより高精度に推定できるアルゴリズムを開発し、言語系統分類情報の自動補完の目的を達成する。

３．研究の方法

(1) 分子系統学を応用した言語系統樹生成

本研究では、基礎語彙のデータとしてはASJPプロジェクトというヨーロッパを中心に活動を展開している言語研究団体が提供しているデータを用いる（図1）。

No.	1	2	3	4	5
Consepts	I	you	we	one	two
EASTERN_ARMENIAN	yes	du	menkh~	mek	yerku
ENGLISH	E	yu	wi	w3n	tu
FRENCH	i3	ti,vu	nu	oe*	de
HINDI	mai*	tu,ap	ham	ek	do
LATVIAN	es	tu,yus	mes	viens	divi
LITHUANIAN	aS	yus,tu	mes	vienas	du,div
NEPALI	ma*			ek	
PERSIAN	man	to	m3	yek	do
RUSSIAN	ya	t3,v3	m3	ody~in	dva
SERBOCROATIAN	ya	ti,vi	mi	yedan	dva
SPANISH	yo	usted,tu	nosotros	uno	dos
STANDARD_GERMAN	iX	du	vir	eins	zvai
WELSH	mi	ti	ni	3n	da3,du3
IRISH_GAELIC	mi~e	tu,hu	midy~,Si5	in	xa,da

図 1: ASJP 発音記号データ（イメージ）

遺伝系統樹の作成手法を応用し、基礎語彙データからアラインメントを行った行列を作って言語間距離を計算する。ASJPの提供しているデータは各言語ごとに40 語の基礎語彙の発音記号を、アルファベットと数字から構成する文字列を用いて記述している。編集距離を用いて言語間距離を計算するが、これには２つの方法で行ってみた。方法１では、図1に示す言語ごとの40語彙の発音記号をつなげ、一つの文字列とみなし、それらの文字列間の編集距離を求め、言語間距離とする。方法２では、同じ意味の語彙の発音記号文字列間の距離を算出し、40語の距離の平均値を言語間距離とする。算出した距離行列（図2）に基づき、近隣結合法を使い言語系

統樹を生成する。系統樹の可視化およびその他の計算用ツールとしては「R」を使用した。インド・ヨーロッパ語族の一部の言語を用いて実験を行い、言語学者が行う言語系統分類との異同についての考察を行い、この系統樹生成方法の有効性をある程度確認できた。

言語	1	2	...	...	m
Language 1	0	$D_{1,2}$	...	...	$D_{1,m}$
Language 2	$D_{2,1}$	0			$D_{2,m}$
⋮	⋮		⋱		⋮
⋮	⋮			⋱	⋮
Language m	$D_{m,1}$	$D_{m,2}$	...	...	0

図 2: 言語間距離行列（イメージ）

(2) 言語間距離計算モデルの提案

基礎語彙の類似度を計算するため、文字列類似度手法として、編集距離以外に、Jaro-Winkler 距離について検討を行った。

また、言語間距離の計算において、基礎語彙に加えて、言語の形式的な特徴を表す言語素性も要因として取り入れることにした。そのため、まず言語学分野で研究されている多くの言語素性のなかで、どのような言語素性がより強く言語の分類に影響を与えている可能性があるかについての分析を行った。言語を特徴づける言語素性のデータとして、一般的に公開されている WALS(The World Atlas of Language Structures)を使い、ランダムフォレストによる特徴量抽出の手法を応用して言語素性の重要度を計算し、言語分類に影響をもたらすと考えられる素性の選択を行った。そのうえで、編集距離および Jaro-Winkler 距離のそれぞれに言語素性を組合せて基礎語彙間の類似度および言語間距離の計算を行い、実験を行った。この二つの方法のいずれについても言語素性を考慮することが有効といえるような結果が得られた。なお、Jaro-Winkler 距離が編集距離に比べてより効果的である、またはその逆の結論を示す結果は得られていない。

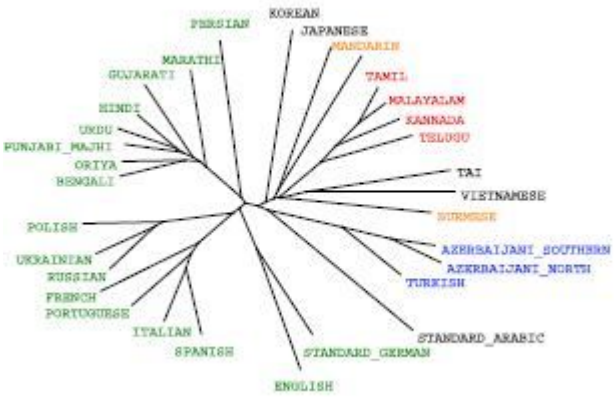

図 3: Jaro-Winkler 距離と言語特徴を組合わせた分類

さらに、基礎語彙の類似度ならびに言語間距離の計算精度を上げるため、発音記号間の相関関係の定量化について調査・研究を行った。ある特定の意味についての基礎語彙の先頭と末尾の子音の変化ついて、ある語派における祖語から各派生言語への変化の数をカウントし、ファイ（φ）係数を使い相関係数を算出する手法を提案した。ただ、ASJP 基礎語彙データが提供しているインド・ヨーロッパ語族のゲルマン語派のデータを使い実験したところ、強い相関を示す音声記号の発見までには至らなかった。

(3) 言語変化・変遷における波紋説および GIS による言語系統分類考察手法の導入

本研究は言語変化のモデルとして系統樹説に基づいて基礎語彙による言語系統樹作成の手法の提案を模索してきた。系統樹説は異なる言語間に共通の特徴が見られた場合、その言語同士の親縁関係をその祖語の同一性により説明しようとし、言語接触による影響をまったく考慮しない仮説であり、現在一般的に採用されている。一方、本研究では採用してこなかった波紋説がある。連続的な人の流動がある地域間においては、それぞれの言語同士が影響しあい、祖語の同一性だけでは説明のつかない言語変化が起こる場合が考えられ、波紋説は言語間の接触による言語の変化・変遷を前提とする学説である。実際の調査で言語接触による影響が予想以上

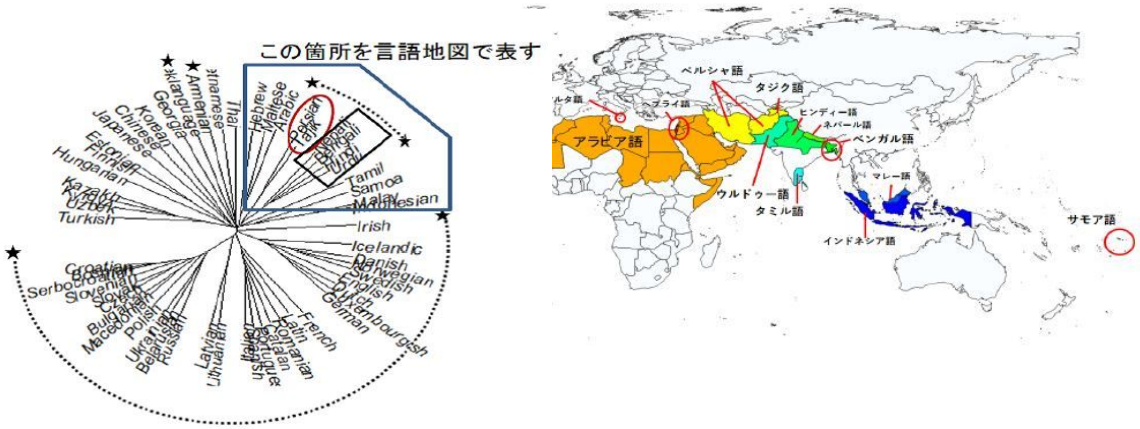

図 4: 単語発音記述データを組み合わせた言語系統樹および GIS による考察

に大きく、無視できないことがわかった。そこで、系統樹説に波紋説を取り入れる形で研究を進めることにした。波紋説に基づく言語分類を試みるための単語発音記号データが見当たらないため、独自の方法で単語発音記述データを作成し実験に供した。このデータは FORVO というWeb サイトを利用し、出来る限り多言語の、多様な意味の単語の発音をローマ字で記述して作成された。この単語発音記述データと基礎語彙データを組み合わせて系統樹を作成し、基礎語彙データのみを用いて作成した系統樹との比較をするため、GIS（地理情報システム）を導入し考察を行った。一部の言語に止まるが、言語接触を前提とした波紋説を取り入れたほうが、言語間の関係をより妥当に解釈でき、系統樹説では見いだせない言語間関係を導出できる可能性があるとの結論が得られた。

(4) 文法データの形式化

言語の文法構造にも着目した。文法を言語分類における要因として処理できるようにするため、まずは言語文法の形式化をターゲットとした。文法理論としては Fillmore の格文法を用いた。Fillmore の格文法は、動詞を中心的に捉え、文の意味構造を格構造によって表現することを提唱し、英語の一部の文型の深層構造から表層構造への変化規則を与えている。我々は文脈自由文法の生成規則を与えることで、Fillmore の格文法による文の生成（表層化過程の記述）を試みた。４言語（日・英・仏・中）の同一の意味の文について深層構造の表層化過程の木を作成し、一部文型の生成規則を与えた（表 1)。Fillmore の格文法による表層化過程が文脈自由文法だけでは記述できないため、さらに単一化文法を適用し、文脈自由文法の補強を行った。これにより、言語系統分類のための、言語特徴としての文法のデータのより一般的な形式化が可能となった。

表 1: ４言語の自動詞文における生成規則

英語	日本語	フランス語	中国語
S → M P	S →M P	S → M P	S → M P
S → O M P	S → O M P	S → O M P O	S → O M P
S → NP M P	S → NP M P	S →NP O M P	S → NP M P
S → NP P	S → NP P	P → V O	S → NP P
P → V O	P → V O	P → V	P → V O
P → V	P → V	O → K NP	P → V
O → K NP	O → K NP	O → se	O → K NP
NP → d N	M → Past	NP → d N	M → Past
M → Past	V → 開く	M → Past	V → 開
V → open	V → 開いた	M → Past etre	V → 開了
V → opened	K → ϕ	M → est	K → 肥
K → ϕ	NP → ドア	V → ouvrir	NP → 門
d → the	NP → ドア-が	V → ouverte	
N → door		K → ϕ	
		d → la	
		N → porte	

４．研究成果

(1) 分子系統学分野の遺伝系統樹の作成手法を言語分野に応用し、言語系統樹の自動生成法を提案し、言語系統樹の自動生成を可能にした。

(2) 基礎語彙に基づく言語間距離計算モデルを提案した。さらに、ランダムフォレストによる

特徴量抽出の手法を応用して言語素性の重要度を計算することによって、言語分類に影響度が高いと思われる言語素性も考慮した言語間距離の計算モデルを提案した。これにより、言語系統樹生成による世界諸言語の自動系統分類研究のひな形が形成された。

(3) Fillmore の格文法を用いて個別言語の文の表層化過程の生成規則を記述し、単一化文法を適用することにより、文法記述における一致の問題を解決し、文法データの形式化を可能にした。

５．主な発表論文等

〔雑誌論文〕（計 3 件）

① R. Wu, Y. Matsuura and H. Matsuno, “On Generating Language Family Trees based on Basic Vocabulary”, Proc. ITC-CSCC2015, 査読有, pp. 272-275, 2015.

② R. Wu, Y. Matsuura and H. Matsuno, “On Computation of Association Coefficients of Phonetic Symbols Based on Basic Vocabulary”, Proc. ITC0-CSCC2016, 査読有, pp. 285-286, 2016.

③ H. Suiji, R. Wu, H. Matsuno, “Classification Method for Comparison of World Languages by Word Pronunciation-based Description and Basic Vocabulary”: Proc. ITC-CSCC2018, 査読有, pp.588-591, 2018.

〔学会発表〕（計 3 件）

① 松浦佑哉、呉靭、松野浩嗣、基礎語彙に基づく言語系統樹生成のための音声記号間の重み付け、電子情報通信学会信学技報、Vol. 115, No. 480, pp. 33-36, 2016.

② 松浦佑哉、呉靭、松野浩嗣、基礎語彙と言語素性に基づく言語分類方法の提案、電子情報通信学会信学技報、Vol.116, No. 525, pp.45-54, 2017.

③ 原添修司、神谷捺希、呉靭、松野浩嗣、格文法を用いた言語系統分類のための形式化手法の提案、電子情報通信学会信学技報、Vol.118, No. 499, pp.69-74, 2019.

６．研究組織

(1) 研究分担者

研究分担者氏名：乾　秀行

ローマ字氏名：INUI HIDEYUKI

所属研究機関名：山口大学

部局名：人文学部

職名：准教授

研究者番号（8 桁）：10241754

研究分担者氏名：松野　浩嗣

ローマ字氏名：MATSUNO HIROSHI

所属研究機関名：山口大学

部局名：大学院創成科学研究科

職名：教授

研究者番号（8 桁）：10181744

(2)研究協力者

研究協力者氏名：
ローマ字氏名：

※科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属されます。