様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ６ 月　 ８ 日現在

機関番号：１５５０１
研究種目：挑戦的萌芽研究
研究期間：2011～2014
課題番号：２３６５０１２９

研究課題名（和文）言語名ゆれと系統分類ゆれを考慮した世界言語系統分類の類似性判定アルゴリズムの開発

研究課題名（英文）Development of an algorithm to decide similarities of world languages with considering differences in language names and classifications

研究代表者
松野　浩嗣（Matsuno, Hiroshi）

山口大学・理工学研究科・教授

研究者番号：１０１８１７４４

交付決定額（研究期間全体）：（直接経費）　　2,800,000円

研究成果の概要（和文）：２つの異なる言語の同一性判定問題は、世界言語照合の研究の主な問題に一つである。本研究では、言語名類似性と言語分類類似性の２つの尺度を用いてこの判定を行う行うアルゴリズムを提案し、実験により88%の正解率の照合結果を得ることができた。さらに、兄弟情報を考慮することで、この正解率を向上させることができた。
この手法のさらなる改良、すなわち、これら２つの尺度のうち、どちらか一方が完全な照合であるときにでも不一致とする問題点を解決するため、さらに２つの基準である、言語情報と兄弟情報による類似性と言語名と分類情報による基準を定め、その効果を実験的に確認した。

研究成果の概要（英文）：Identification of language correspondences between two different sets of language data is one of the main problems in world's languages matching. We proposed a method which enables this identification by using two measures of language name similarity and language classification similarity, having succeeded in searching 88% languages included in one set of language data that relate to another set of language data. We further improved the accuracy by taking into account brother information in a language classification tree.
Their method still has a problem, that is, their method gave an inappropriate decision even if either of these two similarities has a complete matching. To address this problem, we define two kinds of new measures: one is a similarity of languages based on brother information, and the other is a language general similarity that integrates the similarities of language name and language classification. The effectiveness of this method was confirmed by experiments.

研究分野：情報科学

キーワード：言語系統分類

１．研究開始当初の背景

世界中で多くの言語が用いられており、その数は五千を下らないといわれている。この言語の多様性は、歴史的条件や地理的条件を背景にもたらされたものである。言語の分類を行うことは人類の歴史、文化、習慣などの理解につながるため、国内外で多くの言語学者が世界言語の系統分類の研究に取り組んでいる。しかし、個別の言語学者が独自の思考によって分類を行っているため、多様な世界言語系統分類が存在し、その類似性と相違性を明らかにすることが言語研究の大きな課題となっている。

２つの異なる言語系統分類の類似性を検証するためには、言語名の照合作業が必要である。言語は文字列によって表現されるものであるから、文字列照合の道具としてコンピュータを利用し、系統分類の手助けとして活用している取り組みは既になされている。しかし、言語学者の作成した系統分類には、単純に計算処理することのできない、同一言語名の重複出現、言語名のゆれ、系統分類のゆれ等、の問題があり、最終的には人の目によって多量のデータの類似性を確認せざるを得ない状況となっている。

これらを解決するためには、分類を行った言語学者の思考の差異の部分にまで入り込む必要があり、この問題解決は容易ではない。本研究では、この問題に取り組み、世界言語の系統分類の類似性を高効率で検証できるアルゴリズムの開発を試みる。

２．研究の目的

言語学者による言語データでは通常、言語名によって言語を識別している。しかし、１つの言語に対して近隣民族が様々な呼称を用いるため、違った呼び名が複数存在していることがよくある。また、世界諸言語に関するデータでは、データを編成するうえで用語に関し統一した基準となるものが存在しているわけではなく、各々の言語学者は自らの立場により言語名などを用いている。これらのことより、世界諸言語に関するデータでは、言語の名前だけでは言語を識別できないケースが多い。

複数の言語学者によって編成された言語データに含まれる言語の同一性の判断は、言語の一意識別子である言語コードが付与されていれば問題とはならないが、実際には言語コードの付けられていない言語は多い。一般に、世界の言語数は千単位にものぼるため、手作業で言語を特定するのは、莫大な作業量を要するうえ、専門知識も必要とし、大変困難である。

本研究の目的は、言語コードの付与されていない言語の同一性を判定する情報科学的手法を開発することである。その第一段階として、言語データが与えられているものと与えられていないものの２つの言語データの同一性判定アルゴリズムを開発する。すなわち、言語データが付けられていない言語について、自動処理によって言語データが付与されている言語を見つけ出す手法を提案し、なるべく多くの同一言語ペアを検出することを目的とする。

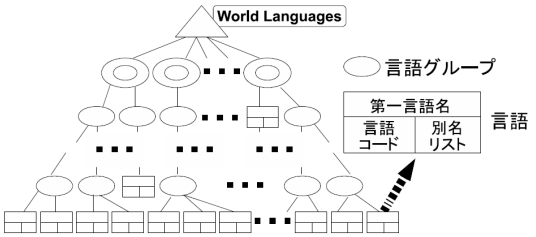

図１ 言語系統木

３．研究の方法

本研究では、アプローチとして言語系統木(図 1)という概念を導入した（言語系統木では，言語は最下位のレベルに位置するリーフノードとなる）。これは、言語名だけでは言語の同一性を判定するための情報が足りないため、別の角度からの情報として、言語系統分類を取り入れるためである。言語系統分類に関するモデルはいくつか提唱されてきたが、そのなかの１ つとして，系統樹モデルがある。系統樹モデルは同じ語族に属する言語は、はるか過去に話されていた１ つの言語から分かれて発展してきたと主張し、言語の分化の過程を一本の樹となる系統樹にたとえている。１ つの系統樹は１ つの語族に含まれる言語から構成され、言語と言語の間の親族関係を表している。本研究では、系統樹モデルに基づき、世界諸言語のデータ構造を系統樹の森となる言語系統木として定義した。

言語系統木の導入により、2 つの表形式の言語データに含まれる言語同一性判定の問題は 2 つの木構造のリーフノードの間のマッチング問題として転化した。また、言語系統分類も言語名と同様に曖昧な性質をもつため、本研究では言語名の類似度と言語系統分類の類似度という概念を提案し、言語類似性の定量化を試みた。

木構造上でのデータマッチングに関する研究は広く行われている。研究対象の概念を明示的に表現し、それらの関係を体系的に記述したオントロジーを構築し、異なるオントロジー間の対応関係を見つけ出すオントロジー・マッピングや、木編集距離などを利用した木構造パターン・マッチングなど数多くの手法が提案されている。しかし、言語学における言語系統分類の学問分野自身がまだ確立された体系を樹立するまでに至っていないため、オントロジー構築が困難である。さらに、本研究では 2 つの言語系統木に含まれる言語（リーフノード）のマッチングのみに限定しており、木構造全体のマッチングまでは考慮する必要がないことから、それらの手法は本研究に必ずしも適しているとはいえない。

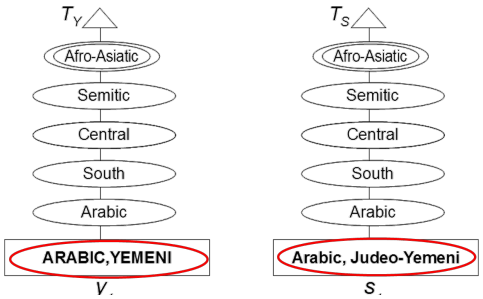

図 2 言語名に変化がある言語

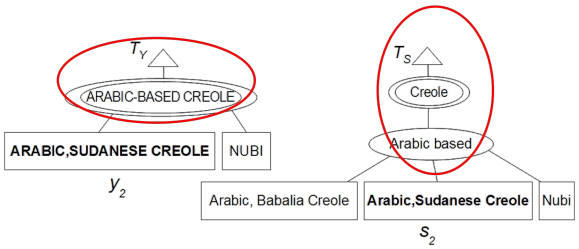

図 3 言語系統分類に変化がある言語

また、木構造上でのデータマッチングに関するテーマではないが、近年ソーシャルネットワークに関連して、人の名前を特定する人名マッチングの研究が盛んである。人名は、本研究が扱う言語名に類似している。オントロジー・マッピング、木構造パターン・マッチングおよび人名マッチングなどに共通して用いられている基本手法がある。それは、文字列類似度に基づく手法である。本研究では編集距離を基本とし、文字列類似度計算構造化手法である Monge-Elkan 法を言語名の類似度計算に取り入れた。

4．研究成果

4.1 木構造と文字列類似度に基づく手法

言語系統分類の類似度についても定量化を行い、同一言語ペアについて、

①言語名と言語系統分類ともに変化がない言語

②言語名あるいは言語系統分類、またはその両方に変化がある言語(ゆれのある言語)。図 2 と図 3 がそれぞれ言語名あるいは言語系統分類に変化がある例である。

を検出する手法を開発し、独自に収集・整理した 2 種類の言語系統木データを用いてその有効性の検証実験を行った。その結果、これら 2 種類のデータの 88%の言語の同一性が判定できた。そのうち、52%は言語名の類似度と言語系統分類の類似度による結果であった。このことから、本研究で提案した言語名の類似度と言語系統分類の類似度とゆれのある言語の検出方法は効果的であることが確かめられた。

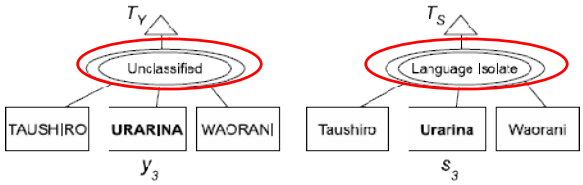

図 4 言語名が同じで、言語系統分類がまったく異なる言語

4.2 言語名と言語系統分類の総合的尺度に基づく手法

4.1 節に示した結果のように、言語名と言語系統分類の曖昧な性質に対して、文字列類似度に基づく言語名の類似度と言語系統分類の類似度を導入し、2 つの言語系統木に含まれる同一言語を検出するための手法を提案し、88%の言語について同一性の判定を行うことができたが、残りの 12%にも同一言語ペアが含まれていることが予測できる。

そこで、木構造と文字列類似度に基づく言語の同一性判定手法を以下に示す 2 つのアプローチにより発展させた。

(1)言語系統分類の類似性評価の改善を試み、兄弟情報を考慮した新しい言語系統分類の類似度を導入した。これは、ある 2 つの言語系統データのそれぞれに含まれる 2 つの言語について、互いに兄弟言語が存在し、その兄弟言語同士が同じ言語であるならば、この 2 つの言語も同一言語である可能性が高い、という考えに基づくものである。図 3 の y_2 と s_2 がその一例である。それらの兄弟言語の中には同じ言語同士が存在しており、ゆえにそれらも同一言語である可能性が大きいことは一般的に考えられることである。この兄弟情報は 4.1 節で述べた手法では考慮されていなかった。ここで新たに取り入れる。

(2)言語総合類似度という概念を新たに導入し、言語名の類似度と言語総合類似度に基づく言語の同一性判定の手法を提案した。これは、言語名の類似度または言語系統分類の類似度の一方が非常に高い場合、他方の類似度がそれほど高くなくても、同一言語である可能性が高いという考えに基づくものである（図 4 がその一例である。言語名はまったく同じである。しかし、言語系統分類の類似度が小さいため、4.1 節の手法では両言語の同一性が判定できない）。言語総合類似度では、4.1 節の言語名の類似度と言語系統分類の類似度を 2 段階で考慮する判定方法を改め、両方の類似性をともに考慮する。さらに、この言語総合類似度には、兄弟類似度も加味される。

この言語総合類似度の導入により、4.1 節で用いたのと同じ 2 つの言語系統データに対して、92%の同一言語を見つけることができた。すなわち、4.1 節の手法では 88%であったから、4%の検出率の増加を得ることができた。例えば、図 3 に示す例の場合は、4.1 の手法では同一言語として判定できないが、

兄弟情報を考慮することによりその同一性が判定できるようになった。このことから、兄弟情報を考慮した言語系統分類の類似度は、言語の系統分類の比較において、よりその特徴を捉えており、有用であることが確かめられた。

4.3 今後の展開

地理情報システム(GIS)は、多角的な時空間検索と分析の機能を持っており、世界諸言語の言語特徴を地図化することができる。この研究で 92%という高率で同一性を検出することができた２つの言語系統データを用いて、語順に着目した GIS 地図を作成し、世界諸言語分類の視覚化を図りたい。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計３件）

① Ren Wu, Hiroshi Matsuno, On parameter setting in identifying the same language involved in different language data, Journal of Robotics, Networks and Artificial Life, 査読有, Vol.1, No.2, 2014, 103－107.

doi:10.2991/jrnal.2014.1.2.1

② 呉 靰、乾 秀行、松野 浩嗣、言語名と言語系統分類の総合的尺度に基づく言語同一性判定、人工知能学会論文誌、査読有、第 28 巻 3 号 C，2014，320－334.
https://www.jstage.jst.go.jp/article/tjsai/28/3/28_320/_article/-char/ja/

③ 呉 靰、松野 浩嗣、木構造と文字列類似度に基づく言語の同一性判定、情報処理学会論文誌：数理モデル化と応用、査読有、Vol.3，No.3，2010，24－35.
http://ci.nii.ac.jp/naid/110007989981

〔学会発表〕（計１件）

① Ren Wu, Hideyuki Inui, Hiroshi Matsuno, New measurement of similarity of language classification, Proc. International Technical Conference on Circuit/Systems and Communications, July 16, 2012, Sapporo Convention Center (Hokkaido, Sapporo-city).

〔図書〕（計　　件）

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者

松野 浩嗣（MATSUNO, Hiroshi）
山口大学・大学院理工学研究科・教授
研究者番号：１０１８１７４４

(2)研究分担者

乾 秀行（INUI, Hideyuki）
山口大学・人文学部・准教授
研究者番号：１０２４１７５４

呉 靰（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