氏　名（本　籍）	**呉　軔**（中国）
生　年　月　日	昭和３９年１２月２６日
授　与　学　位	博　士(理学)
学　位　記　番　号	理工博乙第１２４号
学位授与年月日	平成２５年　８月　９日
学位授与の要件	学位規則第４条２項
研究科，専攻の名称	理工学研究科(博士後期課程)自然科学基盤系専攻
学　位　論　文　題　目	**言語系統木と文字列類似度に基づく言語同一性判定に関する研究** **(A Study on the Identification of the World's Languages Based on Languages Tree and String Similarities)**
論　文　審　査　委　員	主　査　山口大学教授　松野浩嗣 山口大学教授　朝日孝尚 山口大学教授　菊政勲 山口大学准教授　末竹規哲 山口大学准教授　乾秀行

【学位論文内容の要旨】

世界には数千種類の言語がある．世界の諸言語はそれぞれ違い，多様性に富んでいる．言語学者は，世界の諸言語にはどのような多様性が見られ，またその多様性の中にどのような普遍性が潜んでいるのかについて古くから探求してきており，言語類型論という学問体系ができている．近年はIT（情報技術）の著しい発展により，文理融合の新たな手法によってさらなる発見をもたらすことが期待されている．

ITを応用した言語類型論的研究では,特に言語特徴に関する情報が含まれている言語データが欠かせない．また，効果的に研究を展開するためには，しばしば他の言語学者が収集した言語資料をデータ化し，研究に使うことがある．その際，複数の異なる言語学者による言語データを一つにし，新たな言語データを生成してから使うこともよく行われる．

言語学者による言語データ（ここでは2つの表形式のデータを想定する）では，普通，言語名によって言語を識別している．しかし，1つの言語には，複数の名前が付いている場合がよくある．また，言語の別名の存在や表記ゆれなどが含まれているため，言語の名前だけでは言語を識別できないケースが多い．つまり，世界諸言語に関するデータでは言語の一意識別子が含まれていないことがあり，このことが，言語データをマッチングする際に問題となる．

言語の一意識別子として，国際標準化機構による言語コード（ISO639シリーズ）がある．この言語コードの標準化は1980年代から始まったが，以来頻繁にコード体系の変更が発生しており，コード体系設計上一般的に要求される安定性や恒久性が具備されているとはいえない．そのためか，言語コードは未だに，言語学者の間でコンセンサスが得られ，確立されたコード体系が共有され，標準として使われる段階に至っているとはいえない．

言語コードが付与されていない価値ある言語資料は数多く存在する．言語同一性の問題が障害となり，言語研究に活かせないのならば，それは大変残念なことである．人類最大の文化遺産ともいえる言語に関する資料を研究に活かせるようにすることは，重大な意義を持つ．一方，世界諸言語に関する言語データは言語数が千単位にのぼるため，手作業によって言語を特定するのは，莫大な作業量を要するうえ，専門知識も必要とするため，大変困難なことである．そこで，本研究では異なるデータ中の言語同一性をコンピュータ自動処理によって判定することに取り組む．特に，2つの異なる学者による表形式の言語データの一方に言語コードが付けられていない場合における言語同一性の問題に焦点を当て，解決を図る．

本研究が目指す問題解決は，著者の知る限りにおいて，今まで研究が行われていない．本研究では，アプローチとして言語系統木という概念を導入する（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する．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する．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る．

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．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．

文字列類似度にも多くの手法がある．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 Monge-Elkan 法を言語名の類似度計算に取り入れる．また，言語系統分類の類似度についても定量化を行い，同一言語ペアの検出法についても提案を行う．さらに，実験を行い，その手法の有効性を示す．本論文は以下のように構成される．

第1章では，言語学的な背景について述べたうえで，本論文の目的および構成を示す．

第2章では，言語データの例を示し，言語同一性判定の問題点を分析する．

第3章では，準備として，系統樹モデルおよびオントロジー・マッピングなどの関連研究について述べる．また，文字列類似度の指標である編集距離と最長共通部分列および文字列類似度計算構造化手法 Monge-Elkan 法などについて述べる．

第4章では，言語系統木について定義したのち，XML を用いた言語系統木データとその構築について述べる．

第5章と第6章では，言語同一性判定の手法として2つの手法（手法Ⅰと手法Ⅱ）を提案する．第5章では，手法Ⅰとして，まず木構造に基づき，言語名と言語系統分類についてゆれのない完全一致言語の検出法について述べる．次に，言語名や言語系統分類についてゆれのある言語にも対応した木構造と文字列類似度に基づく手法を提案する．実験の方法と結果を提示し，考察を与える．

第6章では，まず手法Ⅰの問題点を指摘する．それは，2つの言語系統木のそれぞれに含まれる2つの言語の言語名，または言語系統分類のどちらか一方が完全に一致しているにもかかわらず，そのことがその2つの言語の同一性判定にまったく考慮されていない，ということである．その問題点をカバーできる言語名と言語系統分類の総合的尺度に基づく手法を提案する．さらに実験の方法と結果を提示し，考察を与える．

第7章では，言語同一性判定問題に関する今後の課題について述べたのち，本論文をまとめる．

【論文審査結果の要旨】

世界には数千種類の言語があると言われている．言語類型論は世界諸言語の様々な言語特徴を考察し，言語の普遍性ならびに多様性を見つけ出そうとする学問である．言語学者は通常，言語データを言語名によって識別している．しかし，一つの言語に対して近隣民族が様々な呼称を用いるなど，複数の言語名が存在し

ている場合がよくある．さらに，名称を決める基準がないため，個々の言語学者が自らの視点や立場で言語名を定めることも珍しくない．そのため，現存するデータだけでは言語を識別できないケースが多く，言語特徴の横断的分析を困難にしている．

言語の一意識別子として，国際標準化機構による言語コード(ISO639)がある．情報科学をベースにした言語研究の始まりに伴って,編成または発表された言語資料にこの言語コードが付与されるようになってきた．しかし，依然，言語コードが無い資料も多く，独自のコードを付ける言語学者もいるのが現状である．現存する多くの言語資料を，言語コードの体系に基づいて統一すれば言語類型研究に役立つ資料となるのは間違いないが，数千に及ぶ言語の照合を手作業で行なうことは非常に困難であるため，未だこの言語データ統合は実現されていない．

本論文では，この言語データ統合作業を計算機を用いて実施するための方法を提案し，実際の言語データを用いて提案手法の効果を検証した結果を報告している．具体的には，言語コードをもつSilGIS-Dataとこれを持たないYamamoto-Dataに含まれている言語の同一性判定を，以下に示しているような，新しく提案したアルゴリズムを用いて実施している．

(1)言語系統木の概念を新たに導入し，Monge-Elkan法に基づく言語及び言語系統分類の類似度を提案したうえで，同一性判定アルゴリズムを開発している．このアルゴリズムは，言語名の一致度だけでなく，言語系統分類の一致度も同一性判定の基準にしており，これにより，Yamamoto-Data の約88%の言語がSilGIS-Dataに含まれる言語と同一であることが検出できている．

(2)上記(1)のアルゴリズムでは，言語名と言語系統分類のどちらかが完全に一致しているにもかかわらず，同一言語ペアとして検出されないという問題点がある．これを解決するため，語族，親，兄弟の情報を考慮した言語系統分類の類似度，並びに，言語名類似度と言語系統分類類似度の加重平均を取った言語総合類似度を導入し，(1)のアルゴリズムを拡張することで，検出率を92．3%まで高めることに成功している．

提案された言語同一性判定手法は，個々の言語及びそれらの連関等の言語学的意味を考慮することなく，言語名の文字列パターンと言語系統の分類パターンの類似度を機械的に比較することだけで，一致度を判定している．言語類型論の研究では，このように言語の意味を考えることなく分類を行なうアプローチはとられておらず，本論文で提案している情報科学的手法に基づく分類手法は極めて独創性の高いものである．さらに，90%を超える検出率を得ていることから，その有効性も高いと言える．

この論文では，地理情報システムを用いた言語類型論研究の新たな可能性についても言及している．地理情報システムは，多角的に整理された時空間情報の高度な検索と分析を可能とするシステムであるが，本論文の提案手法により統合された言語データを利用すれば，これまでに見落とされていた事実の発見が期待され，新たな知見を与える手法として確立する可能性がある．

公聴会における主な質問内容は，未検出の言語名がもつパターンの特徴とその検出のための手法の改善の見通しについて，語ペアによる言語名類似度の規定方法とその同一性判定への影響について，及び言語名の同一性判定の基準に関するものであった．いずれの質問に対しても発表者から的確な回答がなされた．

以上により，本研究は独創性，信頼性，有効性，実用性ともに優れ，博士（理学）の論文に十分値するものと判断した．

審査委員会の副査より，

1．文字列比較を基本とした関連研究との比較を行ない，言語系統分類における本研究の位置づけを明確にすること．

2．言語系統分類研究の課題を必要な背景とともに述べ，さらにその課題を情報科学的に解決する意味について分りやすく説明すること．

の2点の課題が課されたが，口頭試問において必要な資料を示しながら，適切な回答がなされた．また，語学能力については，第一著者の国際会議英文論文を3件執筆し，英語での口頭発表の実績があることから，十分な外国語能力を有するものと判断された．以上，論文内容及び審査会，公聴会での試問応答など総合的に判断して，最終試験は合格とした．

なお，主要な関連論文の発表状況は下記の通りである．（関連論文計7編，参考論文なし）

1)呉靷，山本秀樹，乾秀行，杉井学，松野浩嗣，語順地図作成に必要なデータ及び語順地図に現れる語順分布，一般言語学論叢，第10号，pp．31-49，2008．

2)呉翰，乾秀行，松野浩嗣，Ethnologue15th 言語属性データと系統データの生成及び言語同定における利用，コンピュータ＆エデュケーション，Vol．25，pp．70-73，2008．
3)呉翰，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，Vol．3，No．3，pp．24-35，2011．
4)呉翰，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，28巻3号，pp．320-334，2013．