# 類型論データベースを用いた人類史解明のための言語類似法の構築

松前ひろみ[1] 中井瑞[2], 長谷武志[3,4]
[1]東海大学医学部 [2]慶應義塾大学大学院政策・メディア研究科 [3]東京医科歯科大学 [4]慶応義塾大学薬学部
matsumae.hiromi.g@tokai.ac.jp

## 概要

本研究では、人類社会における言語の多様性とその歴史を解明する目的で、言語類型論のデータベース用いて、文法の類似性に基づいた解析法を新規に考案した。既知の先行研究の課題であった、(1)欠損値を含み、かつ多次元であるような複雑な文法要素のデータ構造の前処理を自動化し、(2)ペアワイズの相関解析に基づき言語間の類似性の推定を行い、(3)文法要素のクラスタリングを行うことで、非独立性要素が要因となり生じる類似度の誤差を推定し、(4)最終的に、木構造に依存しない可視化法としてスパイダーチャートで類似性を採用した。そして北東アジアとヨーロッパという言語多様性の観点から対称的な2つの地域の言語に絞った予備解析を行った。

## 1 背景

ヒトの言語はヒトだけが持ちうるユニークな文化である。しかし世界には 8000 以上の言語が知られており、それらは420以上の言語族に分けられている(1)。つまり人類集団の中で言語は、ヒト言語としての普遍性を保ちつつも、多様性を生み出している。

これら多様な言語の歴史的関係性を定量的に推定することは容易ではない。例えば、基礎語彙に基づく歴史言語学的アプローチは、語族内の関係性の再構成には有力な手法である。またある語族に属する言語とその話者集団の遺伝的関係についても学際的なテーマとしてよく研究されている。例えば、西ユーラシアにおけるインド・ヨーロッパ語族の言語と話者それぞれの起源と拡散史の解明は、ハイスループットな個人ゲノム解析技術の飛躍により、欧米では競争が激しい研究テーマの１つである。しかしこの方法は、借用語の関係などを除けば語族間の関係性の推定には適さない。語彙の他では、音素は語族を超えた比較が容易であるものの、音素の類似性は、近年の言語接触により生じた変化を反映しやすいと考えられており、古い歴史的関係性を保持していない可能性がある(2-3)。したがって、世界中の民族集団の高品質のゲノムデータが得られるようになっている中で、言語族数が少ない西ユーラシアに比べ、多言語族に跨がる東ユーラシアの言語-遺伝子の関係は、大きな後れを取っている。

一方、言語類型論では、文法には言語族間の関係性が反映されているという仮説もあり、文法をデータベース化し、言語の類似性の評価に利用しようという動きにつながった(4)。それら言語の類型論的な特徴は、World Atlas of Language Structures(WALS)といったデータベースに集約されてきた。近年、WALSの文法要素やその他音素などのデータベースを統合し、2,500 以上の言語データを含む新しい言語類型論データベースAUTOTYPが2017年に考案された(5)。

AUTOTYP を用いて言語間の歴史的関係性の推定を行った先行研究として、日本、朝鮮半島、シベリア、北極圏を含む北東アジア 11 言語族に対して、言語の３つの要素（語彙・文法・音素）とゲノムと音楽（歌）を比較した研究が挙げられる(6)。これら地域の民族集団のゲノムはよく研究されており、語族が異なっていても民族集団間にどの程度の遺伝的な連続性があるか知見が集積されているため、もし言語の中にそうした類似性を示す要素があれば、ゲノム史と相関するだろうというアイデアである。この研究では、言語やゲノムに対して、それぞれ言語/民族間の距離を距離行列に変換し、統計的に分析することで文法の類似性とゲノム史の間に有意な相関を見出した（逆に語彙や音素とゲノム史の間には統計的な関連がなかった）。このことは語族間関係の推定に文法が利用可能であることを示唆している。

この先行研究は、文法のデータ解析にいくつかの課題を残している。第一に AUTOTYP 上の文法要素のデータ構造は多次元データであると考えられ、加えて研究者人口の少ない言語では欠損値を多数含むといった複雑な構造になっており、利用にあたり言

語類型論の高い知識が要求されることが挙げられる。例えば、どの文法要素同士が独立なのか、それとも関連があるのかは、現在の AUTOTYP は対応づけできていない。従って、AUTOTYP を用いて、北東アジア以外の地域の語族間の関係性を分析するには容易ではない。第二に、言語間の類似性の可視化の課題である。先行研究(6)では、主成分分析や遺伝子系統解析で用いられるネットワーク系統樹 Neighbornet を用いることで言語間の関係性を可視化した。しかし系統解析の基本である木構造は、人間には直観的に理解しやすいものの、言語間の関係を単純化しすぎる懸念がある(7-8)。語族間の定量解析は始まったばかりであり、語族同士の系統関係についてはそもそもよく分かっていないため、こうした関係性は様々な方法で検証されるべきである。さらに木構造以外で言語のような複雑な変化を示す文化の類似度を可視化できれば、言語に限らず、幅広い文化的指標に適用できる可能性が生まれる。

そこで本研究では、最終的に言語史研究に応用する目的で、AUTOTYP のデータを、事前の知識をなるべく必要としない形で自動的に抽出・解析し、木構造以外の形式で可視化した。本研究では、汎用性を高めるため、perl および R を用いたパイプラインを構築した。

# 2 AUTOTYP データの抽出と前処理

## 2.1 地域集団データセットの作成

本研究は、言語間の関係を最終的に人類史に照らし合わせることを目的としている。そこで、テストデータセットとして 2 つの言語セットを作成した。一つ目は北東アジアに焦点を当て、先行研究(6)の 14 言語に対して、より広域の 52 言語/23 言語族を含むデータセットを作成した。この 52 言語には、地域的に遠く、言語学的な知見からも外群的な（コントロール的な）役割を期待できそうな言語族も含めた。二つ目は言語多様性の低いヨーロッパのデータセットを作成した。AUTOTYP に記載されている地域情報(“Region”)から Europe を抜き出したところ、先印欧語も含む 107 言語/6 言語族が含まれていた。これらのデータセットを用いて以下の解析を行った。

## 2.2 文法要素の coverage の産出

各言語は、AUTOTYP 上に存在する全ての文法要素に対してデータを持っているわけではない。これは、例えば、特定の言語族にしかないような稀な文法要素の変数が存在するためである。例えば、アジアの特定地域にしかない稀な文法要素を使って、ヨーロッパの言語を分析するのは、多くの場合、ナンセンスであることが想定される。さらに研究者がほとんどおらず、言語記録が限られている消滅危機言語では、類型論的研究が十分になされていないケースは多い。実際、北東アジアにはアイヌ語を始めとした消滅危機言語が多い。しかし人類史的な観点からは、そうした稀少な言語こそが重要であるため、一律に排除することは避けたい。

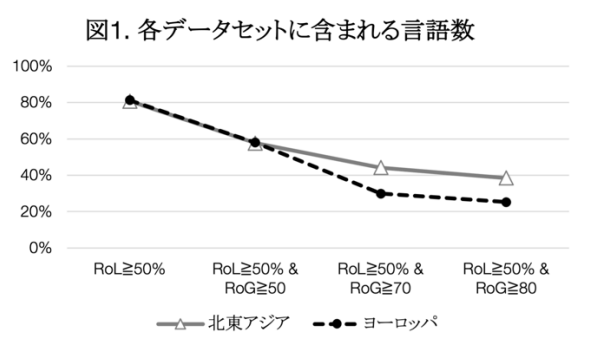
そこで、本研究では与えられたデータセットごとに欠損値の取り扱いを考慮し、任意の欠損率をもつデータのみを選択できるようにするため、2 つの閾値を設定した。1 つ目は、RoL=データセット内において各文法要素が共有する言語の割合(a Ratio of Languages sharing each grammatical factor in a dataset)である。RoL では、文法要素の頻度に閾値を設けた。例えば、RoL を低く設定すれば、その地域では稀である文法要素を、全体的な言語間の類似性の判定に用いることができる。逆に共通要素だけを使いたい場合には RoL を高くする。2 つ目は、RoG=データセット内において各言語が共有する文法的要素の割合(a Ratio of Grammatical factors sharing each language in a dataset)である。例えば、RoG を低くすれば、欠損値ばかりの言語を除去したい場合に用いることが出来る。RoG の低い言語を含めると、その後の言語間の関係性の推定精度が大きく下がるからである。

2 つの欠損値率を組み合わせた場合のデータ量について、テストデータセットで調べた。RoL を 80%以上とした場合、北東アジアもヨーロッパも言語が残らなかった。そこで、RoL を 50%以上の場合に絞り込むと、言語族の数が大きく異なる北東アジアでもヨーロッパでも 80%程度の言語が残ることが分かったが、文法要素数の割合は異なっていた(表 1)。さらに、RoL ≧50%の場合から、RoG に閾値を与えたときの言語数の影響を検討した（図 1）。その結果、RoG≧50%では、どちらの地域でも言語が残る確率は変わらなかったが、RoG を上げると、ヨーロッパの言語は北東アジアに対して言語が絞られる傾向があった。このことは、ヨーロッパのデータセットの特性に由来すると考えられる。ヨーロッパの言語はインド・ヨーロッパ語族が 88 言語(82%)を占めており、しかもその内訳は方言に近いような関係の言語も多い。そのため、高い RoG ではインド・ヨーロッ

パ語族に偏った要素が抽出されると考えられる。なお、RoG の値の選択では、文法要素の数は変わらなかった。このように、データセットに適した欠損値の扱いが必要である。

**表 1. RoL≧50%の言語数と文法要素数**

| | 元のデータセット | | RoL≧50% | |
|---|---|---|---|---|
| | 言語数 | 文法要素数 | 言語数(%) | 文法要素数(%) |
| 北東アジア | 52 | 506 | 42(80.8%) | 52(10.3%) |
| ヨーロッパ | 107 | 437 | 87(81.3%) | 19(4.3%) |




## 3 言語間の類似性解析

言語間の類似度解析を行うために、One hot vector encoding を行い言語の文法要素データベースの数値化を行った。この数値化により、各言語の文法要素の情報は、文法要素の存在の有無を 0 と 1 で表したバイナリベクトルで表される。ただし、言語要素の有無に関する情報がない場合、null とした。今回、このバイナリベクトルを用いて、以下の様に、言語間のペアワイズな類似性解析を行った。

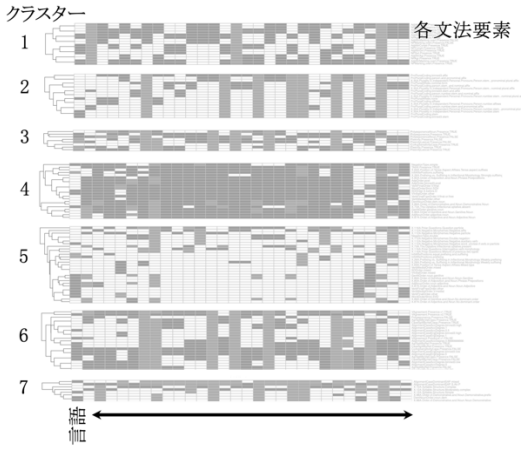
２つの言語ペアに対して、存在の有無の情報を有している要素（null でない要素）のみに着目し、バイナリベクトルの一致率を計算した。この一致率の計算を全ての言語ペアに対して行い、一致率の行列を作成した。その後、各言語ペアの全言語ペアに対する相対的な類似度を調べるために、この一致率の行列からペアワイズに相関係数を算出した。この相関係数を、各言語ペアの類似度として用いた。

## 4 文法要素間のクラスタリング

言語の文法要素において、複数の文法要素が強い関連性を持つことがある（非独立性文法要素クラスター）。このような文法要素クラスターは、言語間の類似度の計算に強いバイアスを生じさせる原因となる。例えば、ある２言語が非独立性文法要素クラスターを共有する場合、この２言語の類似度は、相対的に高く見積もられてしまう。

そこで、文法要素のクラスタリングを行い（図 2）、このようなバイアスが原因となって生じる類似度の誤差を推定した。具体的には、言語間の類似度解析と同様に One hot vector encoding によるデータの数値化を行い、そのデータに対して kmeans 法で文法要素間のクラスタリングを行った。最適なクラスターの数は、Gap 統計量をもとにして算出した。その後、各クラスターからランダムに１つの文法要素の抽出を行い、ペアワイズな言語間の類似度を算出した。この類似度の算出を 100 回行い、標準偏差を計算した。この標準偏差の値を用いて、非独立性文法要素によるバイアスから生じる言語間類似度の誤差を推定した。

**図 2. 文法クラスターの例。**RoL≧50%以上、RoG≧50%以上のとき、北東アジアセットでは、文法要素は7つのクラスターに分けられた。各言語（縦）に対して、横のラベルが各文法要素で、変数の値で色づけした。

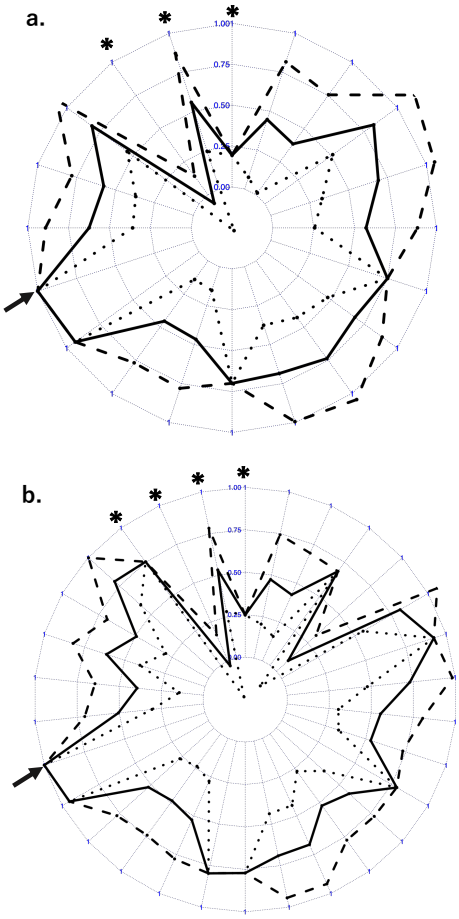



## 5 言語間の類似性の可視化

求めた類似度と誤差率に対し木構造に依存しない可視化法として、本研究ではスパイダーチャート（レーダーチャート）を採用した(図 3,4)。ある言語に対して、データセットに含まれる言語が円周上に配置されており、円の中心からの距離で類似度（直線）、その誤差率（点線）を示した。円の中心に近いほどその類似性が低く、円周に近いほど類似度が高くなる。北東アジアのデータセットにおける日本語の例

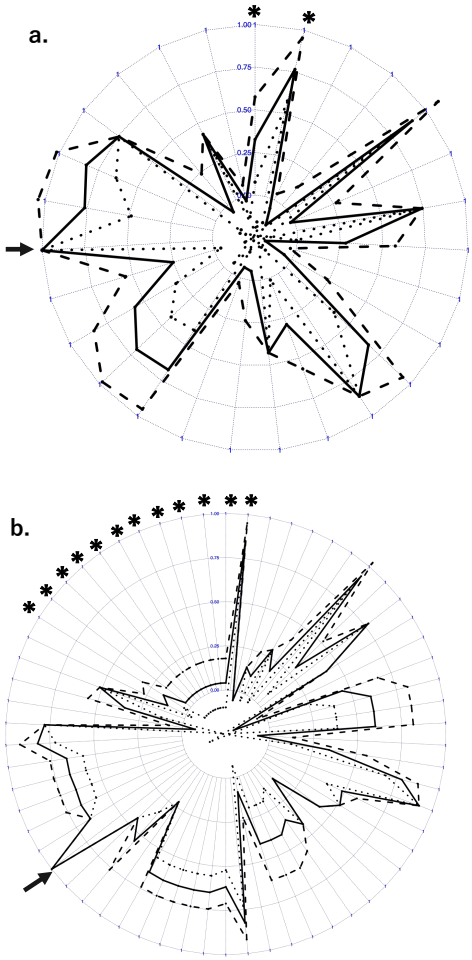
を図 3 に示した。日本語と日本語の類似度は 100%一致するので、円周上に位置する。日本語に対して、北東アジア外の言語は、50%程度の類似度があっても誤差率が高い言語や、誤差率は低いが類似度も25%と低い言語などがあった。こうした事例は、例えば、類似度が 60%以下だったり、誤差率の高い類似度を除外して、より信頼性の高い語族間の距離の分析に利用できる可能性がある。

**図 3.** RoL≧50%における、日本語とその他北東アジアの言語との類似度（a:RoG≧80%、b:RoG≧50%）。矢印は日本語自身。アスタリスクは北東アジア外の言語、無印は日本語以外の北東アジアの言語を示す。




一方、英語とヨーロッパの言語を比較したところ（図 4）、インドヨーロッパ語族以外の言語との類似性は、1 言語を除き 25%以下の類似度となった。同じインド・ヨーロッパ語族同士の比較では、類似度や誤差率には大きな幅があった。これには、前述の通り、ヨーロッパの言語の偏りが関係している可能性がある。また、研究の進んでいる語彙など他の指標に基づく距離との比較を行い、結果の妥当性を検証したい。

**図 4.** RoL≧50%における、英語とその他ヨーロッパの言語の類似性（a:RoG≧80%、b:RoG≧50%）。矢印は英語自身(100%)。アスタリスクはインド・ヨーロッパ語族以外の言語を示し、記号のない言語はインド・ヨーロッパ語族に該当する言語。




## 6 まとめ

本研究では、言語の歴史的側面を解明する目的で、文法データベース AUTOTYP を用いた言語の類似性解析法の構築を行った。任意の言語群に対して欠損率を制御した上で、言語の類似性を推定する方法を新たに構築し、北東アジアとヨーロッパの言語データで検証した。今後は、AUTOTYP 上でラベリングされている 24 の地域ごとに言語を分け分析を進めると同時に、言語学者と共に結果の検証を行う。これらの一連のパイプラインは、perl および R で公開予定であるが、より幅広い利用を想定し、web アプリケーションの構築を検討中である。

## 謝辞

本研究は JSPS 科研費 JP18H05080, JP20H05013, および JST 創発的研究支援事業 JPMJFR2060 の助成を受けたものです。

## 参考文献

1. Glottolog 4.5（オンライン）（引用日：2022 年 1 月 12 日.）https://doi.org/10.5281/zenodo.5772642
2. N. Creanza, O. Kolodny, M. W. Feldman, How culture evolves and why it matters. Proceedings of the National Academy of Sciences, 114 (30) 7782-7789 (2017).
3. J. Nichols, Linguistic Diversity in Space and Time (University of Chicago Press, 1999).
4. B. Bickel, J. Nichols, Oceania, the Pacific Rim, and the theory of linguistic areas. Annu.Meet. Berkeley Linguist. Soc. 32, 3–15 (2006).
5. B. Bickel, J. Nichols, T. Zakharko, A. Witzlack-Makarevich, K. Hildebrandt, M. Rießler, L. Bierkandt, F. Zúñiga, J. B. Lowe, The AUTOTYP typological databases. Version 0.1.0(2017); https://github.com/autotyp/autotyp-data/tree/0.1.0
6. H. Matsumae, P. Ranacher, P. E. Savage, D. E. Blasi, T. E. Currie, K. Koganebuchi, N. Nishida, T. Sato, H. Tanabe, A. Tajima, S. Brown, M. Stoneking, K. K. Shimizu, H. Oota, B. Bickel, Exploring correlations in genetic and cultural variation across language families in northeast Asia. Sci. Adv. 7, eabd9223 (2021).
7. R. Boyd, M. Bogerhoff-mulder, W. H. Durham, P. J. Richerson, Are cultural phylogenies possible? Hum. Nat. Biol. Soc. Sci. , 355–386 (1997).
8. A. Mesoudi, Cultural Evolution: How Darwinian Theory Can Explain Human Culture and Synthesize the Social Sciences (University of Chicago Press, 2011).