Vol.2019-CH-121 No.7
2019/8/1

言語間語彙比較に基づく野生動物の生息域推定の試み

大林武、山田和範、長野明子†
†東北大学情報科学研究科

概要：本研究では、野生動物の生息域と当該動物を表す語彙の関係を網羅的に調査することで、比較言語学的アプローチに基づく、野生生物の生息域推定を試みる。各種野生生物に対応するWikipediaの見出し語と言語間リンクに基づき、言語間の語彙の類似度を算出したところ、各種野生動物の生息域の広さ、その動物を示す語彙の多様性、ならびに単語長が相関していることを見い出した。

キーワード：野生動物、借用語、生息域

1. はじめに

出アフリカ以降、人類は様々な極限環境に適応しながら生息域を拡大してきた。寒冷地への進出を例にすれば、ネアンデルタール人との混血による変異獲得などのゲノムレベルの適応に加えて[1]、裁縫技術の進歩も重要だったとされる[2]。これをゲノム進化の視点から見ると、技術（あるいは文化）により自然選択圧が弱まり、それにより生息域の拡大に成功したと解釈できる。これとは逆に、文化が選択圧の一つとなる例も少なくない。潜水とPDE10A遺伝子変異[3]、牧畜とLCT遺伝子変異[4]、飲酒とALDH2変異[5]などである。エネルギー摂取は生物の基幹的活動であるため、数千年から数万年の時間スケールの中で、食文化とゲノム変異は互いに影響を与えつつ進化してきたと言える。しかしこのような文化とゲノム進化との関連は十分に解析されているとは言えない。これは、ゲノム配列が一次元のデジタル情報であり全人類規模で解析されつつある[6]のに対して、食文化を始めとする地域文化は、その多様性を画一的手法で記載することすら困難であることが一因であろう。

地域環境は極めて多くの要素から構成されるが、これを地学的、生物学的、文化的な側面に分類する（図1）。地学的要素は長期にわたり安定しており、網羅的な測定を行いやすい。この地学的環境を基盤として各地域特有の生物相があり、生物学的な環境を形成する。人類にとって食料を得ることは生存の根幹であるため、生活スタイルはこの生物学的環境に強く影響を受ける。文化的な要素は食文化をはじめとする各地域の文化的風習の全体であり、単純なデータ化は容易ではない。

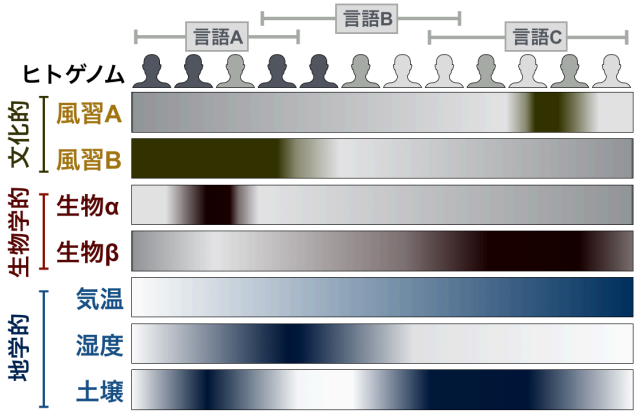

図 1　地域環境の構成要素

言語の語彙は、そのものが文化の一つであるだけでなく、その地域で暮らすために必要な単語（モノ・概念・行為の名前）を漏れなく含む基盤情報であり、そのため、語彙を網羅的に解析することにより、各地域の文化の相違を定量的に評価できる可能性がある。ここでいう「網羅的解析」とは、英語や日本語といった特定の個別言語を解析するのではなく、多数の個別言語を集め、通言語的 (cross-linguistically) に解析することを指す。書き言葉はデジタルデータであるため網羅的解析を行いやすく、死語となった言語に対してもある程度解析が可能である。

さらに、言語の語彙には、借用という言語接触の現象が広く観察されることから、地域間での人や動物、モノや概念の交流の足跡をたどる指標にもなる。語彙の借用とは、ある単語が、話し手の交易活動や各種メディアを通じて、源泉言語 (source language) から受容言語 (recipient language) へと移動することをいう[7]。例えば、「ウマ」の源泉は古代中国語であり、古代社会において日本にもたらされた単語である[8]。借用による単語の共有は、例えば英語とドイツ語がbook-Buchを共有しているというような、同じ語派 (language family)に属する言語間にみられる共有とは質的に異なる。後者の場合は、英語とドイツ語が同一の祖先的言語（祖語）に由来することから生じる単語共有である。

そこで本研究では地球規模での文化的環境のデータ化を念頭に、その当面の目標として、各種野生動物について、その動物を表す単語からその動物の生息域を予測する問題を考える。例として、各種言語におけるシャモアとコアラの名称（表1：Wikipediaの見出し語）と生息域（図2：IUCNの公開データ[9]）の関連を取り上げる。ヨーロッパ生息のシャモア (*Rupicapra rupicapra*) の名は、個々の言語、特にヨーロッパの言語において異なる形態をとるのに対し、オーストラリアに生息するコアラ (*Phascolarctos cinereus*) の名は、広範囲に同一形態をとるとわかる。この

例は、動物の原産地・生息域が、その動物の名前の形態的多様性と関係していることを示唆している。つまり、ヨーロッパ生まれの動物には、ヨーロッパの各種言語で固有の本来語 (native word) が用意されているため、形態的な差が生じる。表1でいえば、Chamois vs. Gems vs. Isard vs. Sarrio という対立である。他方、コアラはヨーロッパには生息していないため、その名前は借用によってもたらされる。その結果として、図1のように、語彙には形態的差異が生じない。要するに、名前の異なり度は、それが表す動物自体の歴史的足跡を反映するのである。

表 1　各種言語におけるシャモアとコアラの名称

言語	シャモア	コアラ
英語	Chamois	Koala
ドイツ語	Gämse	Koala
オランダ語	Gems	Koala
フランス語	Chamois	Koala
スペイン語	Isard	Coala
バスク語	Sarrio	Koala
ポルトガル語	Camurça	Coala
ロシア語	Серна	Коала
アラビア語	شامواه	كوالا
ベトナム語	Sơn dương Chamois	Koala
韓国語	알프스산양	코알라
インドネシア語	Chamois	Koala

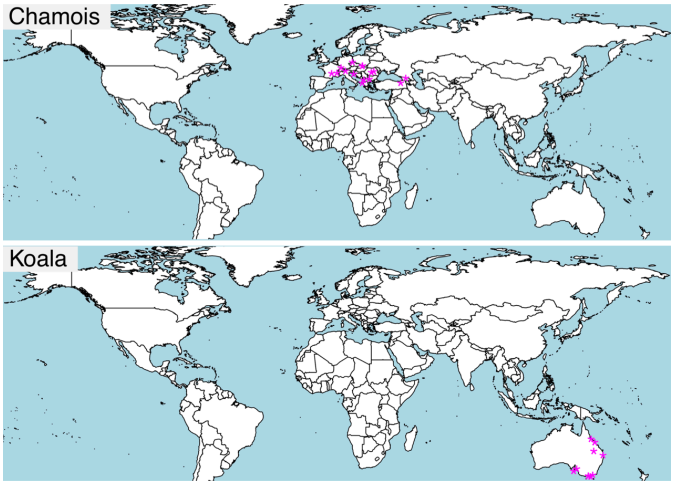

図 2　シャモアとコアラの生息域

網羅的な語彙の比較から、その語彙の地理的起源を推定するという問題を取り組むにあたり、生物の生息域は測定可能な良い教師データとなりえる。一方で、生物の生息域の推定そのものも生態学的に重要なトピックであり、直接観測の困難な過去の生息域の推定などの応用が考えられる。本稿では、表2にあげる20言語における語彙を用いて、表3の83の陸上生息野生動物の生息域との関係を調べ、語彙の多様性が原産地の情報を持つことを示す。

表 2　解析対象言語

ロマンス語派		ゲルマン語派		スラブ語派	
ca	Catalan	de	German	bg	Bulgarian
eu	Basque	en	English	ru	Russian
fr	French	nl	Dutch	pl	Polish
pt	Portuguese	sv	Swedish		
ウラル語派		**西アジアの言語**		**東アジアの言語**	
fi	Finnish	ar	Arabic	ko	Korean
		fa	Persian	id	Indonesian
		he	Hebrew	ms	Malay
		tr	Turkish	vi	Vietnamese

表 3　解析対象生物種の分類

目	対象科数	対象生物種数
Carnivora (食肉目)	7	31
Cetartiodactyla (鯨偶蹄目)	4	24
Dasyuromorphia (フクロネコ目)	2	2
Diprotodontia (双前歯目)	1	1
Lagomorpha (ウサギ目)	1	3
Monotremata (カモノハシ目)	1	1
Perissodactyla (奇蹄目)	3	7
Pilosa (有毛目)	1	1
Primate (霊長目)	6	7
Rodentia (齧歯目)	4	6
合計	**30**	**83**

2. 手法

2.1. 任意の言語ペアの文字の対応

借用関係を抽出するためには、比較する言語間の表記規則を対応づける必要がある。同一の対象物に対する多言語横断コーパスとして、Wikipediaの言語間リンクを利用した。百科事典であるWikipediaには大量の固有名詞が収録されており、言語間の文字の対応を定量化するのに適している。またWikipediaは集合知データベースであるため、将来的なデータ規模の拡大により解析精度の向上が期待できる。

一般に言語が異なると単語長が異なるため、単語を比較するには文字列の整列が必要となるが、ここでは先頭文字の比較を行うことで、文字列の整列の問題を回避した。言語 A と言語 B を比較する場合、両言語共通に存在するWikipedia見出し語ペア集合$\{(W_{A1}, W_{B1}), \cdots, (W_{AN}, W_{BN})\}$を対象に、言語Aの先頭文字$i$と言語Bの先頭文字$j$の対応強度$C_{ij}$を、同時観察確率$N_{ij}/N$ $(N = \sum_{i,j} N_{ij})$とそのランダム確率の対数尤度として定義する。低頻度値の安定化のため、擬似カウント1を全ての文字ペアの頻度に加えた。

$$C_{ij} = \log_2 \frac{N_{ij} \cdot N}{\sum_i N_{ij} \cdot \sum_j N_{ij}}$$

表2にあげる言語の全てのペアについて、文字の対応づけを行なった。C_{ij}の算出に用いた単語ペアは、最も多い英語-フランス語間で1048860単語ペア。最も少ないヘブライ語-マレー語間で24718単語ペアであった。

2.2. 単語の類似度

次に単語の類似度を算出する。言語Aと言語Bの任意の見出し語ペア(W_{Ak}, W_{Bk})に対して、W_{Ak} $[a_1, \ldots, a_n]$（単語長n）と$W_{Bk}[b_1, \ldots, b_m]$（単語長m）の類似度S_{ABk}は、2.1で求めたC_{ij}をもとに、動的計画法を用いた局所的アラインメントにより算出した。ここで、先に算出した先頭文字の一致傾向C_{ij}は、先頭以外の文字の一致傾向とは異なる可能性があるため、先頭文字同士の一致スコアは5倍の重みで評価し、スコアC_{ij}を上限3、下限0の打ち切りスコアとした。ギャップスコアは-1とした。また、文法の異なる言語の比較において、複合語を構成する要素の順序が入れ替わることがあるため、言語Bの単語を周期的に並べることで語順入替に対応した。

このように算出した単語の一致スコアは、単語長と文字組成の影響を受ける。それを補正するため、当該単語の文字をランダムに並び替えたランダムスコアR_{ABk}（5回施行の平均値）と、当該単語が取りうる最大スコアM_{ABk}を用いて、スコアが0から1の値になるように規格化した。規格化した単語の一致率の例を表4に示す。文字の種類が異なっていても妥当な類似度が産出されていることがわかる。

表 4　単語一致率の例

言語1	言語2	類似度
(英) Lion	(露) Лъв	0.69
(英) Lion	(独) Löwe	0.48
(英) Lion	(韓) 사자	0.00
(尼) Singa	(韓) 사자	0.53

2.3. 生息域データ

生息域データは国際自然保護連合にて公開されている陸上哺乳類の生息域データを用いた[IUCN19]。同データに2回以上の報告があり、かつ表2の20言語中で18言語以上にWikipediaの項目がある83生物種を解析対象とした（表3）。生息地のデータは次の8地域[10]に粗視化し、生息地域数に変換した。

- R1: Sub-Saharan Africa
- R2: Middle East and North Africa
- R3: Europe
- R4: Central and Eastern Europe
- R5: East Asia and the Pacific
- R6: South Asia
- R7: North America
- R8: Latin America and Caribe

図1の例では、シャモアの生息地域数は3（R2、R3、R4）、コアラは1（R5）となる。

3. 結果

3.1. 平均単語類似度による各種言語の相対関係

今回得られた単語類似度の傾向を得るため、83生物種の各々の単語類似度の平均をとり、言語間距離D_{AB}を求めた。

$$D_{AB} = 1 - S_{AB} = 1 - \sum_k S_{ABk} / \sum_k$$

図3はD_{AB}のヒートマップ（濃いほど距離が近い）を、群平均法で階層的クラスタリングを行った結果をともに示したものである。ゲルマン語派 (en, sv, de, nl)、ロマンス語派 (eu, ca, fr, pt)、スラブ語派 (pl, bg, ru) が各々クラスターを形成しており、同一祖語を共有する語派の内部で単語が共有される傾向が確認できる。一方で、インド・イラン語派のペルシャ語は、語族の異なるアラビア語やヘブライ語と高い単語類似度を示すことから、話者集団の地理的な近さから生じる言語の接触によって、単語の借用が起こっていることがわかる。また、マレー語 (ms) とインドネシア語 (id) が英語 (en) と近いことは、大英帝国の植民地主義の歴史、具体的にはマレーシア地域が英植民地であったことの結果として、英語を源泉とする借用が多いためと考えられる。英語 (en) は様々な言語と類似性を示しており、現代におけ

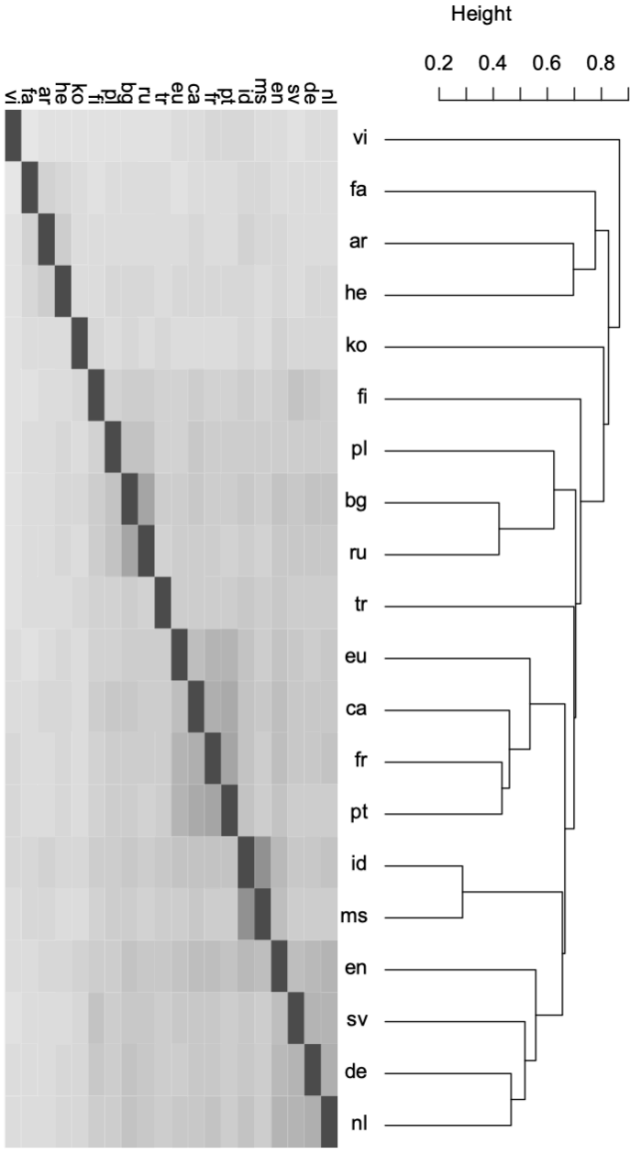

図 3　語彙の類似性

るグローバル言語として、代表的な借用の源泉言語であることを反映している [11]。

3.2. 単語と生息域の関係

次に各生物種の生息域の広さと語彙の特徴について調べる。言語Aと言語Bにおける単語kの類似度S_{ABk}から、各生物種の平均単語類似度S_kを求めた。

$$S_k = \sum_{AB} S_{ABk} / \sum_{AB}$$

合わせて関連する別指標も導入した。例えば、good spell (良い知らせ）がgospelに短化し、be going to がgonnaに短化したことからわかるように、単語の使用頻度とその長さは負の相関をもつ [12]。とすると、単語長が短い生物種名はその言語にとって重要であることを示唆するといえる。そこから、様々な地域の言語で短い生物種名が使用されていれば、生育域が広いと推定できることになる。平均単語長L_kの算出にあたっては、言語ごとに単語長の傾向が異なるため、言語ごとに平均と分散で正規化した単語長L_{Ak}を算出し、その単語ごとの平均値として算出した。

$$L_k = \sum_{A} L_{Ak} / \sum_{A}$$

83生物種の各々について、平均単語長（x軸）と平均類似度（y軸）をプロットしたのが図4である。濃い色が示す生息域の広い生物種では、平均単語長が短く、平均類似度も低い傾向が観察される。地元に昔から存在する生物種はその地域に重要であり、より短い単語になりやすい。また、昔から各地域に存在する生物種は各地域で独自の呼び方があり、単語のバリエーションが高い（すなわち単語の類似性の低い）と考えらえる。

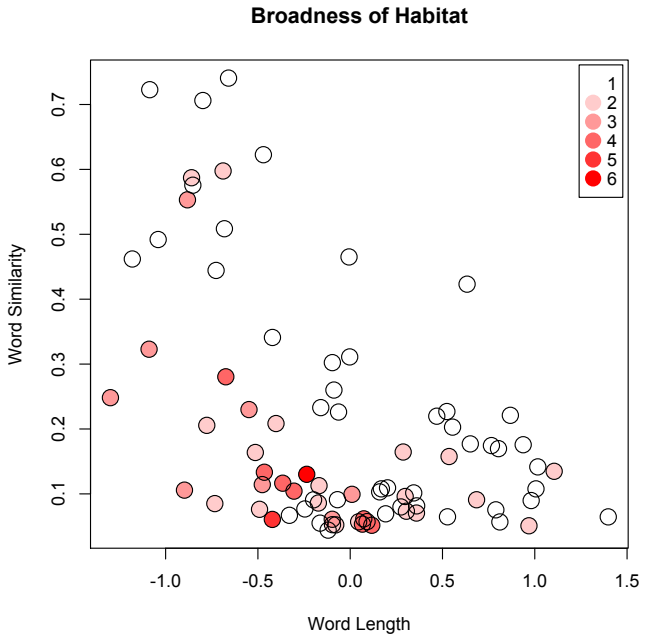

図 4　単語の特徴と生息域の広さ

なお、平均単語長と平均類似度は緩く逆相関している。これは今回の単語ペアの類似度の計算において、最大可能スコアで規格化しているため、単語帳の長い単語では、類似度が低く算出されていると考えられる。

本報告における文字の対応強度の算出においては、語頭、語中、語尾の区別が不完全である、本来語と翻訳借用の区別がないなどの技術的課題があるものの、大枠としては、各言語の語彙は生息域の広さと関連しているいえる。

4. まとめ

本研究では、各種言語に含まれる語彙の地理的起源を推定するため、野生動物を対象に、語彙と生息域に関連があることを示した。生息域の広さではなく生息域そのものを予測することが次のステップとなる。

生物種名の伝播に関しては、印象の強さ（獰猛な動物、巨大な動物）や有用性（毛皮や牙の供給）などの人間社会との関係性も重要な要素となる。このような観点を推定モデルに取り込んでいくことは、地域環境を形成する生物相が、人間社会にどのように受容され、伝播したかを解析するのに有効だと思われる。

語彙から地理的起源を推定する本アプローチは、動植物全般に留まらず、生物種以外の語彙に対しても適用可能であり、意味ドメインと伝播の関係も興味深い。また、生息域を予測するモデルの言語学的な特徴を解析することで言語接触についても俯瞰的な理解を得ることを目的とする。

近年の集団遺伝学的解析は人類集団の移動、融合、消滅の歴史[13]や、野生動物の家畜化や野生植物の栽培化の歴史を明らかにしつつある。本研究がゲノムで捉える人類史と、言語で捉える文化の動きをつなげる視点となることを期待する。

参考文献

1. Sankararaman S. et al. The genomic landscape of Neanderthal ancestry in present-day humans. Nature 2014 vol. 507, p. 354-357.
2. 関野吉晴. 2006.『グレートジャーニー全記録 1 1993-2002 移動編』毎日新聞社.
3. Ilardo M.A. et al. Physiological and Genetic Adaptations to Diving in Sea Nomads. Cell 2018 vol 173, p. 569-580.
4. Tishkoff S.A. et al. Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet 2007 vol 39, p. 31-40.
5. Okada Y. et al. Deep whole-genome sequencing reveals recent selection signatures linked to evolution and disease risk of Japanese. Nat Commun 2018 vol 9, 1631.
6. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature. 2010, vol. 467, p1061-73.
7. Haspelmath P., et al. (eds.) 2009. World Loanword Database. Leipzig: Max Planck Institute for Evolutionary Anthropology.
8. 小松英雄. 2001.『日本語の歴史：青信号はなぜアオなのか』. 東京：笠間書院.
9. IUCN 2019. The IUCN Red List of Threatened Species. Version 2019-1. http://www.iucnredlist.org. Downloaded on 3 June 2019.
10. Esty D.C. et al. 2008 Environmental Performance Index. New Haven: Yale Center for Environmental Law and Policy.
11. Crystal, David. 2003. English as a Global Language, 2nd edition. Cambridge: Cambridge University Press.

12. Mahowald K. et al. Word Forms Are Structured for Efficient Use. Cogn Sci. 2018, vol. 42, p. 3116-3134.
13. Pagani L. et al. Genomic analyses inform on migration events during the peopling of Eurasia. Nature 2016 vol. 538, p. 238-242.