E-004

n グラムを利用した琉球諸語の定量化

Quantification of Ryukyu languages using n-gram

松本 和樹 † 岡崎 威生 ‡
Kazuki Matsumoto Takeo Okazaki

1. はじめに

琉球諸語では Karimata [1] によって、方言間の系統関係と言語的多様性が生じた要因と過程を明らかにするために言語系統樹が作成されている。しかし、琉球諸語において系統樹作成にあたり観測地点の特徴を示す定量化の方法が定まっていない。

インド・ヨーロッパ語の系統樹作成は Schleicher [2] から始まり、言語系統樹作成において統計的手法が使われだしたのは 1990 年代からである。Gray and Atkinson [3] は系統樹解析にベイズ統計の手法を利用している。そして系統樹作成において統計的手法のために使われるものの多くが語彙を利用して定量化された言語データである。これらは言語ごとの語彙の構成を定量化したものであり、定量化された各言語の語彙構成を比較し、系統樹が作成されている。Pellard [4] は語彙ではなく個々の音変化そのものを特徴とみなし，言語間での特徴の共有を調べている。

定量化にあたり、本研究では地点間の発音変化の関係性に対して注目した。発音変化自体を観測地点の特徴として定量化し、地点間の比較を行う。本研究で行う定量化である発音変化自体を観測地点の特徴方法は言語ごとの構成を利用した定量化と比べて少ない。定量化のため、発音記号列を使用し、n-gram によって分解された発音記号列ごとの変化を要素として定量化を行う。また、区分された言語分類と本研究によって定量化された観測地点の分類を照らし合わせて定量化の手法の評価を行う。

2. 発音変化

一般的に発音は大きく分けて子音 (consonant) と母音 (vowel) の 2 種類に分類することができ、発音はこの子音と母音レベルの発音記号列によって表現されている。特に琉球諸語の発音において子音は 81 種類、母音は 35 種類が存在する。この表現を表 1 に示す。表 1 は”今”に対する地点ごとの発音を子音、母音レベルの発音記号列で表記している。

Karimata により琉球諸語では発音変化が起こるさいに発音記号列においては、単一の発音記号の変化、前または後ろの発音記号が影響した変化、前後の発音記号が影響した変化が確認されている。これは音の変化が単一の発音記号だけでなく、前後の発音記号にも依存していることを意味している。このことから、本研究では n 並びの発音記号列の変化の出現頻度を変化後の観測地点の特徴とした定量化を提案する。

表 1: 発音記号列”今 “

伊仙町伊仙	nj	a:				
知名町知名	n	a	m	a		
宮古島市池間	n	a	m	a		
石垣市富野	m	e:	m	a		
竹富町小浜	m	i	n	a	m	a
竹富町古見	m	i	n	a		
与那国町祖納	n	a	i			

3. n-gram による発音変化の定量化

音の並びを考慮して出現頻度を得るため、n-gram を利用する。n-gram は Shannon[5] が 1948 年に提唱した、自然言語処理学において単語や文字列を比較するにあたって使用される手法である。n-gram は テキストや単語に対し、前から順に隣り合った n 個の 文字列または単語の組み合わせで分割を行う。n が 1 の場合を uni-gram、2 の場合を bi-gram、3 の場合を tri-gram とよぶ。tri-gram による n-gram の使用は次のようになる。次の言葉「nature」を文字ごとで分割する場合、「nat」「atu」「tur」「ure」と分解される。特に分解によって出現する n 並びの文字列の集合 n-grams の種類のことをタイプと呼び、タイプごとの出現頻度を要素とし、特徴ベクトルが作成される。その後、特徴ベクトル間を比較して類似度を求めることがある。また、語頭と語尾の順序の情報も得るために bi-gram 以上の場合は先頭と末尾にダミーシンボルを追加して使用する。

各地点の発音情報のみによる頻度を特徴とした定量化のための定式化を示す。入力される地点 l における単語 w の発音記号列 $\boldsymbol{x}_{l,w}$ は次式のように定義される。

$$\boldsymbol{x}_{l,w} = \{x_0, x_1, \cdots, x_i, \cdots, x_{N_{l,w}}, x_{N_{l,w}+1}\} \quad (1)$$

$$x_i = \begin{cases} <b> & i = 0 \ \ (n > 1) \\ \in \boldsymbol{C} \ or \ \boldsymbol{V} & i = 1, 2, \cdots, N_{l,w} \\ <e> & i = N_w + 1 \ \ (n > 1) \end{cases} \quad (2)$$

発音記号 x_i は子音集合 $\boldsymbol{C}$ か母音集合 $\boldsymbol{V}$ のいずれかから入力されている。また、発音記号列の長さは $N_{l,w}$ と定義でき、地点、単語で異なる。また、bi-gram 以上で発音記号列の分解を行う場合、語頭と語尾の情報を得るためにダミーシンボル < b > と < e > が追加される。この発音記号列 $\boldsymbol{x}_{l,w}$ が n-gram により分解された発音記号列 $\boldsymbol{y}_{l,w}$ は次式で定義される。

$$\boldsymbol{y}_{w,l} = \{y_1, y_2, \cdots, y_{N_{l,w}-(n-1)+2}\} \quad (3)$$

†琉球大学大学院理工学研究科情報工学専攻, Graduate School of Engineering and Science, University of the Ryukyus
‡琉球大学工学部工学科知能情報コース, Computer Science and Intelligent Systems, University of the Ryukyus

表 2: アライメントにより対応のある発音記号列” 今 “

地点名	語形	1 子音	1 母音	2 子音	2 母音	3 子音	3 母音
旧笠利町笠利	nama	-	-	n	a	m	a
旧名瀬市金久	nama	-	-	n	a	m	a
龍郷町中勝	nama	-	-	n	a	m	a
竹富町小浜	minama	m	i	n	a	m	a
与那国町比川	nai	-	-	n	a	-	i

使用する n-gram が $n=2$ 以上の場合は $N_{l,w}-(n-1)+2$ 個に分解されるが、$n=1$ の場合は語順が存在しないことからダミーシンボルを追加しないため、発音記号列 $\boldsymbol{x}_{l,w}$ は $N_{l,w}-(n-1)$ 個に分解される。全地点 L、全単語 W で n-gram を適用した結果、$\boldsymbol{y}_{l,w}$ によって出現する n 並びの発音記号列の集合 n-grams のタイプは G 個存在する。これにより特徴数を G とした地点 l における n-grams のタイプ g の出現頻度 $N_{l,g}$ をまとめた特徴ベクトル $\boldsymbol{NG}_l$ は次式で定義される。

$$\boldsymbol{NG}_l = \{N_{l,1}, N_{l,2}, \cdots, N_{l,G}\} \tag{4}$$

これにより、各地点の発音情報のみによる頻度を特徴として定量化を行うことができる。

ところで、琉球諸語では発音変化が起こるさい、発音記号列においては単一の発音記号の変化、前または後ろの発音記号が影響した変化、前後の発音記号が影響した変化が確認されている。これらの変化は n 並びの発音記号列の変化を見ることによって表すことができる。上述の従来の定量化では n 並びの発音記号列の変化を見ることはできなかった。発音変化を特徴とするためには同単語で対応のある n 並びの発音記号列の変化の種類を得なければならない。そこで、発音変化自体を特徴とした定量化の定式化を行う。発音変化を確認するためには単語 w の各地点の発音記号列に対応がある必要がある。そのため、本研究では同単語内でアライメントが行われた発音記号列を適用する。例としてアライメントが行われ、発音記号列に対応のある単語”今 “を表 2 に示す。アライメントの適用により、個々の発音記号において対応する発音の変化をみることができる。その結果、単語 w における地点 l のアライメントが行われた発音記号列 $\boldsymbol{xa}_{w,l}$ は次式で定義される。

$$\boldsymbol{xa}_{w,l} = \{xa_0, xa_1, xa_2, \cdots, xa_{N_w}, xa_{N_w+1}\} \tag{5}$$

$$xa_i = \begin{cases} <b> & i=0 \quad (n>1) \\ \in \boldsymbol{C} & i=1,3,\cdots,N_w-1 \\ \in \boldsymbol{V} & i=2,4,\cdots,N_w \\ <e> & i=N_w+1 \quad (n>1) \end{cases} \tag{6}$$

アライメントが行われた発音記号列 $\boldsymbol{xa}_{w,l}$ には従来の定量化の方法と同じく、$n>1$ の場合、語頭と語尾の情報を得るためにダミーシンボル $<$ b $>$ と $<$ e $>$ が追加される。奇数位置の場合は x_i には子音 $\boldsymbol{C}$ のいずれかが、偶数位置の場合には母音 $\boldsymbol{V}$ のいずれかが入力されている。そしてアライメントの適用によって発音記号列の長さは地点に関係なく、単語ごとに長さ N_w で固定されている。この発音記号列 $\boldsymbol{x}_{w,l}$ が n-gram により分解された発音記号列 $\boldsymbol{ya}_{w,l}$ が次式である。

$$\boldsymbol{ya}_{w,l} = \{y_1, y_2, \cdots, y_{N_w-(n-1)+2}\} \tag{7}$$

地点 l の単語 w の発音記号列は n-gram によって、$N_w-(n-1)+2$ 個に分解される。$n=1$ の場合は語順が存在しないことからダミーシンボルを追加しないため、発音記号列は $N_w-(n-1)$ 個に分解される。n-gram により分解された集合は同単語内で対応する発音記号列を変化前、変化後としてパターンを作成する。これにより n-grams のタイプは全単語、全地点中で G 種出現するため、発音の変化パターンは最大で G^2 種存在する。

発音変化の例として単語”今 “の発音変化を示す。n-gram として tri-gram を適用した際、表 3 のように発音記号列は 3 つ並びに分解される。これらの対応のある 3 つ並びの発音記号列のタイプの組み合わせが発音変化となる (図 1)。

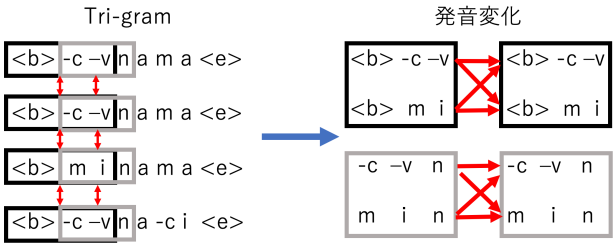

図 1: 発音記号列”今 “の発音変化

発音変化後の地点 l を基点とした変化前の n-gram のタイプ a に対して変化後にタイプ b となる発音変化の発生回数は $N_{l,a.b}$ である。特徴数を G^2 とし、発音変化の発生回数は $N_{l,a.b}$ をまとめた発音変化後の地点 l に対する発音変化の特徴量 NP_l は次式で定義される。

$$\boldsymbol{NP}_l = \{N_{l,1,1}, N_{l,2,1}, \cdots, N_{l,G,1}, \cdots, N_{l,G,G}\} \tag{8}$$

これにより、発音変化自体を用い、頻度を特徴量とした定量化が行われる。

以上より、従来行われている各地点のみの発音情報の頻度を特徴とする定量化と、本研究で提案する発音変化の頻度を特徴とした定量化を定式化した。

表 3: 発音記号列” 今 “に対する tri-gram による分解

旧笠利町笠利			旧名瀬市金久			龍郷町中勝			竹富町小浜			与那国町比川		
< b >	−c	−v	< b >	−c	−v	< b >	−c	−v	< b >	m	i	< b >	−c	−v
−c	−v	n	−c	−v	n	−c	−v	n	m	i	n	−c	−v	n
−v	n	a	−v	n	a	−v	n	a	i	n	a	−v	−n	a
n	a	m	n	a	m	n	a	m	n	a	m	n	a	−c
a	m	a	a	m	a	a	m	a	a	m	a	a	−c	i
m	a	< e >	m	a	< e >	m	a	< e >	m	a	< e >	−c	i	< e >

表 4: 定量化による分類と言語区分との一致率

定量化の方法	発音構成			発音変化		
n-gram	$n=1$	$n=2$	$n=3$	$n=1$	$n=2$	$n=3$
一致率	0.768	0.957	0.957	0.800	0.957	0.968

4. 定量化妥当性検証

提案した定量化の妥当性は直接評価することは困難である。そこで、定量化の結果を利用してクラスタリングを行い、既知の言語分類との整合性を検証することにより、間接的に評価する。

本研究で使用する琉球諸語の発音データは単語数が 130 語、地点数が 95 地点存在する。地点には言語学的に分類された区分によってラベルづけが可能であり、琉球諸語は琉球音声言語データベース [6] により、北から順に奄美徳之島諸方言、沖永良部与論沖縄北部諸方言、沖縄中南部諸方言、宮古諸方言、八重山諸方言、与那国方言の 6 つの区分に分けることができる (図 2)。この言語分類に対して、提案した定量化の結果を利用して各地点を分類し、その結果との一致率を利用して検証する。

定量化には大きく分け、従来法である各地点の発音情報のみを特徴とした定量化と提案した発音変化を特徴とした定量化の 2 種があり、さらに発音変化で単一の変化、前または後ろの発音の影響を受けた変化、前後の音の影響を受けた変化を特徴とするために $n = 1, 2, 3$ の uni-gram、bi-gram、tri-gram による 3 種類の発音記号列の分解を行う。このことから n を変えた各地点の発音情報のみを特徴とした 3 種類の定量化と、発音変化を特徴とする 3 種類の定量化の計 6 種類の定量化において、どの定量化が妥当かを検討する。

言語区分による分類と比較するため、定量化による各地点の分類を階層的クラスタリングによって行う。クラスタリングのため、定量化の結果として得られる各地点の特徴ベクトルを比較した距離行列は cos 類似度によって計算を行い、分類法には分類感度が良いとされる ward 法を使用した。実験結果を表 4、図 3〜8 に示す。

実験の結果、各地点の発音情報のみによる定量化と発音変化による定量化ともに uni-gram の一致率が他に bi-gram、tri-gram に比べて一致率が低く、この中でも各地点の発音情報のみによる定量化の一致率が 0.768 と最も低かった。uni-gram による発音情報のみの定量化の分類結果である図 8 からは、八重山諸方言に分類されるべき観測地点の多くが宮古諸方言に分類されてしまっている。このことから、uni-gram によって、単一の発音記号の頻度をみたさい、言語区分の間で出現頻度の差が少ないことを表し、bi-gam、tri-gram で音の並びに明確な差が出ることがわかる。また、最も一致率が高かったのは tri-gram による発音変化の定量化であり、一致率は 0.968 である。最も一致率が高い理由は他の方法では八重山諸方言と分類されてしまう与論島の観測地点が唯一正しく沖永良部与論沖縄北部諸方言に分類が行うことができているためである。このことから、発音変化を特徴とすることで、各地点の発音構成の特徴の比較では認識できない情報を比較することができている。

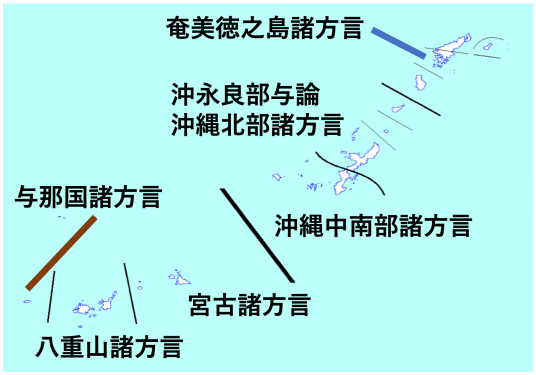

図 2: 地図上での言語区分

5. 終わりに

本研究では発音変化自体を要素とした n-gram による発音変化の定量化を行なった。検証実験の結果、従来の各地点のみの定量化に対して今回の提案方法が良いとする結果となった。特に前後の発音記号の影響を考慮した tri-gram による発音変化の定量化は最も言語区分との一致率が高かった。

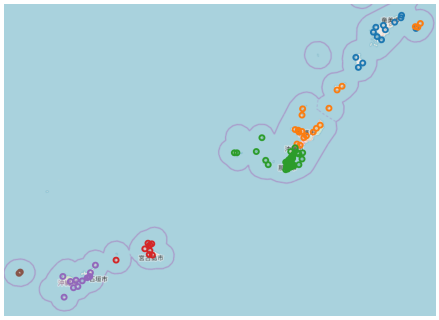

図 3: tri-gram による発音変化の分類結果

図 4: bi-gram による発音変化の分類結果

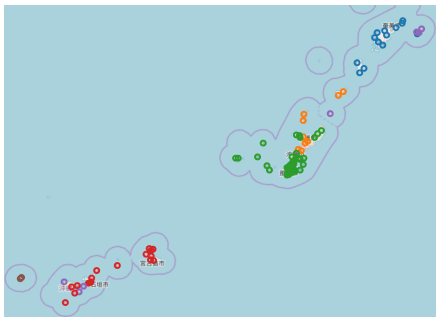

図 5: uni-gram による発音変化の分類結果

図 6: tri-gram による発音構成の分類結果

図 7: bi-gram による発音構成の分類結果

図 8: uni-gram による発音構成の分類結果

参考文献

[1] Shigehisa Karimata. Challenges and methodologies in ryukyuan language phylogenetic tree studies. *International Review of Ryukyuan and Okinawan Studies*, 48(7):1–14, 2018.

[2] August Schleicher. Die ersten spaltungen des indogermanischen urvolkes. *Allgemeine Monatsschrift für Wissenschaft und Literatur*, 3:786–787, 1853.

[3] Russell D Gray and Quentin D Atkinson. Language-tree divergence times support the anatolian theory of indo-european origin. *Nature*, 426(6965):435–439, 2003.

[4] Thomas Pellard. *Ōgami: Éléments de description d'un parler du Sud des Ryūkyū*. PhD thesis, Ecole des Hautes Etudes en Sciences Sociales (EHESS), 2009.

[5] Claude Elwood Shannon, Claude ElwoodShannon. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423, 1948.

[6] 琉球語音声データベース-琉球語概説. `http://ryukyu-lang.lib.u-ryukyu.ac.jp/intro/index.html`.