

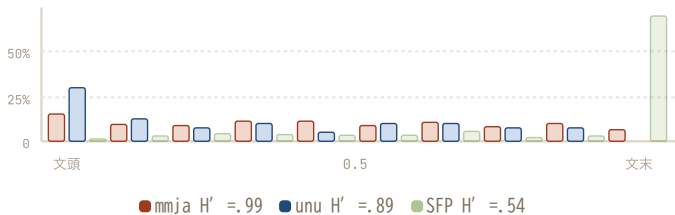
## §1 研究の問い

宮古語伊良部島方言の **mmja** / **unu** / **naugara** は同じ「フィラー」か？

(1) aidu **mmja** **unu** **naugara** kiban kiban nu pitu  
すると **DSC** **FIL** **FIL** 貧しい 貧しい GEN 人  
「すると、MMJA, UNU, NAUGARA, 貧しい人は（友達だったから）」

(2) par-i-ba **mmja** maju nu **unu** tabemunu u ba...  
行く-CSL **DSC** 猫 NOM **FIL** 食べ物 ACC TOP  
「行くからMMJA、猫が食べ物を（みんな食べてしまって）」

文内相対位置:



H' = 正規化エントロピー (1.0=均等)。mmja は文中に平坦分布。SFP は文末 70% に集中 (対比用)。

パツと見ると似た分布をしているにもかかわらず、**mmja** には **unu** や **naugara** にはない位置規則性や談話機能があるという「エセ内省」が、私には**ずーっと**ある (下地 2018)。しかし「ここで **mmja** は変ですか?」のような内省判断の引き出しは困難で検証手段がない。

私の抱えるこの問題は、より一般的には、フィールド言語学者が「フィラーっぽいやつ」をすべて **DSC** **FIL** と雑にグロスを振って終わりにしがち、という問題につながる。区別の妥当性を評価する方法論がない。

## 提案

圧縮法 (NLP) とサプライザル (情報理論) により、内省に頼らずコーパスから差異を検出する。

下地理則 (2018) 『南琉球宮古語伊良部島方言』くろしお出版

## §2 位置規則性

## 圧縮シャッフル検定

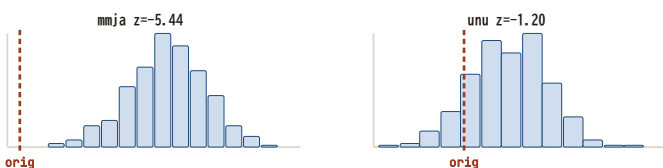
対象形式の文内位置を一様ランダム (uniform) に並べ替え500回→gzip圧縮サイズを比較。元データが有意に小さければ、位置にパターン (規則性) がある。Bonferroni 補正 (k=3) 適用。

$$z = (C_{\text{orig}} - \mu_{\text{shuf}}) / \sigma_{\text{shuf}}$$

	Z	\Delta/N	
<b>mmja</b>	<b>-5.44</b>	<b>0.40</b>	<b>***</b>
<b>unu</b>	-1.20	0.21	n.s.
<b>naugara</b>	-1.54	<b>0.39</b>	n.s.

|\Delta/n|=1トークンあたりの圧縮効果。naugara は mmja 並みだが n=23 で検出力不足。

シャッフル分布 (500回)



赤破線=元データ。mmja は分布外、unu は分布内。

例: 発話 V mmja N CP V

元データ	V_	<b>TGT_</b>	N_	CP_	V_
shuf 例1	V_	N_	<b>TGT_</b>	CP_	V_
shuf 例2	<b>TGT_</b>	N_	V_	CP_	V_

語を入れ替えるのではなく、**TGT\_** マーカーの位置をランダムに選び直す。非ターゲット語の POS 列は固定。gzip はバイト列の繰り返しパターンを検出するため、**TGT\_** が特定の POS 文脈に偏って出現するほど圧縮が効く。シャッフルでその噛み合いが壊れると圧縮サイズが増大する。

## mmja の POS 文脈

直前: V 26% · ISP 24% · CJP 8%

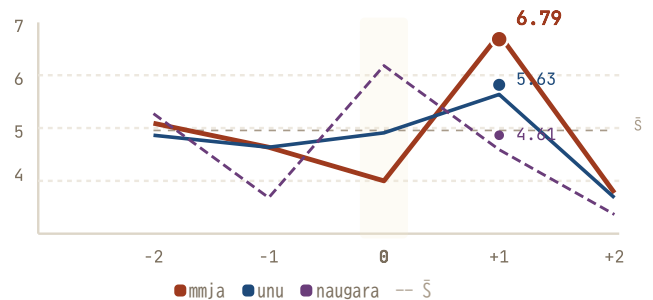
直後: N 30% · V 25% · PRN 13%

文中 89%。節境界 (V/ISP 後 → N/V 前) に挟まり後続内容語を予告 → 圧縮で検出。

## §5 情報密度 → 機能分離

## サプライザル

$$S(w_i) = -\log_2 P(w_i | w_{i-1}) \quad \bar{S}_{\text{corpus}} = 4.95 \text{ bits}$$



	POST $\bar{S}$	D	P
<b>mmja</b>	<b>6.79</b>	<b>0.80</b>	<.0001
<b>unu</b>	5.63	0.29	<.0001
<b>naugara</b>	4.61	-.15	<.0001

## mmja - 談話標識 (負荷アラート)

直後に予測困難な内容語。情報密度の遷移点を標示 (d=0.80)。

## unu - フィラー (語彙検索)

後続は名詞集中 (45%)。d=0.29 で有意だが mmja より弱い。

## naugara - フィラー (アラート解除?)

mmja と逆方向 (d=-.15) でアラート解除? n=22 で要追加データ。

「フィラーっぽいやつら」を雑に扱わず記述すると、なんかわかってきた気がする (!?)

## 近隣方言との比較 - 着目ポイントが大きく違う -

池間 mmya (川田他 2009)

← 左側に着目  
すでに言われた内容を背景化

本研究 mmja

→ 右側に着目 →  
これから言う内容の予測困難さを予告

+ 本研究は **unu** · **naugara** との体系的対立を新たに示した

川田他 (2009) 『琉球語宮古池間方言における mmya の談話機能』『言語文化のクロスロード』111-130