

超語彙パターンに基づく連辞型響鳴連鎖の自動発見手法の提案

吉川 正人^{†‡} 伊澤 宜仁[†][†] 慶應義塾大学大学院 [‡] 日本学術振興会

1 はじめに

日常会話を観察していると、(1)のような、類似形式の発話連鎖を目にすることが少なくない:

- (1) Joanne: It's kind of like you Ken.
Ken: That's not at all like me, Joanne.
(SBC015: 870.750–874.220)¹⁾

この連鎖は、[質問–応答]等、先行発話が後続発話のタイプを予期的に拘束する種の連鎖(「隣接ペア(adjacency pairs)」)とは言い難く、意図的・非意図的に関わらず、後続の話者が発話を形式的に先行発話と「類似させる」ことによって、両発話間にある種の関係性を形成する連鎖と言える。

このように、会話において、ある発話者 A の発話 u_i に対し、別の発話者 B が、音韻・形態・語彙・統語・意味・機能など、様々な側面で u_i と類似した発話 u_j を後続させ、「発話間の類似性を活性化させる(activation of affinities across utterances: Du Bois 2001)」現象は、「対話統語論(Dialogic Syntax: e.g., Du Bois 2001; 崎田 2010)」という枠組において「響鳴(Resonances)」と呼ばれ、議論されている。

響鳴は、1) 会話というマルチタスクにおける参加者の認知処理負担を軽減し、発話の産出/理解を易化すると考えられる、認知科学的に興味深い性質(e.g., 崎田 2010)と、2) 連鎖を繰り返すことで、対話を通して文法構造/知識の蓄積がなされると考えられる、会話と言語構造とを結びつける性質(e.g., Du Bois 2001; 崎田・岡本 2010)が見出され、ヒトの言語運用と認知処理のメカニズムの解明に関わる、重要な現象であると考えられる。

響鳴連鎖のタイプは多岐に渡るが、(1)のような、(語彙)要素の連続・不連続な配列(e.g., _'s _ like _)の共通性に基づく、「連辞型(syntagmatic)」の響鳴連鎖は上に挙げた2つの特徴双方に強く貢献すると考えられ、特に興味深い。

従って本研究では、連辞型の響鳴連鎖の分析が会話・文法研究、及び、言語使用を足掛かりにした認知の研究に有益であると考え、言語学的分析をサポートする

ツールとして、会話コーパスから連辞型響鳴連鎖(以下 SRSN)を自動抽出する手法を考案する。アルゴリズム・手法の詳細は3節で述べるが、簡単に言えば、1) 会話コーパスより話者交代(turn-taking)部分の発話連鎖 $\langle u_i, u_j \rangle$ を抽出し、2) 上に述べた要素配列、即ち、「超語彙パターン(super-lexical patterns)」²⁾を自動生成し、3) その共通度で u_i, u_j の類似度 $s(u_i, u_j)$ を評価し、4) $s(u_i, u_j)$ が一定以上の値の場合、発話連鎖 $\langle u_i, u_j \rangle$ を SRSN と定義し、5) その精度を手で当該コーパスに付した「響鳴タグ」との一致度に基づき評価する。

次節では、自動発見手法の詳細に入る前に、理論的背景となる対話統語論と響鳴連鎖について概説する。

2 響鳴と対話統語論

2.1 対話統語論の概要

日常会話には、類似要素を繰り返す「反復(repetition)」現象が頻繁に観察される。この一見冗長とも思える現象は複数の観点から議論がなされ、参加者の会話に対する「関与(involved)」を示す(e.g., Tannen 1989)、先行発話の修正や会話の促進などの対話調整機能を果たす(e.g., Beun 1985)、といった分析が行われている(小磯・伝 2011)。

対話統語論の研究対象は、対話において使用される様々な「記号の連携の構造(the structure of engagement between and through signs, Du Bois 2001: 1)と定義されている。対話統語論は、反復現象をこの「連携の構造」の顕著な表れと見なし、議論の中心に据え、発話の連鎖など、会話上に現れる単文を超えた単位に基づき文法を捉えようとする試みである(Du Bois 2001: 1)。中でも、対話統語論の支柱が、(1)に見たような「響鳴(Resonances)」現象である。

響鳴とは、「発話間の類似性を活性化させること」(Du Bois 2001: 1)と定義されており、その「類似性」の種類としては多様なものが想定されているようであるが、議論の中心は、上に挙げたような形式的な響鳴のうち、³⁾ 連辞型のもの(SRSN)である。このような類似性の活性化は、基本的には「隣接発話(juxtaposed utterances)」間で、かつ、先行発話と後続発話の発話者が異なる発話対に対して認定されている。⁴⁾

対話統語論が響鳴を分析の中心に据える理由は、こ

¹⁾ この対話は3.2節で紹介する *Santa Barbara Corpus of Spoken American English* (SBCSAE) (Du Bois, Chafe, Meyer, & Thompson 2000; Du Bois, Chafe, Meyer, Thompson, & Martey 2003; Du Bois & Englebreton 2004, 2005) から引用した。括弧内の表記は、本コーパスにおける、15 という id の付されたファイルのデータのうち、音声データにして 870.750 ミリ秒から 874.220 ミリ秒の間に行われた対話であることを示している。以下、同様の表記を用いる。

²⁾ この定義は3.1節で述べる。

³⁾ 先行研究では議論されていないが、他者による「言い換え(paraphrase)」などは、意味上の響鳴ということも可能と思われる。

⁴⁾ しかし、対話統語論の射程としては、非隣接の発話間の響鳴や、同一話者による響鳴(self-resonance)も同様の現象として扱われると考えられている(Du Bois 2001: 35–37)。

れが上に述べた反復の効用(会話への参与の表明、会話促進)を持つだけでなく、言語の使用及びコミュニケーションに、より根本的に重要な貢献をするものと考えられるためである。「会話」という営みは、状況の理解、先行発話の理解、発話のプランニング、対話相手との調整など、様々なタスクを並列的に進行させなくてはならない認知的に複雑・高負荷のプロセスであるが、これを我々はいとも容易く実行できてしまう(Garrod & Pickering 2004)。⁵⁾ 響鳴は、このような対話の高度な認知負荷を軽減している側面が認められる。当然ながら、自身の発話と先行発話の類似性の度合いが高くなるほど、発話/理解にかかる処理不可は軽減されよう。

また、認知的な側面だけでなく、文法構造の獲得や構築・創発という側面から見ても、響鳴という現象は興味深い。このような連鎖を通して、共通の、対話/談話機能を持った文法パターンのようなものが蓄積されていき、その蓄積こそが、文法知識を構成していくと考えられる側面も存在する(Du Bois 2001; 崎田・岡本 2010)。

2.2 対話統語論の方法

対話統語論は、従来の文法研究にありがちな、母語話者の直観に基づく(自)作例ベースの分析は重視しない。それは、自然会話に現れる対話的な現象に基づき文法研究を行うという方向性が故である。従って、分析の方法としては、対話/会話コーパスを利用し響鳴連鎖等を記述的に分析するというアプローチが用いられる。

響鳴の記述には、「ダイアグラフ(diagraph)」という表記法が用いられる。これは、発話間に見られる、要素の対応(mapping)関係を明示するための表記法で、例えば、(1)の連鎖は以下のように表示される。先頭行の数字は両発話間で対応する列の番号を示している。⁶⁾

- (2)
- | | | | | | | |
|----|------|----|------------|------|-----|---------|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| J: | It | 's | kind of | like | you | Ken. |
| K: | That | 's | not at all | like | me, | Joanne. |

このダイアグラフから、以下の対応関係が認められる: 1. $\langle It, That \rangle$; 2. $\langle 's, 's \rangle$; 4. $\langle like, like \rangle$; 5. $\langle you, me \rangle$; 6. $\langle Ken, Joanne \rangle$ 。1, 5 の対応は形式的には異なるが、直示的な指示表現であるため、同一の性質を持つと見做せる。また、6 に関しても、呼格の呼びかけ表現で、直示性を持つため、同様に考えて差支えないと思われる。

重要な点は、両発話の対応は、単純な語彙の共通度に基づくものではなく、その配列の情報(= タテ方向の対応)も含んでいるということである。従って、SRSN の自動発見には、1) 語彙の共通度と、2) その配列(e.g., 登場順序、連続/不連続性)(及び、可能であれば3)直示表現の対応)が適切に抽出できることが前提となる。

また、要素の対応関係は、単に形式的に一致していればよいわけではない。例えば、以下の連鎖の 1, 3 の対を考えてみる。⁷⁾

- (3)
- | | | |
|---|----------|-------------------------------------|
| 1 | ALICE: | [I think it's more] of a flu . |
| 2 | ANNETTE: | [they aren't so] |
| 3 | | That's what I think it was though . |
- (SBC043: 1445.498–1448.425)

これはダイアグラフで以下のように表記できる:

- (4)
- | | | | | | |
|-----|---|------|-------|------|----|
| | 1 | 2 | 3 | 4 | 5 |
| AL: | | I | think | it | 's |
| AN: | | That | 's | what | I |

2. $\langle I \text{ think}, I \text{ think} \rangle$; 3. $\langle it, it \rangle$; 4. $\langle 's, was \rangle$ の 3 対に関して対応関係を認めることができるが、対応箇所の発話全体の構造に対する機能は大きく異なっており、連辞的な構造の対応関係を認めるのは困難である。

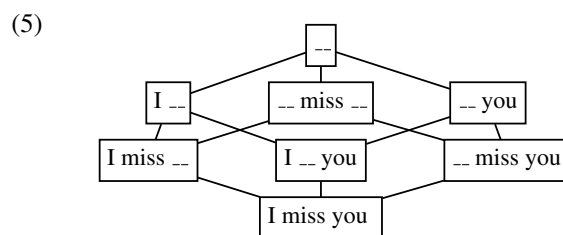
以上から、対話統語論の中心的分析対象である SRSN は、要素の配列の関係がダイアグラフ上に要素の(タテの)対応関係ととして明示できるような、隣接発話の連鎖を差すものと規定できよう。本研究が対象とするのは、このような連鎖である。

3 連辞型響鳴連鎖の自動発見

本節では、本稿で提案する連辞型響鳴連鎖の自動発見手法を詳述する。以下、1) アルゴリズムの詳細、2) 今回適用するデータの詳細、3) 処理手順、4) 処理結果・評価結果、を順に提示する。

3.1 アルゴリズム

SRSN の発見には、「パターン束モデル(Pattern Lattice Model, PLM: e.g., 黒田・長谷部 2009)」と呼ばれる計算モデルを応用した。PLM は、1) (単語等に)分節化された要素配列 $e = [w_1, w_2, \dots, w_n]$ を入力とし、2) e の分節 w_i を網羅的かつ再帰的に変項“ $_$ ”で置換し、3) 定項・変項の配列(=パターン p)の半順序集合、パターン束 $L(e) = \{p_1, p_2, \dots, p_m\}$ を出力するものである。例えば、 $e = [I, \text{miss}, \text{you}]$ として、 $L(e)$ は以下ようになる:



この“ $I _ \text{you}$ ”などの配列を「超語彙パターン(superlexical patterns)」と呼ぶ。尚、(5) はパターン束のハッセ図で、エッジによって示されている半順序関係は、超語彙パターンの「継承(is-a)」関係である。またこの図では、連続する変項を単一の変項に縮約している(e.g., “ $I _ _$ ” \rightarrow “ $I _$ ”).⁸⁾

このような超語彙パターンは、戦術の SRSN を特徴づける要素配列の対応関係の抽出に適している。本研究では、隣接する発話対 $\langle u_i, u_j \rangle$ からそれぞれパターン

⁵⁾ それなりの困難が伴う場合はあるが、それでも独話との比較で考えてみれば、会話の容易性は明らかである。認知的な観点からはより単純であるはずの独話は、明らかに、対話よりも困難である(崎田 2010: 619)。

⁶⁾ Du Bois (2001) などでもこの表記法は統一されておらず、様々な表記法が提案されている。本稿の表記は先行研究の表記に基づく独自の表記である。

⁷⁾ ブラケットは発話の重複を表す。

⁸⁾ 今回の目的とする SRSN の発見には、PLM の他にも、単語/文字 n -gram 集合や、単語/文字列集合の共通度を用いる方法などが考えられるが、いくつかの指標(単語集合共通度、文字 3-gram 集合共通度)を試験的に適用してみた結果、PLM が最も高い精度を示したため、採用した。

束 $L(u_i), L(u_j)$ (以降 L_i, L_j) を生成し、両者に共通のパターン、即ち、 $L_i \cap L_j = L_{ij}$ の情報を活用し、SRSN を探索した。

アルゴリズムの詳細は以下の通りである。先行発話を u_i 、後続発話を u_j 、先行・後続を問わない、発話一般を u で、発話の長さ = 語数を $l(u)$ で示す：

- (6) a. 発話対 $\langle u_i, u_j \rangle$ からパターン束 L_i, L_j を生成する；
 a'. この時、発話の長さ $l(u)$ がある閾値 θ を超える場合、 u を $n = \theta$ とする n グラム、 $G(u_i) = [g_1, g_2, \dots, g_m]$ に分解した上で、それぞれの n グラム g_k に対してパターン束 $L(g_k)$ を生成し、それらの和集合 $L(g_1) \cup L(g_2) \dots \cup L(g_m) = \bigcup_{k=1}^m L(g_k)$ を $L(u)$ と定義する；
 a''. また、発話 u もしくは n グラム g の前後に、ダミーの変項 “_” を追加する；⁹⁾
 b. $L_{ij} = L_i \cap L_j = \{p_1, p_2, \dots, p_l\}$ のうち、定項の数、Rank (R) が最大 ($rankmax$) のパターン群、 $M_{ij} = \{p \in L_{ij} \mid R(p) = rankmax\}$ を抽出する；
 b'. この時、 $\#M_{ij} \geq 2$ の場合、任意の s, t ($1 \leq s, t \leq \#M_{ij}; s \neq t$) について、対 $\langle p_s \in M_{ij}, p_t \in M_{ij} \rangle$ を取り出し、 p_u is-a p_s, p_u is-a p_t となるようなパターン p_u が L_j (=後続発話から得られたパターン束) 中に存在する場合、 $rankmax$ に 1 を加える ($rankmax = rankmax + 1$)；
 c. u_i と u_j の響鳴度 $s(u_i, u_j)$ (以降 s_{ij}) を、

$$s_{ij} = \frac{\log rankmax}{\log l(u_i)} \quad (1)$$

で与え、¹⁰⁾ $s_{ij} > 0.5$ の時、発話対 $\langle u_i, u_j \rangle$ を SRNS と認定する；

- c' この際、 $l(u_i) = 4, rankmax = 2$ で、かつ $l(u_j) < 4$ であれば、この場合も SRSN と認定する。

例えば、(7) の連鎖であれば、 $M_{ij} = \{“_ can _ do that _”\}$, $rankmax = 3, l(u_i) = 5$, 従って $s_{ij} = \log 3 / \log 5 \approx 0.68 \dots$ となり、SRSN と認められる。

- (7) 1 PHIL: and we can [go do that].
 2 BRAD: [Can I] do that
 (SBC010:326.310–327.310)

3.2 データ

分析には、*Santa Barbara Corpus of Spoken American English* (SBCSAE) (Du Bois et al. 2000, *et seq*) というアメリカ英語の会話コーパスを利用した。SBCSAE は、総語数約 25 万語で、Part 1 ~ 4 の 4 部からなり、各々が約 15 のファイル (=会話) を含み、合計で 60 のファイル (id 001~060) から構成される。ファイル毎に参与者の人数や性別、会話のタイプ (e.g., タスク指向, 講義, 日常会話) は異なるが、1 ファイル (id: 028) を除いて全て対面会話の書き起こしである。¹¹⁾

⁹⁾ この処理の理由は紙面の関係上詳しくは述べられないが、発話冒頭・発話末の要素を適切に処理するためのものである。

¹⁰⁾ この数式には計算的根拠があるわけではなく、経験則である。

¹¹⁾ 詳細は <http://www.linguistics.ucsb.edu/research/sbcorpus.html> も参照されたい。id:028 の会話は、電話による対話である。

SBCSAE のデータは、「発話 (Utterances)」(\approx 文) という文法上の単位と、下位構造の「韻律単位 (Intonation Units)」(以下 IU) という単位が存在し、二重の分節構造を持つ。以下に、発話を U , IU を u として、模式図を示す：

(8)

話者 A:	U_1	$u_{1,1}$
		$u_{1,2}$
話者 B:	U_2	$u_{2,1}$
	U_3	$u_{3,1}$
		...

3.3 手順

SRSN は、IU の対に対して認定することとした。また、対象とした会話は、多人数会話と、講義形式のものを除き、(基本的に) 2 人による対話形式のものに限定した。利用ファイルの id は、005, 006, 007, 010, 016, 017, 022, 024, 028, 029, 034, 041, 043, 044, 045, 046, 047, 058, 060 の 19 ファイルである。

また、連辞的な対応関係を考える際には、発話にある程度の長さが求められるため、認定の対象を、先行 IU が 3 語以上のものと限定した。

以下、 i 番目の発話単位を U_i , U_i の内 k 番目の IU を $u_{i,k}$ と表す。手順は以下の通りである：

- (9) 前処理
 a. 発話クリーニング: 1) 非言語情報を表す記号、2) 文末の句読点 (“,”, “?”), 3) 重複記号 (“[”, “]”), 4) 単語途中の長音記号 (“:”) の除去;
 b. 省略の復元: 助動詞 + “n’t” → “助動詞 not” の書き換え (e.g., “don’t” → “do not”, “can’t” → “can not”);
 c. アポストロフィーの剥離: アポストロフィー付き文字列の分離 (e.g., “I’m” → “I ’m”, “it’s” → “it ’s”);
 d. 直示表現置換: $\langle I, you \rangle$, $\langle it \text{ this, that} \rangle$ 等、話者交代で入れ替わる表現の共通記号への置き換え¹²⁾

(10) IU 対の取得

- a. 隣接した発話対 $\langle U_i, U_j \rangle$ の内、それぞれの話者を s_i, s_j として、 $s_i \neq s_j$ のとなっている対を「ターン交代対」 $T_{i,j} = \langle U_i, U_j \rangle$ として取得;
 b. $U_j = [u_{j,1}, u_{j,2}, \dots, u_{j,n}]$ の内、1 番目の IU $u_{j,1}$ と、 U_i の IU 全てについて、IU 対を作成 ($U_i \bullet u_{j,1} = \{\langle u_{i,1}, u_{j,1} \rangle, \langle u_{i,2}, u_{j,1} \rangle, \dots, \langle u_{i,m}, u_{j,1} \rangle\}$);
 c'. ただし $l(u_{j,1}) < 3$ の場合、 $u_{j,2}$ も IU 対作成の対象とする;
 c. 全ての IU 対に対して (6) を適用する (ただし $\theta = 7$)。

(11) 結果の評価

- a. (10) の結果を、第二著者によって人手で付けられた「響鳴タグ」との一致率によって評価する。

¹²⁾ 置き換えのパターンは以下の通り: $I, you, we, me, us \rightarrow \langle 1st2nd \rangle$; $my, your, our \rightarrow \langle 1st2ndPos \rangle$; $myself, yourself, ourselves, yourselves \rightarrow \langle 1st2ndRefl \rangle$; $it, this, that \rightarrow \langle DemSg \rangle$; $they, them, these, those, \rightarrow \langle DemPl \rangle$; $am, are \rightarrow \langle Am/Are \rangle$;

3.4 結果

自動発見及び評価の結果を Table. 3.4 に示す。数値は IU 対の数である。

Table. 1 自動発見結果と評価結果

atm_det	mnl_det	matched	precision	recall
195	168	127	0.651	0.756

(amt_det: 自動発見数; mnl_det: 手動発見数)

単純な評価は難しいが、再現率はある程度の水準に達しているものと思われるが、適合率はさほど高くない。次節で、失敗例を中心に要因の考察を行う。

4 考察

本研究で提案した手法は、NLP 的な応用を目指したものと言うよりは、言語研究のサポートツールとしての活用を念頭に置いたものである。従って、精度の評価には適合率よりも再現率を重視して考えるべきである。従って以下では、再現率を低下させている要因を中心に考察する。

4.1 検出漏れの検討

人手で発見した SRSN は 168 対で、自動発見できたのはそのうち 127 対であったため、計 41 対の SRSN が検出できなかったことになる。事例を見てみると、(12, 13, 14) のような「共通部分に対し発話が長すぎる」事例が多数 (27 対) であった (共通部分をボールドで示す):

- (12) 1 PAME: (...) I haven't read the book so **I don't know** .
2 DARR: Yeah but **I do know** .
(SBC005:230.790–232.290)
- (13) 1 KARE: I'm so **tired** .
2 SCOT: (...) **Ti:red** .
(SBC034:13.158–15.155)
- (14) 1 DARR: (...) **That** [does hurt] .
2 REED: [y- (.) it] **it hurts** .
(SBC046:538.604–540.530)

このような事例からは、(13) のように後続 IU が先行 IU の一部を反復している (「抽出」タイプ) が多数見られた。このタイプに関しては、先行研究 (e.g., 小磯・伝 2011) でも取り上げられており、比較的頻繁に観察される現象と言えるため、例えば u_j の一部が u_i に包含される場合には $\langle u_i, u_j \rangle$ を SRSN と認めるようにアルゴリズムを改定すれば、再現率の上昇が見込める (ただし適合率は低下する可能性がある)。

また (14) に見られる (does) *hurt* と *hurts* のように、文法上の屈折によってほぼ同型の表現が異なった表現として認識され、結果 SRSN と認定されないという事例も散見された。この種の事例は簡単な形態素解析を組み合わせることでもうまく拾うことができると考えられる。今後の課題とする。

4.2 過剰検出の検討

過剰検出は 68 例であった。この低減は第一目標ではないが、ある程度の傾向をとらえるため、簡単に事例を見る:

- (15) 1 JEFF: Oh honey
2 **I miss you** .
3 JILL: **I miss you** too .
(SBC028: 746.415–751.040)

この事例の 2, 3 が SRSN として抽出されたが、実際はこの反復は慣習化されたものであり、後続発話を「類似させる」側面があるとは考えづらい。このような事例の抽出を防ぐには、慣用的な反復をリスト化し、フィルタリングするなどの対処が考えられる。

5 総括

概して自動発見の精度は良好であり、検出漏れの傾向もはっきりしており、先述の対策を施せば響鳴研究の支援ツールとしては十分活用可能であると言える。NLP への応用は現時点ではやはり難しいが、有効な応用手段が見いだせれば、目的に合わせてチューニングしていくことである程度の応用は可能であると思われる。

今後は、TalkBank の他の会話コーパスに本手法を適用し、手法の精緻化を図ると共に、これを活用し響鳴の分析を進めていく。

参考文献

- Beun, R. J. 1985. The function of repetitions in information dialogues. *IPO annual progress report*(20), 91–98.
- 小磯花絵・伝康晴. 2011. 会話における (部分) 復唱発話の分析. 日本認知科学会第 28 回大会発表論文集 (pp. 250–255).
- Du Bois, J. W. 2001. *Towards a dialogic syntax*. Department of Linguistics, University of California, Santa Barbara. (Unpublished Manuscript)
- Du Bois, J. W., Chafe, W. L., Meyer, C., & Thompson, S. A. 2000. *Santa Barbara corpus of spoken American English, Part 1*. Philadelphia: Linguistic Data Consortium.
- Du Bois, J. W., Chafe, W. L., Meyer, C., Thompson, S. A., & Martey, N. 2003. *Santa Barbara corpus of spoken American English, Part 2*. Philadelphia: Linguistic Data Consortium.
- Du Bois, J. W., & Englebreton, R. 2004. *Santa Barbara corpus of spoken American English, Part 3*. Philadelphia: Linguistic Data Consortium.
- Du Bois, J. W., & Englebreton, R. 2005. *Santa Barbara corpus of spoken American English, Part 4*. Philadelphia: Linguistic Data Consortium.
- Garrod, S., & Pickering, M. J. 2004. Why is conversation so easy? *Trends in Cognitive Sciences*, 8(1), 8–11.
- 黒田航・長谷部陽一郎. 2009. Pattern Lattice を使った (ヒトの) 言語知識と処理のモデル化. 言語処理学会第 15 回大会発表論文集 (pp. 670–673).
- 崎田智子・岡本雅史. 2010. 言語運用のダイナミズム. 東京: 研究社.
- 崎田智子. 2010. 対話促進のメカニズム: 情報処理と談話語用論の視点から. 言語文化, 12(4), 619–663.
- Tannen, D. 1989. *Talking voices: Repetition, dialogue, and imagery in conversational discourse*. Cambridge: Cambridge University Press.