S-6

フィールドデータとコーパスに基づく日琉諸語の研究

中川奈津子

九州大学 nakagawanatuko@gmail.com

1 はじめに

本稿では、日琉諸語の情報構造と名詞標識の研究を例に、コーパスやフィールド調査によって得られるデータで研究することの利点と問題点を議論する。

2 背景

本発表ではコーパスとフィールドデータに基づく日琉諸語の研究が主眼であるが、筆者の研究対象が情報構造であるので、まず背景を共有するために筆者のこれまでの研究を概観する。

筆者は、情報構造を前提と断定という伝統的な概念と結びつけて議論してきた (Nakagawa, 2016, 2020)。前提現象は以下のように定義されている (Russell (1905); Strawson (1950); Stalnaker (1974, 1998) なども参照。); "the phenomenon whereby speakers mark linguistically information as being taken for granted, rather than being part of the main propositional content of a speech act." (Beaver et al., 2021) [1] 古典的な例を (1), (2) に示す。(1) には「自分の妻を殴っている」という前提があり、重要なことに、この質問に「はい」と答えても「いいえ」と答えても、前提を取り消すことはできない。すなわち、率直には訂正しがたい (unchallengeable である)。また、「自分の妻を殴るのをやめていない」のように否定しても、前提は否定することができない。

(1) 自分の妻を殴るのをやめたのか？

(2) には「フランス国王が存在する」という前提が存在する。元の例文は英語であるが、ここでは日本語の訳に「フランス国王は」とハが含まれていることにも注目したい。

(2) フランス国王はハゲだ。

日本語の主題助詞と言われるハもこの前提現象と密接に関わっている。例えば以下の (3-a) と (3-b) を比べてみると、アイスの存在を話し手も聞き手も共有していることまでは同じだが、聞き手が今アイスのことを考えているとは想定しづらい (3-a) の状況のほうが、聞き手が冷凍庫を覗いているのでアイスのことを考えていると想定しやすい (3-b) よりもハが使いにくい。

(3) S と H はルームメイトで、S は H が冷凍庫にアイスクリームを入れたのを見ていた。S はそのアイスクリームを H がいない間に食べた。

1) 本稿では断定は扱わないので定義は省略する。

a. [次の日に学校で H に会って S が] アイス{**??は/Ø**}食べちゃったよ。
b. [家で H が冷凍庫を覗いているのを見て S が] アイス{**は/Ø**}食べちゃったよ。

これはハが後続する要素が、話し手も聞き手も談話に導入済みであると了解している事項であるという話し手の前提があるからではないかと考えた (Nakagawa, 2020, Ch.3–4)。

実際、ハの後続する要素は、聞き手が「へー」など驚きを示す感嘆詞の後にニュースとして繰り返しづらく、また簡単には訂正しがたい (unchallengeable である)。例えば (4-A2) では、「へー、太郎は！」のように「太郎は」の部分をニュース（断定）として繰り返しづらい。一方、「へー、教授！」は問題なくニュースとして繰り返しうる。また、「太郎は」の部分は訂正しづらく、(4-C)「違うよ、（太郎じゃなくて）次郎だよ」のような訂正は奇妙である。これは (2) の「フランス国王は (*the king of France*)」に「フランス国王が存在する」という前提があるのと似ている。ただし、(3) の例も考え合わせると、英語の *the* とは異なり、「話し手も聞き手も導入済みの事項であると了解している」という前提があると思われる。

(4) A1: 太郎は何してる人？
B: 太郎は教授だよ
A2: へー、{教授！/太郎は！}
C: 違うよ、{助教だよ/次郎だよ/次郎は教授だよ}

ただしこれはあくまでも筆者の直感に基づくものであり、どのように実証できるのかを考えると、その方法論は自明ではない。

3 コーパスに基づく研究

Nakagawa (2016, 2020) では、日本語話し言葉コーパス (Maekawa et al., 2004) を用いて、実例に基づきこの仮説を実証しようと考えた。しかしながら、「話し手も聞き手も導入済みの事項であるという話し手の前提がある」ことそのものを実例の中から見つけて数え上げるのは困難であることに気づいた。これも結局は筆者の直感に基づくものであり、前節であげた問題点が解決されるわけではない。資金があって複数のアノテーターに依頼し、「話し手も聞き手も談話に導入済みの事項であるという話し手の前提がある」要素を見つけてくださいという課題を出せば客観的にはなるかもしれないが、多くのアノテーターが一致してそのような要素を見つけられるかというと疑問が残る。例えば図 1 は筆者が用いたデータの一部 (S00F0014) であるが、どの要素にそのような前提があり、どの要素にはないのか、はっきり判断できるだろうか？

そこで筆者は、談話に「導入済み」の要素か否かは客観的に判断できると考え、その情報をコーパスに付与していくことにした。コーパスにもともと付与されている文節の ID (図 1 C 列) を利用し、各名詞の ID とした。談話を話された順に追っていき、前に出てきた名詞と指示対象が同じである名詞が現れた場合は、その名詞に同じ ID を付与した。例えば図 1 では、文節 ID 5 の「旅」と ID 11 の「旅行」の指示対象を同じと考え、同じ ID を付与している (R 列)。

アノテーションの集計結果を図 2 に示している。“Anaphoric” が「導入済み」の名詞、“Non-anaphoric” がそうでない名詞である。図が示す通り、ハが後続する名詞の 6 割ほどが導入済みであり、4 割は仮説に反する結果となった。しかし、実例をよく観察してみると、ハが後続している導入済みでない名詞は、全く新規の名詞とも言えないことがわかった。例えば (5-c) の「試験」は導入済みでないのにハが後続している例だが、(5-a) のように直前に「入社」という語が出現している。入社に試験はつきものであるという一般的知識により、「試験」も導入済みであるという調整 (accommodation) が働いて「試験は」が可能になると考えられる。

図 1　アノテーションの例

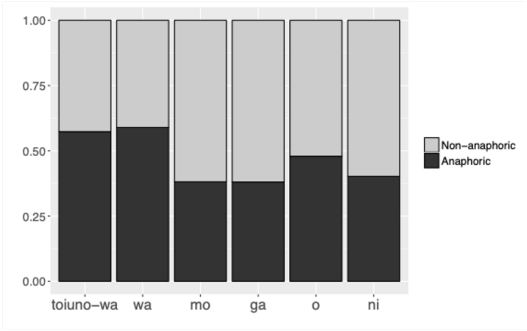

Figure 4.1: Particle vs. information status (ratio)

図 2　アノテーション結果

(Nakagawa, 2020, 95)

(5)　a.　えーとある旅行社にあの一応入社決まりました
　　b.　…
　　c.　非常に**試験は**難しかったと今も覚えております　(CSJ S01F0038: 231.34-241.96)

これは Prince (1981) のいう inferable information、あるいは Clark (1975) の bridging anaphor と類似している。例えば (6-a) では、先に導入された *a bus* には運転手がつきものであるため、*the driver* は初出だが確定記述になると説明される。同様に日本語でも、「バス」が導入済みなら「運転手さん」にハが後続しても良いと考えられる。

(6)　a.　I got on a bus yesterday and **the driver** was drunk.　(Prince, 1981, 233)
　　b.　昨日バスに乗ったら、**運転手さん{は/が}**酔っ払っていた。

ただし日本語の場合は、ハもガも後続できる。これは単に導入済みの背景情報として「運転手さん」に言及するか、驚きをもって断定の一部として言及するかの違いであると思われる。

　また、導入済みであってもハが使いづらい場合もある。(7-f) の「仕事の部分」ように、導入から時間が経った、あるいは「名声」という別の事項について語ったあとでは、「仕事の部分は」と言うのは困難で、「ですけ (れ) ど (も)」や「ですが」のような別の標識のほうが適切である。

(7)　a.　これからの あの 目標っていうのが ありまして
　　b.　ま それは 大きく 分けて 二つ あるんですけども
　　c.　ま 名声の部分と 仕事っていう部分が ありまして
　　d.　一番目の **名声の部分は** やはり あの 今まで 受けた コンクールの最高順位が あの ま 日中友好国際音音楽コンクールっていうのがあってそれが まー 一般のピアノ部門で 四位で 奨励賞だったんですね
　　e.　で それを 超えて あの 三位以内に 入賞することが まず 一つで
　　…
　　f.　で 後は **仕事の部分{なんですけれども/??は}** あの 一回の ギャランティーが えーと 勿論 アップするように　(CSJ S00F0209: 495.77-539.19)

このことは、ハは「話し手も聞き手も導入済みの事項であるという話し手の前提がある」だけでなく、「聞き手も今まさにそれが念頭にある」という前提もあることを示す。

直感だけではわからない傾向があることもわかった (中川, 2023)。同じ導入済みの名詞でも、主語 (A と S) のほうが目的語 (P) よりもハが後続しやすい (図 3)。導入済みは “Given” と表示されており、いわゆる主題標識 (ハ) の割合を “Top” で表している。一方、南琉球八重山語白保方言ではこの傾向は見られない (図 4)。名詞標識 *=ja* が主題標識に当たる。

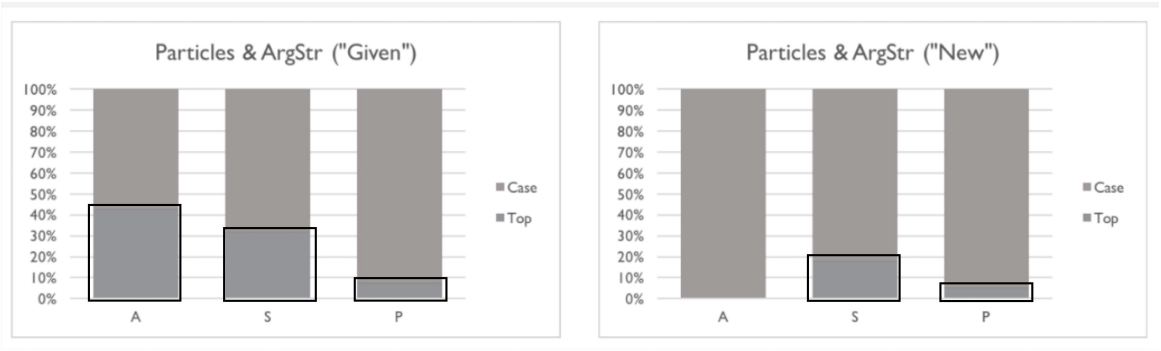

図 3　項構造と情報ステータスによる主題標識の分布の違い（東京方言）

(中川, 2023, 107)

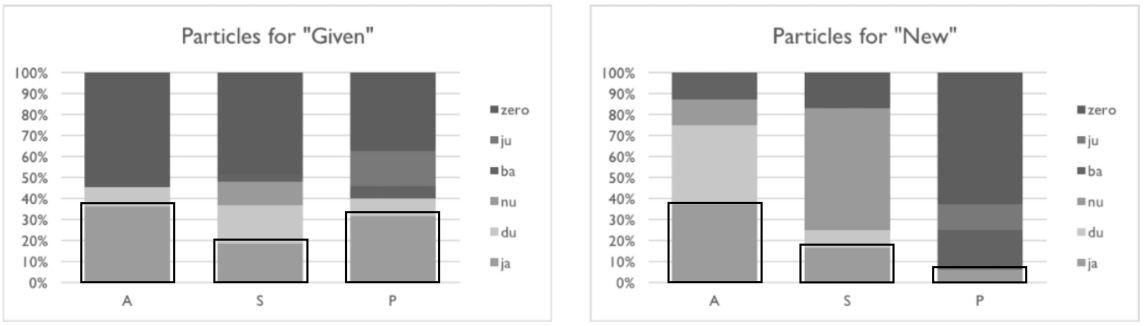

図 4　項構造と情報ステータスによる主題標識の分布の違い（白保方言）

(ibid.)

コーパスデータに基づく言語分析の際には、定量的な傾向を見ることも重要だが、傾向の例外となる事例に関しても詳細に分析し、別の要因が働いているのか、あるいは同じ原理で説明できるのかを議論する必要がある。このように議論を進めることで、量的な傾向、その背後に働く原理、提案の原理では説明できない例外を区別でき、残された課題を明確にすることができる。

4　フィールド調査に基づく研究

筆者は、コーパス調査だけでなく、南琉球八重山でフィールド調査も行っている。フィールド調査においても、主題標識とされる標識 (琉球の多くの言語では *=ja*) が日本語のハと同様の機能を持っているのではないかと考えているが、このことをそれぞれの地域の母語話者の方々に直接尋ねて回答を得るのはかなり難しい。

まず従来の方法通り、翻訳による聞き取り調査が考えられる。この方法論では、典型的には研究者が調査票を準備し、共通語（日本語）を自分の母語に翻訳してもらうという方法である。例えば下地理則氏のウェ

ブサイトで、日琉諸語を調査するための調査票が公開されている。[2] 「誰が泣いているの？」「弟が泣いているんだよ」など、質問と応答のペアで比較的自然な会話になっており、母語話者の方々も答えやすい調査票になっている。また、母語話者の中には言語の直感が鋭い方々がいるので、そのような人を見つけて調査すればかなりのことがこの調査票で明らかになるかもしれない。Max Planck Institute for Evolutionary Anthropology のウェブサイトにおいて、数多くの調査票が公開されている。[3] 情報構造に特化した調査票としては Skopeteas et al. (2006) が存在する。ただし、ある言語（共通語）から自分の母語に翻訳してもらうことによる共通語の影響は避けがたいと考えられる。例えば「雨に降られた」のような迷惑受け身は少なくとも八重山語では一般的ではないと思われるが、形態論的に受け身を作ることは可能なので、迷惑受け身を直訳できる話者もいるかも知れない。しかしこれがこの言語の使用法を反映しているかどうかは慎重に考える必要がある。

この問題を解決するために、翻訳を回避する手法もある。例えば、絵や絵本を母語話者に見せて、それを描写してもらい、どのような表現が出てくるかを調べる方法が存在する。情報構造の研究の場合は、継続して出てくる人物や物がどのような表現で言及されるか、あるいはゼロ代名詞になるか、初めて出てくる人物や物がどのような表現や構文で言及されるかが注目される。Berman and Slobin (1994) など、*Frog where are you?*という文字のない絵本を用いた研究が典型的である (frog-story project)。類似の手法で、動画を見せてそれを描写してもらう方法もある。少年が収穫中の梨を盗む Pear Story が代表的な例である (Chafe, 1980)。同じ登場人物が出てくる同じ展開の物語が数多く収集できるという点で、多言語比較に向いている。Pear Story は話の展開が文化依存にならないように、あえて筋のよくわからない動画になっているが、そもそも登場人物が（おそらく）西洋人であることや、梨、自転車など、文化依存性は排除できていない。筆者らが八重山で Pear Story を見せたとき、母語話者は少年が家族のパパイヤ収穫を手伝う話として語った。このようにして得られた談話は、前節のコーパス言語学的な手法で分析されることが多いだろう。したがって、前節にあげた問題をはらむことになる。

5 聴取・産出実験に基づく研究

心理言語学的な実験は、殆どの場合には実験室で行われるが、フィールドで行われる実験もある。地域の母語話者がコンピュータに馴染んでいない場合、コンピュータのキーを押してもらったり、画面を見せたり、場合によってはヘッドセットを着用してもらったりするのも困難な場合がある。しかし、それでもある種の実験は可能である。例えば Rohde and Kehler (2014) が採用しているような、1 文を見せてその続きを考えてもらうようなタスクはフィールド調査でもやりやすいだろう。話者がどのように代名詞を用いるか、というような特定の問いに対して、比較的統制された環境による文の産出が期待できる。また、中川・小川 (to appear) では、助詞の部分にノイズが入った音声を聞かせ、それを繰り返してもらうことで関西方言における助詞ガの出没の条件を明らかにしようとした。この種の実験手法も、聞こえにくかった発話を繰り返すという比較的自然な文脈の中で行われているため、フィールド調査でも適用しやすい。他にも空間指示枠 (Frames of Reference) の実験など、フィールドでこそ可能な実験手法もある (O'Meara and Pérez Báez, 2011)。

心理言語学的な手法では、結果を統計的に分析することがしばしば前提となっているため、多くの被験者が必要である。しかし話者が少数しかいない言語では、多くの参加者を集めることが困難である。また、1

2） https://sites.google.com/view/japoniclanguages-questionnaire/ (2024/5/13 最終アクセス)
個人のウェブサイトで公開するのはデータの持続可能性、著作権、引用可能性の観点から問題があるので、作者を明示したうえでしかるべき機関リポジトリで公開されることを強く要望する。

3） https://www.eva.mpg.de/lingua/tools-at-lingboard/questionnaires.php (2024/5/13 最終アクセス)

人 1 人の話者と関係を築いているため、1 つの実験にかなり時間がかかる。世間話なしに実験を開始し、終了したらすぐにその場を立ち去るのはその後の人間関係にも影響を及ぼしかねない。この問題を統計的に、あるいは手法として解決するにはどうすれば良いか、読者に問いかけたい。

6 (大規模) 言語モデルを用いた研究

翻訳による聞き取り調査などによって日本語共通語と目的の言語の対訳コーパスがあるなら、既存の大規模言語モデル（この場合は日本語）からの転移学習とファインチューニングによって、低資源言語においても ChatGPT などの生成 AI の精度が向上する可能性が示唆されている (坂井, 2024)。そして言語モデルを認知モデルと見なして研究対象にする試みもすでにある (日本語における試みは例えば Kuribayashi et al. (2022) など)。言語モデルとヒトの認知モデルが類似の振る舞いをする条件を見つけ、日本の諸方言・言語に応用するならさらにそれが転移学習可能であることも示す必要がある。課題は数多く存在するが、フィールド調査によって得られたデータに基づく大規模言語モデルを用いて研究する日もそう遠くないうちに訪れるかもしれない。

7 おわりに

本稿では、情報構造と名詞標識の関係という、言語の表層から抽出しづらい研究課題を例に、コーパスやフィールド調査で得られたデータを用いてどのような方法で研究ができるか、またその方法にはどのような問題があるのかを議論した。

参考文献

Beaver, David, Bart Geurts, and Kristie Denlinger (2021) “Presupposition,” *The Stanford Encyclopedia of Philosophy (Spring 2021 Edition)*.

Berman, Ruth A. and Dan I. Slobin (1994) *Relating Events in Narrative: A Crosslinguistic Developmental Study*: Routledge.

Chafe, Wallace ed. (1980) *Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production* in , Advances in Discourse Processes, No. 3, New Jersey: Ablex.

Clark, Herbert H. (1975) “Bridging,” in Schank, R. C. and B. L. Nash-Webber eds. *Theoretical Issues in Natural Language Processing*, pp. 169–174, New York: Association for Computing Company.

Kuribayashi, Tatsuki, Yohei Oseki, Ana Brassard, and Kentaro Inui (2022) “Context Limitations Make Neural Language Models More Human-Like,” in Goldberg, Yoav, Zornitsa Kozareva, and Yue Zhang eds. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 10421–10436, Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, December, DOI: 10.18653/v1/2022.emnlp-main.712.

Maekawa, Kikuo, Hideaki Kikuchi, and Wataru Tsukahara (2004) “Corpus of Spontaneous Japanese: Design, Annotation and XML Representation,” in *Proceedings of the International Symposium on Large-Scale Knowledge Resources (LKR2004)*, pp. 19–24, Tokyo: Tokyo Institute of Technology.

Nakagawa, Natsuko (2016) “Information Structure in Spoken Japanese: Particles, Word Order, and Intonation,” Ph.D. dissertation, Kyoto University, Kyoto.

——— (2020) *Information Structure in Spoken Japanese: Particles, Word Order, and Intonation*, Berlin: Language Science Press.

中川奈津子 (2023) 「日琉諸語における主題標識の類型と格標識」,『日本語文法学会 第 24 会大会予稿集』, 105–112 頁.

中川奈津子・小川雅貴 (to appear) 「産出実験による関西方言の格配列と情報構造の調査」.

O'Meara, Carolyn and Gabriela Pérez Báez (2011) "Frames of Reference in Mesoamerican Languages (Special Issue)," *Language Sciences*, Vol. 33, No. 6.

Prince, Ellen (1981) "Toward a Taxonomy of Given-New Information," in Cole, Peter ed. *Radical Pragmatics*, pp. 223–256, New York: Academic Press.

Rohde, H. and A. Kehler (2014) "Grammatical and Information-Structural Influences on Pronoun Production," *Language, Cognition and Neuroscience*, Vol. 29, No. 8, pp. 912–927, September, DOI: 10.1080/01690965.2013.854918.

Russell, Bertrand (1905) "On Denoting," *Mind*, Vol. 14, pp. 479–493.

坂井美日 (2024) 「生成 AI を用いた鹿児島方言生成: 日琉諸語の低資源言語・方言の生成に向けた試み」,『言語処理学会 第 30 回年次大会 発表論文集』, 234–238 頁.

Skopeteas, Stavros, Ines Fiedler, Anne Schwarz, Ruben Stoel, Gisbert Fanselow, Caroline Féry, and Manfred Krifka (2006) *Questionnaire on Information Structure: Reference Manual*, Interdisciplinary Studies on Information Structure, Potsdam: Universitätsverlag Potsdam.

Stalnaker, Robert C. (1974) *Pragmatic Presuppositions*, pp. 197–214: New York University Press.

——— (1998) "On the Representation of Context," *Journal of Logic, Language and Information*, Vol. 7, pp. 3–19.

Strawson, Peter F. (1950) "On Referring," *Mind; a quarterly review of psychology and philosophy*, Vol. 59, pp. 320–344.