情報構造の類型論に向けた談話データのコーディング

児倉徳和 [1]　中川奈津子 [2]　佐藤久美子 [3]　吉村大樹 [1]
[14] 東京外国語大学アジア・アフリカ言語文化研究所　[2] 九州大学　[3] 国立国語研究所
koguran@aa.tufs.ac.jp　nakagawanatuko@gmail.com
satok@ninjal.ac.jp　taiki_y@aa.tufs.ac.jp

概要

本発表では、情報構造に関する類型論的研究に向けた談話データのコーディング作業の概要と、その作業の課程で認識された理論的な問題点を報告することを通し、情報構造の研究のための談話資料のあるべきコーディングについての議論を行う。

1　研究目的

本研究は、ユーラシア大陸の広範な地域で話されるアルタイ諸言語と日本語諸方言を対象に、主語卓立 (subject prominent) ／主題卓立 (topic prominent) という類型論的枠組みからの比較対照を行い、対象言語の言語的多様性を明らかにするとともに類型論的枠組みの精緻化を目指すものである。

アルタイ諸言語とはチュルク諸語 (トルコ語など)、モンゴル諸語 (モンゴル語など) とツングース諸語 (エウェン語、満洲語、シベ語) からなる言語群である。アルタイ諸言語と日本語とは、対格型の格標示、膠着的形態構造、後置詞構造卓立、SOV 語順といった系統関係によらない共通の文法的特徴を有し、それゆえに「アルタイ型」言語という一つの類型として扱われることがある。

主語卓立／主題卓立という類型論的枠組みは Li and Thompson (1976) [1] によって提案されたものである。 Li and Thompson (1976) [1] は主語を「述語との (文法的) 関係によって規定される」、主題を「述語との (文法的) 関係によって規定されない」という特徴によって定義し、主語卓立／主題卓立という類型を提案している。この類型論的枠組みでは、英語は主語卓立型、中国語は主題卓立型、日本語は主格標識「が」と主題標識「は」を有するなどの根拠により主語・主題卓立型に分類されている。

Li and Thompson (1976) [1] の類型論的枠組みに対しては主に二点の問題点を指摘することができる。まず一点目として、Li and Thompson (1976) [1] は主格標識と主題標識が別個に存在することを日本語が主語・主題卓立型に分類する根拠としているが、(1) 日本語には主格標識「が」があるのに対しアルタイ諸言語には一般に主格の標識がない、(2) 日本語には述語への主語人称標示がないのに対しアルタイ諸言語 (特にチュルク諸語) では義務的な人称標示が行われる、という重要な相違点が存在する点である。さらに近年の研究 (木部・竹内・下地 (編) 2022) [2] では日本語の話し言葉や一部の方言において主語が無助詞で表されるのが無標であるなど、琉球語を含めた日琉諸語全体での格標示の多様性が明らかにされている。格助詞「が」の存在は Li and Thompson (1976) [1] の類型論的枠組みでは日本語の主語卓立性の根拠とされていることから、話し言葉や一部の方言において格助詞「が」が存在しないことは Li and Thompson (1976) [1] の類型論における日本語の類型の再検討が必要であることを意味する。

二点目は、「主語・主題卓立型」の類型の言語において、主語卓立的特徴と主題卓立的特徴がどのようにして一つの文法体系に共存しているのか、という点である。

本研究では、主に以上の二点の点から対象言語の多様性の解明と、より精緻な類型論的枠組みの提案を目指している。

2　研究方法

2.1　対象とする言語

本研究では、アルタイ諸言語のうちシベ語（ツングース諸語）とトルコ語（チュルク諸語）、日本語諸方言のうち青森県弘前方言と茨城北部方言を主な対象とする。対象言語はそれぞれ主語卓立・主題卓立の類型論的観点からみて多様な文法的特徴を持つ。

- シベ語：二つの主題標識 =ni’ と da を持つ一方、主格はゼロで標示される。主格標識がゼロとし

て存在するか、存在しないかは主語卓立・主題卓立の類型において重要な特徴である。

- トルコ語：シベ語と同様に主格がゼロで表示されるが、シベ語とは異なり述語動詞に主語人称が標示される。また焦点を表す分節的形式を持つ一方、動詞直前位置という統語的手段によっても焦点が標示される。
- 日本語青森県弘前方言：主格・目的格ともにゼロ標示が最も頻度が高く、主格と混同されそうな目的語にだけ特別な標識が現れる (大槻, 2018) [3]。また東京方言には頻出する主題標識の頻度も低い。
- 日本語茨城方言：目的格は、「=o」とゼロ標示が同程度の割合で現れ、目的語となる名詞句が動詞と隣接していない場合、および目的語となる名詞句の特定性が高い場合（当該の名詞句が指示詞の場合や、修飾を伴っている場合）「=o」の出現率が極めて高くなる（Kibe et al. 2020 [4]）

2.2 使用するデータ

本研究では、対象言語の談話を録音・録画し書き起こしたものに主語卓立・主題卓立の文法的特徴の分析に必要な情報を追加する形でデータを作成している。元になる談話資料はシベ語とトルコ語については発表者が話者との協働により採録した対話データを使用している。また日本語諸方言については日本語諸方言コーパス (COJADS) https://www2.ninjal.ac.jp/cojads/index.html のうち公開されたデータを使用している。

2.3 コーディングのフォーマット

本研究では、上記のデータにまず情報構造に直接的に関わる以下の情報をコーディングしている。

- 節内部における名詞句の文法関係：主語・直接目的語を中心に他の文法関係についてもコーディングを行っている。
- 情報構造上（定性）のステータス：Prince (1981) [5] による Given/New Taxonomy, Chafe (1994) [6] による Activation State のコーディングを行っている。
- 情報構造標識：対象とする各言語において主題標識、焦点標識とされているもののコーディングを行っている。

これらの情報は互いに独立してコーディングされており、それぞれの組み合わせの頻度を算出することによりある情報構造標識がどのような文法関係の要素に付加されているか、どのような情報構造上のステータスの要素に付加されているか、情報構造上のステータスと文法関係との間に相関関係があるか、といった情報を得ることが可能である。

情報構造上のステータスと文法関係の相関関係は、新情報が他動詞の直接目的語ないし自動詞の主語で現れやすく、旧情報が他動詞主語で現れやすいという Du Bois (1987) [7] の研究がよく知られているが、このような相関関係が対象とする言語に等しく見られるか否かはそれぞれの言語の主語卓立・主題卓立的特徴を考える上で重要な指標である。

このほか、対象の各言語ごとに情報構造に関連するとされる文法的特徴をメンバー間で共有し、他の対象言語でもコーディングを行うための議論を行っている。詳細は 3 節で述べる。

2.4 データマネジメント

本研究では各メンバーがそれぞれの対象言語について独自に蓄積したデータに本研究に必要なデータを追記する形をとっている。このような研究手法から、メンバーが扱うデータのフォーマットがそれぞれに異なっており（Excel/CSV、Praat、ELAN）、言語間の比較対照を行う際のデータ処理の障壁となることが懸念される。このような問題に対し、本研究では ELAN のデータフォーマット（XML をベースとした独自仕様）と ELAN の CSV データインポート/エクスポート機能を中心に、Excel/CSV-ELAN-Praat 形式のデータを一元化できるよう、データフォーマット変換プログラムを作成して対処している。

3 コーディングに際しての問題

本節ではコーディング作業において直面した問題を報告し、情報構造に関する談話データのコーディングの方法論について議論する。

3.1 省略

まず、省略された要素をどのようにコーディングするか、という問題が存在する。情報構造の観点からみると、ある要素が省略可能であるのはその指示対象が話し手と聞き手の双方にとって同定可能である（旧情報である）場合に限られる。また、Du Bois (1987) [7] は新情報は他動詞目的語または自動詞主語として現れ、旧情報は他動詞主語として現れる傾

向にあると述べていることから、どの文法関係の要素が省略されているかを見ることにより新情報・旧情報という情報構造上のステータスと文法関係に関係があるかどうかを検討することが可能である。

しかし、実際の談話データについて省略された要素をコーディングするにあたっては検討すべき問題が存在する。これは大きく、(1) 本来存在するはずの要素が省略されているのか、それとも元々存在しないのかをどのように判断するか、という問題と、(2) 省略された要素をどの位置にコーディングするかという問題に分かれる。

以下の日本語の例 (1) において、動詞「食べる」が対格の目的語を取ると考えると、(1b) の応答の文「うん、食べたよ」についても「うん、[ごはんを] 食べたよ」と対格の目的語が省略されていると考えることは可能である。

(1) a. ごはんを食べた。
b. ごはん食べた？ — うん、食べたよ。

しかし、以下の (2) の場合、「食べる」の目的語として何かが省略されているから、また「飲んで」についても目的語が省略されているとすると、省略された要素が膨大に想定することになり、実際のデータと乖離が大きくなるように見える。

(2) 正月は食べて飲んで寝て過ごした。

このような例を見ると、省略とコーディングすべきでない要素がある、という可能性を検討する必要がある。

また、省略された名詞句が存在すると判断し、それをゼロなどの形でコーディングしようとする場合には、その「省略された」名詞句のもとの位置を判断し難い。例えば、日本語の「太郎がたべたよ」という文において「ごはんを」という名詞句が省略されていると考えると、少なくとも以下の２つの構造が考えられる。

(3) a. 太郎が（ごはんを）食べたよ。
b.（ごはんを）太郎が食べたよ。

このような場合、名詞句の位置を特定しようとするよりも、述語（動詞）「食べ (る)」に対し「主語省略」「直接目的語省略」という形でコーディングするのが有効かもしれない。

シベ語とトルコ語にも省略された要素を想定すべきか否か、という問題が存在する。これらの言語では、所有物を表す名詞句に後続して所有者の人称を表す要素が存在する。

シベ語において接語 =ni’ は所有物を表す名詞句に後続し、3 人称の所有者を表す。以下の (4) において =ni’ は xenahce’「おじの妻」の所有者が moN bo=i guruN「うちの妻」であることを表している。

(4) tume=da moN bo=i guruN
そうして=主題 1 複. 除外 家=属格 人々
xenahce’=ni’ eN taw
おじの妻=主題 間投詞 ちょうど
se-me=da tere yinglaN
いう-副動詞. 同時=主題 それ （人名）
se-Xe nane.
いう-完了 人
「それで、うちの妻のおじの妻がちょうどその英蘭という人だ。」

トルコ語にも同様の機能をもつ接辞-I が存在する。以下の (5) において-ıは fiyat「値段」の所有者が kravat「ネクタイ」であることを表している。

(5) kravat-ın fiyat-ı
ネクタイ-属格 値段-所有
「ネクタイの値段」

しかし、これらの要素は以下の (6,7) のように所有者を表す名詞句が同一文中に現れることなく用いられる場合がある。

(6) tese=ni’ bo eme hale guruN.
彼ら=所有 1 複. 除外 ひとつ 姓 人々
eme hale guruN o-ci’ bo=ni’ meji’
ひとつ 姓 人々 なる-条件 家=所有 やや
haNci.
近い
「彼らは、我々は一つの姓（氏族）の人だ。一つの姓（氏族）の人なので家がやや近い。」

(7) Bu kravat çok güzel. Fiyat-ı ne
この ネクタイ とても よい 値段-所有 何
kadar?
ほど
「そのネクタイ、いいですね。（そのネクタイの）値段はいくらですか？」

このような場合でも当該の要素が所有者を表すと仮定すると、どこかに所有者名詞句が存在すると

考える必要がある。このとき、同一文内に存在する所有者名詞句が省略されていると分析しコーディングすることが可能かもしれない。しかし、シベ語の=ni’ については、児倉 (2007) [8] は談話全体のトピックの所有物に後続して二次的な談話の主題を表す、つまり=ni’ の表す所有関係は同一文内の要素の間ではなく談話全体で形成されると述べていることから、同一文中に所有者名詞句が省略されているという分析は妥当でない可能性がある。

このように、ある要素が省略されているか否かという判断はしばしば困難であり、どのようにコーディングすべきか、という問題が存在する。

3.2 照応

照応は言語処理の分野でも研究が進められている [9] が 、本研究でも照応関係をどのように判断しコーディングするべきか、という問題が存在する。3.1 で取り上げたシベ語やトルコ語の所有標識は所有者との照応要素であると考えると、所有者が何であるか、どの名詞句と照応しているかという情報をコーディングする必要がある。この照応関係は、3.1 でみた (4, 5) のように所有者ー所有物の所有関係を特定しやすい場合には問題とならないものの、(6, 7) のように所有者名詞句が同一文中に存在しない場合にはどの要素と照応関係にあるのか、そもそも他の要素との照応関係があるのかどうかの判断すら困難である場合がある。

日本語でも照応関係の有無が問題となる場合がある。Nakagawa (2020) [10] が指摘するように、日本語の「は」はそれ自体が談話に導入済みでなくとも、他の談話に導入済みの要素から推論、想起可能な要素であれば用いることが可能である。以下の (8,9) において、(8) ではバスから運転手が比較的容易に想起されるのに対し、(9) では一般人の自宅から消防車を想起することが比較的困難であるため、(9) の容認度が (8) に比べて低くなっている。

(8) （バスに乗ったら）運転手は酔っ払っていた。

(9) （自宅の話で）??消防車は新しい。

これらの例は照応関係が情報構造要素の使用に影響しており、コーディングする必要がある一方で照応関係が存在するか否かの判断が困難であることを示している。

3.3 統語的位置

名詞句の統語的位置も情報構造に関わることが知られており、この統語的位置をどのようにコーディングするかも問題となる。例えばトルコ語では焦点となる要素が述語動詞の直前の位置に置かれる。

(10) Naomi Türkçe bil-iyor.
（人名） トルコ語 知っている-現在
「奈緒美はトルコ語がわかる。」

日本語でも談話で初出の要素が述語の直前に現れやすいことが Nakagawa (2020) [10] により指摘されている。これらのことから、ある要素が動詞の直前にあるか否かという情報をコーディングする必要がある。

3.4 韻律

プロミネンスやピッチパターンなどの韻律情報も情報構造に関わると考えられるため、コーディングが必要である。例えば、茨城県北部方言では疑問詞などの焦点の有無がピッチパターンの実現に影響を及ぼす。以下の例では、{ } で表した韻律単位に山型のピッチパターンが生じる。

(11) {オナミカ゜ }{リンコ゜ クッタヨ}

(12) {ダレカ゜ リンコ゜ クッタノ}

韻律的特徴のコーディングの方法としては ToBI (Tones and Break Indices) が提案され、英語、日本語、韓国語、ドイツ語などに応用されている。ToBI では音調を H/L 等のトーンの連続と捉え、それらのトーンを韻律的単位に基づき発話の要所要所に配置していく。日本語の自然発話のコーディングに応用した X-JToBI も発案されている。

コーディングのためには、当該言語においてどういったトーンのセットが存在するか、韻律単位がどういった階層を成しているかを明らかにしておく必要があるが、茨城県北部方言では未だ明確な分析ができておらず、コーディングの必要性はあるものの、まだよい方法ができていない。

4 おわりに

以上、本発表では情報構造の類型論に向けた談話データのコーディングにおける問題点を報告した。これらの課題について引き続き議論をしつつコーディングを進めるのが本研究の課題である。

謝辞

本研究は JSPS 科研費 JP24K00061, JP24K03836, JP18K12395 の助成を受けたものです.

参考文献

[1] Charles Li and Sandra Thompson. Subject and topic: A new typology of language. In Charles Li, editor, **Subject and Topic**, pp. 457–489. New York: Academic Press, 1976.

[2] 木部暢子, 竹内史郎, 下地理則（編）. 日本語の格表現.

[3] 大槻知世. 青森県津軽方言の情報の表示をめぐって. Phd thesis, 東京大学, 2018.

[4] Nobuko Kibe, Kumiko Sato, Taro Nakanishi, and Kohei Nakazawa. Corpus-based study of japanese dialects regional differences in accusative case marking system. In **Proceedings of Methods XVI**, pp. 197–207, 2020.

[5] Ellen Prince. Toward a taxonomy of given-new information. In Peter Cole, editor, **Radical pragmatics**, pp. 223–256. New York: Academic Press, 1976.

[6] Wallace L. Chafe. **Discourse, Consciousness, and Time**. Chicago: The University of Chikago Press, 1994.

[7] John W. Du Bois. The discourse basis of ergativity. **Language**, Vol. 63, No. 4, pp. 806–855, 1987.

[8] 児倉徳和. シベ語の名詞接尾辞-ni についての若干の考察. 満族史研究, Vol. 6, pp. 141–161, 2007.

[9] 笹野遼平, 飯田龍（共著）, 奥村学（監修）. 文脈解析：述語項構造・照応・談話構造の解析. 自然言語処理シリーズ ; 10. コロナ社, 2017.6.

[10] Natsuko Nakagawa. **Information structure in spoken Japanese**. No. 8 in Topics at the Grammar-Discourse Interface. Language Science Press, Berlin, 2020.