日本語話し言葉コーパスを用いた情報構造のタグ付けとその分析方法の提案

中川 奈津子[†‡]
nakagawanatuko@gmail.com
[†]京都大学大学院 人間・環境学研究科
[‡]ニューヨーク州立大学 バッファロー校 言語学科

1　はじめに

本研究では、『日本語話し言葉コーパス』の模擬講演を用い、情報構造のタグ付けとその分析手法を提案する。

情報構造 (information structure) は理論言語学、歴史言語学や類型論、自然言語処理など、幅広い分野で問題となる。理論言語学では情報構造の定義やその心的な位置づけ、歴史言語学では情報構造が言語変化に及ぼす要因や変化のパターン、類型論では情報構造の標示手段の傾向を明らかにすることなどが中心的な課題である (e.g., [1, 3, 6])。一方、自然言語処理の分野では、話題抽出や文書要約などのために情報構造研究が役立つと考えられる。

情報構造の研究では、ある談話において話題やそれに追加される新情報 (焦点) が表現される言語的・非言語的な手段を調べ、またそのような情報の心的な状態をモデル化することを目標としている。このためにはまず、どの言語コーパスにでも適用可能な話題や焦点のタグ付け方法が問題となる。本研究は、言語学の観点から情報構造と言語的・パラ言語的形式の関連を明らかにすることを最終的な目的とし、そのためのタグ付けと分析手法を提案する。

2　コーパス

コーパスは『日本語話し言葉コーパス』([8], 以下 CSJ) の模擬講演を使用した。表 1 に本研究で使用した模擬講演一覧をまとめる。話し言葉を研究対象とするのは、従来の研究では書き言葉を対象としたものが多く話し言葉を対象としたものが少ないからであり、また書き言葉には現れないイントネーション、言いよどみなど情報構造のオンラインの処理を調べる豊富な手がかりを含んでいるからである。会話ではなく独話を対象としたのは、会話のタグ付けはすでに試みられており ([7, 9])、将来的にはこれと比較する意図からである。

表 1: 利用した模擬講演

ID	性別 (年齢)	テーマ	長さ (sec)
S00F0014	F (30-34)	ハワイ旅行	1269
S00F0209	F (25-29)	ピアニスト	619
S00M0199	M (30-34)	コソボ紛争	580
S00M0221	M (25-29)	サラ金	654
S01F0038	F (40-44)	人生の幸運	628
S01F0151	F (30-34)	ヒマラヤ	765
S01M0182	M (40-44)	ボクシング	644
S02M0198	M (20-24)	犬の死	762
S02M1698	M (65-69)	犬の死	649
S02F0100	F (20-24)	難病	740
S03F0072	F (35-39)	イランでの 1 年	816
S05M1236	M (30-34)	茂原の思い出	832

ある。

3　タグ付け

本節では情報構造のタグ付けの手順を概観する。まず照応関係のタグ付けを行い (3.1)、情報構造タグは照応関係タグから自動生成する (3.2)。

3.1　照応関係のタグ付け

日本語の照応関係のタグ付け手順はすでに書き言葉でも話し言葉でも提案されている ([5, 7]) ので、それにしたがってタグを付与する。本研究では話し言葉を対象としているため、主に [7] に沿ってタグ付けを行った (一部変更がある)。具体的な手順は以下である。

(1) a. 述語の認定
　　b. 項構造とゼロ代名詞の同定: それぞれの述語に関してガ、ヲ、二格に相当する名詞を同定し、述語にこの要素が必要だが欠けているも

Copyright(C) 2013 The Association for Natural Language Processing.
All Rights Reserved.

のをゼロ代名詞とした。

c. 談話要素 (名詞・代名詞・ゼロ代名詞) の分類: ある名詞の指示対象を話し手、聞き手、外界照応、一般事物などに分類した。本研究では一般事物を指す名詞を対象とする。

d. 照応関係の同定: すべての名詞に関して、その指示対象が以前に言及されていれば先行詞を同定した。

[5] では連想照応や、共参照と照応の区別などの情報も含まれているが、本研究ではこのような情報はまだ付与しておらず、今後の課題とする。

3.2 情報構造タグの生成

次に照応関係タグから情報構造タグを生成する ([7])。情報構造タグは大まかに、ある談話要素とそれ以前の談話要素の関係に関するタグ (情報の新旧) と、それ以後の談話要素の関係に関するタグ (継続性) の 2 種類に分けられる。旧情報で継続的なものが話題性 (topicality) が高いと考える ([4])。

情報の新旧 (information status) は、新情報 (new)・旧情報 (given) の 2 種類がある。先行詞を持たない談話要素を新情報とし、それ以外のものを旧情報とする。

継続性 (persistence) には継続的 (persistent)・非継続的 (non-persisntent) の 2 種類をもうける。談話内で指示対象が 2 回以上言及されるものは継続的、そうでないものは非継続的である。また、継続性を図る別の因子として、継続回数 (number of mentions) ももうけた。これは談話要素が言及された、それ以後に、その指示対象が何度言及されたかを示す。

4 分析

タグ付けの結果、7697 の談話要素を同定した。このうち話し手・聞き手などを除外すると残りは 4614 で、これが分析の対象となる。

本節では、情報構造と助詞 (4.1)、語順 (4.2)、韻律 (4.3) の関連をそれぞれ分析する。

4.1 助詞

この節では助詞と情報構造の関連を見る。ここでは、結果が明確に出た、いわゆるトピック・マーカーと情報の新旧・継続回数の関連を紹介する。図 1 に示すように、トイウノハで言及された要素は、ハやモよりもわずかながらに旧情報が多い。また、図 2 に示すように、トピック・マーカー (トイウノハ、ハ、モ) のうち、トイウノハで言及された要素はその後続けて言及されることが最も多い。これはトイウノハが話題性の高い談話要素を標示していることを示すと考えられる。

4.2 語順

図 3, 4 はそれぞれ、語順と情報の新旧、継続性の関係を示したものである。語順は、問題の談話要素を含む文節が文中の何番目にあるかを示す、CSJ に付随する情報 (nth) を利用した。1 は文頭にあることを示し、2 は 2 番めにあることを示す。21 以上は 21+としてひとまとめにした。

図からわかるとおり、文頭には旧情報と継続的な談話要素の割合がやや高い。これは「トピックは文頭に現れやすい」という伝統的な観察が実際に定量的に示されたものであるといえる。

4.3 韻律

最後に、情報構造と韻律の関連を見る。ここでは、韻律単位の基準として [2] を採用する。[2] は、揺れのない明確な手続きによって長・短二種類の発話単位を認定する基準を提案しており、この手続きによって CSJ の模擬講演にもほぼ自動で韻律単位を付与することが可能である。まず「長い単位」(統語・語用論的な単位) を認定し、「長い単位」を分割する形で音響・韻律的な基準で「短い単位」(韻律単位) を認定する。韻律単位は、0.1 秒以上の休止とピッチリセットによって分割される。

韻律単位の構成要素のうち、ここでは述語とその項に注目し、項となる談話要素が述語と同じ韻律単位内で発話されているか (節型: (2-a))、異なる韻律単位で発話されているか (句型: (2-b)) の 2 種類を区別する。(2) はこの区別を図式的に表したもので、1 つの 枠で囲まれた部分 が 1 つの韻律単位に相当する。

(2) a. 節型 (clausal unit) : 項　述語

b. 句型 (phrasal unit) : 項　述語

図 5 はこの 2 種類の韻律単位と継続回数の関係を示す。図 5 からは特に関連があるようには見えないが、これをトピック・マーカーで標示された主語に限定すると、継続回数が多いほど句形の韻律単位が多くなるという関係がありそうである。

5 おわりに

本稿では CSJ の模擬講演の照応関係のタグ付けを行い、そこから情報構造タグを生成しそれを分析する手法を提案した。分析結果の統計分析、独話と会話の比較などは今後の課題とする。

Copyright(C) 2013 The Association for Natural Language Processing.
All Rights Reserved.

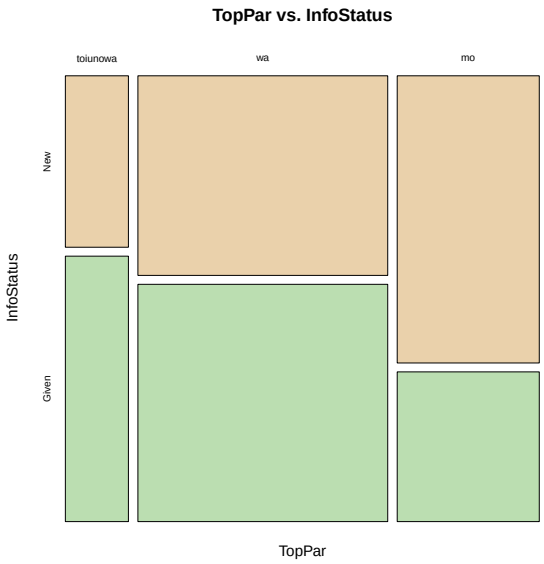

図 1: トピック・マーカー vs. 情報の新旧

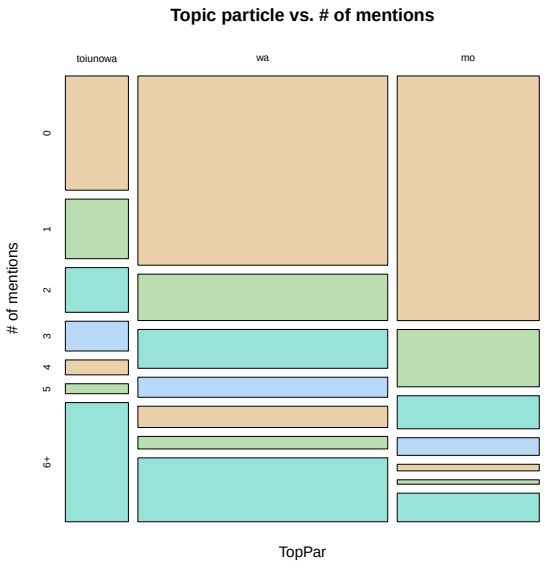

図 2: トピック・マーカー vs. 継続回数

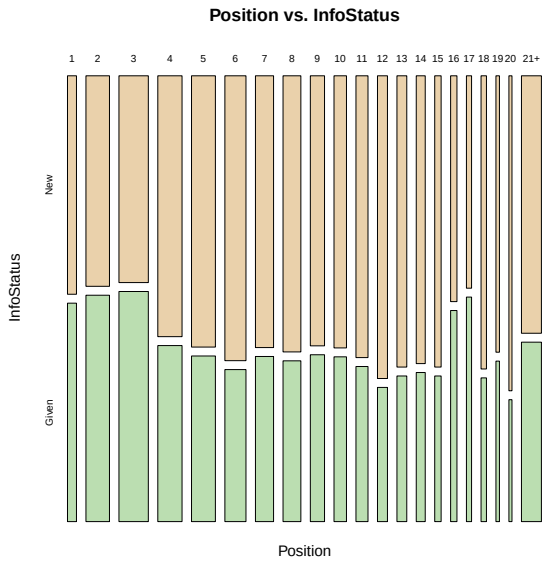

図 3: 談話要素の位置 vs. 情報の新旧

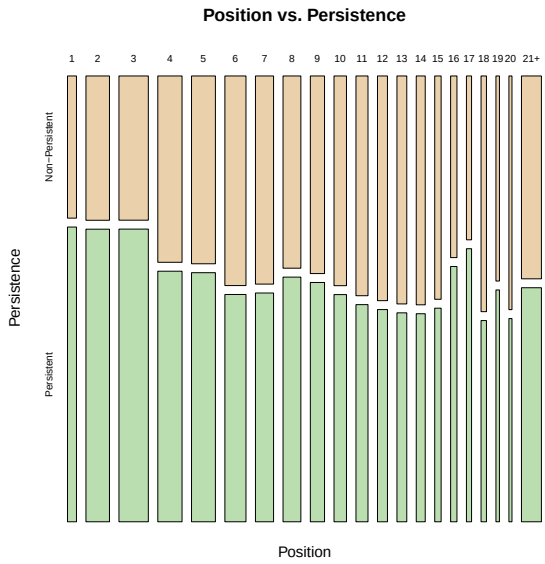

図 4: 談話要素の位置 vs. 継続性

Copyright(C) 2013 The Association for Natural Language Processing.
All Rights Reserved.

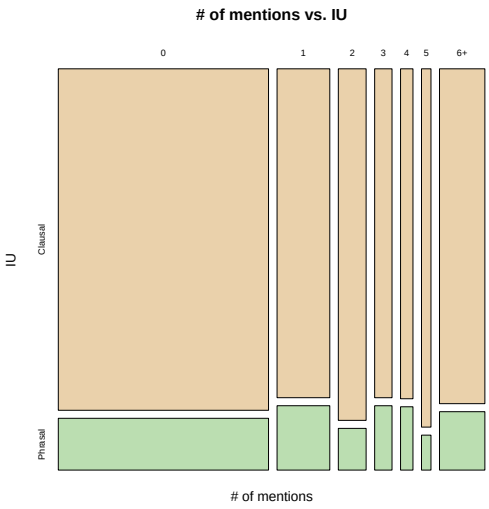

図 5: 韻律 vs. 継続回数 (全体)

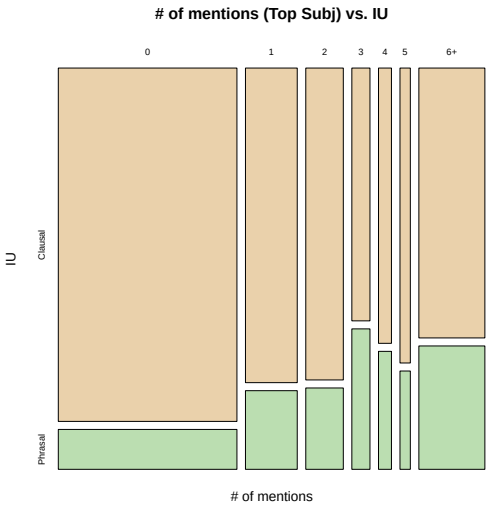

図 6: 韻律 vs. 継続回数 (トピック主語)

参考文献

[1] Sasha Calhoun, Mark Nissim, Malvina Steedman, and Jason Brenier. A. framework for annotating information structure in discourse. In *Frontiers in Corpus Annotation II: Pie in the Sky*, Michigan, 2005. Association for Computational Linguistics.

[2] Y. Den, H. Koiso, T. Maruyama, K. Maekawa, K. Takanashi, M. Enomoto, and N. Yoshida. Two-level annotation of utterance-units in Japanese dialogs: An empirically emerged scheme. In *Proceedings of the 7th Language Resources and Evaluation Conference (LREC2010)*, pp. 2103–2110, Valletta, Malta, 2010.

[3] Stefanie Dipper, Michael Götze, Manfred Stede, and Tillmann Wegst. ANNIS: a linguistic database for exploring information structure. In S. Ishihara, M. Schmitz, and A. Schwarz, editors, *Interdisciplinary Studies on Information Structure 01*, pp. 245–279. Potsdam, Universtätsverlag Potsdam, 2004.

[4] T. Givón, editor. *Topic Continuity in Discourse*. John Benjamins, Amsterdam, 1983.

[5] R. Iida, M. Komachi, N. Inoue, K. Inui, and Y. Matsumoto. 述語鋼構造と照応関係のアノテーション: NAIST テキストコーパス構築の経験から. 自然言語処理, Vol. 17, No. 2, pp. 22–50, 2010.

[6] Knud Lambrecht. *Information Structure and Sentence Form: Topic, Focus and the Mental Representations of Discourse Referents*. Cambridge University Press, Cambridge, 1994.

[7] N. Nakagawa and Y. Den. Annotation of anaphoric relations and topic continuity in Japanese conversation. In Nicoletta Calzolari et al., editor, *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC2012)*, pp. 179–186, Istanbul, 2012.

[8] 前川喜久雄. 『日本語話し言葉コーパス』の概要. 日本語科学, Vol. 15, pp. 111–133, 2004.

[9] 中川奈津子, 伝康晴. 対話における情報構造と韻律・統語構造の関係の分析. 人工知能学会研究会資料 SIG-SLUD-B201, pp. 43–48, 2012.

Copyright(C) 2013 The Association for Natural Language Processing.
All Rights Reserved.