対話研究にふさわしい統語的単位の認定基準 —対話節単位の設計—

丸山 岳彦 *　　高梨 克也 †　　吉田 奈央 ‡ *

* 国立国語研究所 言語資源研究系 / コーパス開発センター
† 京都大学 学術情報メディアセンター
‡ 東京工業大学大学院 総合理工学研究科

1　はじめに

筆者らは、平成 20 年度から科研費基盤研究（B）「対話における発話単位と機能の認定に関する研究」を開始している。このプロジェクトは、対話において、文や発話に相当する単位とそれが担う機能を、揺れのない明確な手続きによって認定する基準を策定し、公開することを目的としており (伝ほか, 2008) 、統語・韻律・語用論という異なる観点から、それぞれ独自に発話単位を認定し、各単位の相互関係から対話研究にふさわしい発話単位を規定するという手順を採っている。

本研究では、上記のプロジェクトの一環として、対話研究にふさわしい統語的な発話単位「**対話節単位（**Dialog Clause Unit**）**」を認定する基準について示す。

2　先行研究

対話の分析単位には、これまでにも、様々な単位が提案されている。例えば、一定以上の無音区間の持続によって単位を区切る「間休止単位（IPU）」は最も安定的に単位を認定することができるが、しかし必ずしも統語的・意味的に望ましい単位が取り出されるとは限らない。また、話者交替の生じる可能性に着目して単位を区切る「ターン構成単位（TCU）」(Sacks, Schegloff, & Jefferson, 1974; 坊農・高梨, 2009) は、統語情報や韻律情報が単位の認定に影響はするものの、第一義的には、話し手と聞き手による相互行為としての単位という側面が強く、統語的単位と見なすことはできない。

独話を対象とした統語的単位として、『日本語話し言葉コーパス（以下、CSJ と記す）』の構築過程で設計された「**節単位（CU）**」がある (丸山・高梨・内元, 2006; 坊農・高梨, 2009)。これは、独話における談話レベルの文法情報を付与するための基本単位として設計された、統語的単位である。ただし、対象が独話であることを前提としているため、対話における聞き手との相互行為の存在は想定されていない。

そこで本研究では、対話研究にふさわしい統語的単位を設計することを目的として、従来の節単位を対話分析用に拡張した「**対話節単位（DCU）**」を提案する。

3　「節単位」の考え方

3.1　節単位とは何か

節単位とは、CSJ の構築時に設計された発話単位であり、係り受け構造情報、重要文情報、談話境界情報など談話レベルの情報を付与するための単位として考案された。基本的には特定の節境界の直後で発話を分割することによって得られる「節」が節単位を構成するが、非流暢現象などが原因となり、単独の名詞句や言いさされた部分、発話の断片などが節単位を構成することもある。

そもそも、1 人の話し手が発話し続ける独話という発話様式は、最初に想定された発話プランに従って、統語的に適格な構造を備えた言語形式を漸進的に（実時間上に）構築していこうとする一回的な行動である。しかし、発話プランの動的な変更や発話生成過程におけるトラブルなどにより、適格な統語構造が破綻することも頻繁に起こる。このような発話産出の結果から統語的単位を取り出す際に手がかりになるのが、**節境界（**Clause Boundary**）**である。SOV の語順を持つ日本語の場合、基本的には、述語句の出現によって統語的・意味的な完結点が生じる。そこで、様々な形態を取る述語句の末尾（＝節境界）の位置と種類を認定し、発話の分割点として柔軟に利用することで、統語的にも意味的にもまとまりを持つ単位を取り出すことができる。

CSJ での節単位認定は、形態論情報を入力とする自動節境界解析の結果を、人手で修正するという手順が採られた。節境界解析によって検出される節境界は 49 種類あり、統語的な切れ目の大きさという観点から、「絶対境界」「強境界」「弱境界」という 3 段階の別が設けられている。このうち、絶対境界と強境界をデフォルトの発話分割位置とした上で、必要に応じて弱境界や節境界以外の箇所でも発話を分割するなど、あらかじめ定められた手順に従って、人手による修正を実施した。

3.2 対話節単位の認定基準

本研究では、独話用に設計された節単位を、対話用に拡張することを考える[*1]。先に述べたように、独話は 1 人の話し手が発話を産出し続けるという様式を取り、聞き手からのフィードバックは極めて少ない。一方、対話の場合、話し手による一方的な語り（narrative）だけでなく、話し手と聞き手の相互行為によって談話が動的に構築されていくという側面が生じる。また、あいづちを独立した単位と見なすかどうか、言いさし表現やいわゆる「引き取らせ」の形式をどう扱うかなど、対話に特有な現象に対処する必要がある。

そこで、対話節単位のアノテーションを試行するために、以下のような認定基準を定めた。

■認定方針

- 音声を書き起こした転記テキストに対して、特定の箇所に**節境界ラベル**を付与し、その直後で発話を分割する。倒置・引用・挿入表現などによって不自然な発話分割が生じる場合、**操作記号**を付して適切に修正し、対話節単位を認定する。
- 転記テキストには、話者情報、フィラー、語断片、感動詞のタグが付いていることを前提とする。
- ラベリングは、音声を参照しながら実施する。

■節境界ラベルに関する認定基準

- 図 1 に示す節境界ラベルを、当該箇所に付与する。

■操作記号に関する認定基準

- 節境界ラベルを付与した後、図 2 に示すような理由で発話分割位置として適切でないと判断された場合、操作記号を当該箇所に付与して修正する。

[*1] ここで言う対話とは、2 者対話だけでなく、3 者対話、あるいはそれ以上による複数人対話を含む。

/AB **絶対境界**（文末表現に相当）の直後に付与する
- 行ったことあります/AB
- なんだろうね/AB
- 新婚旅行は海外って決めてるもん/AB

/SB **強境界**（従属度の低い従属節）の直後に付与する
- 有名な所ばっかりだったけど/SB
- もう松屋すぐきれるしね/SB
- 楽器としてファゴットがあるんですが/SB

/WB 発話の分割点になると判断された**弱境界**（従属度の高い従属節）の直後に付与する
- 別の名前があるらしくて/WB
- でもドイツは夜怖くなかったんで/WB
- コミュニケーションとれないからね/WB

/NB 節境界以外で**発話の切れ目となる箇所**（1 語文、体言止め、言いさしなど）の直後に付与する
- よろしく/NB
- 何で/NB
- 最高で何人とか/NB
- ウイーンよりもっと下/NB
- 勝手に日本人が命名したん/NB

/MB 発話の冒頭・末尾に現れる、**独立した感動詞**の直後に付与する
- (I_うん)/MB
- (I_あー)/MB

/FB 発話の冒頭・末尾に現れる、**独立した語断片**の直後に付与する
- (D_ハッ)/FB

図 1 対話節単位で付与する節境界ラベルの一覧

4 分析

4.1 分析対象データ

分析対象データとして、以下の 2 種類のデータから、8 対話を用いた。CSJ はインタビュー形式の対話であり、Chiba は 3 人の雑談である (Den & Enomoto, 2007)。

CSJ: CSJ 対話データ （4 対話 × 5 分、4,382 語）
Chiba: 千葉大学 3 人会話データ （4 会話 × 5 分、5,201 語）

いずれも、対話開始後 1 分から 6 分までの 5 分間を切り出して分析対象とした。転記テキストは、CSJ の転記基準を拡張し、話者の別、単語境界時間情報、フィラー (F)、語断片 (D) に加えて、「感動詞 (I)」のタグが付与されている。

倒置要素　節境界ラベルの直後に倒置要素が出現した場合、ラベルを「|AB+」と変更し、倒置要素を「<<***>>」で囲む。
– なんかもたっとしない |AB+<< あれ >>/NB

引用表現　引用表現の末尾にある節境界ラベルを「|AB+Q」と変更する。
– 生徒の中からだったんだ:|AB+Q と思います/AB
– やめようか:|AB+Q とも思ったんですが/SB

挿入表現　挿入表現の末尾にある節境界ラベルを「|AB」と変更し、挿入表現の範囲を、「{I***}+」で囲む。
– 後はちょっと {I 何だろう |AB}+ 小心者:だから
– 工場によってまた (F_あの:){I 何て言うかな |AB}+ 持ち場って言うか仕事内容なんかは違うらしいですけどね/SB

試行的提示[a]　試行的提示の末尾にある節境界ラベルを「|AB」と変更し、試行的提示の範囲を、「{T***}+」で囲む。
– バイト先でもう十年目 {T ぐらいかな |AB}+ なる人がいんの |AB+<< バイトで >>/NB

[a] 「試行的提示」とは、「発話に際して用いる表現の選択に躊躇があり、とりあえずある表現を選んで発話した（＝試行的に提示した）」現象を指す (北野, 2005)。

図 2　対話節単位で付与する操作記号の一覧

4.2　ラベル付与

図 1、2 に示した基準に従って、対象データに対して節境界ラベルおよび操作記号の付与を行なった。作業は当初 2 名の作業者で同時に実施し、一致しなかった箇所を協議の上で統一した。作業結果がほぼ一致するようになった段階で、作業者 1 名の作業結果をもう 1 名が添削する方式に変更した。

4.3　分析 1: 節境界ラベルの種類

付与された節境界ラベルの種類と数を、表 1 に示す。

表 1　付与された節境界ラベルの種類と数

	CSJ		Chiba	
/AB	160	(21.80%)	245	(21.29%)
/SB	41	(5.59%)	50	(4.34%)
/WB	50	(6.81%)	84	(7.30%)
/NB	99	(13.49%)	200	(17.38%)
/MB	334	(45.50%)	470	(40.83%)
/FB	50	(6.81%)	102	(8.86%)
合計	734	(100%)	1,151	(100%)

節境界ラベルの数を見ると、CSJ、Chiba とも、/MB が圧倒的に多いことが分かる。これは、「あいづち」として機能する感動詞であり、対話でいかにあいづちが多用されているかを示している。それ以外の各節境界については、CSJ と Chiba の間でそれほど大きな開きは見られない。独話ではあいづちは原則的に生起しないことから、独話と対話の統語的単位の違いを考える上で最も考慮しなければならないのは、あいづちの扱いであると言える (吉田・高梨・伝, 2009)。

4.4　分析 2: 操作記号の種類

次に、付与された操作記号の種類と数を、表 2 に示す。

表 2　付与された操作記号の一覧

	CSJ	Chiba
倒置要素	6	37
引用表現	44	58
挿入表現	3	4
試行的提示	0	3

倒置要素が、CSJ に比べて Chiba で多く出現している。丁寧な口調が用いられているインタビュー形式の CSJ に比べて、友人同士の雑談である Chiba に倒置が多く現れるということは、発話場面のフォーマリティが倒置の出現に影響していると考えることができる。

4.5　分析 3: 対話節単位の長さ

付与された節境界ラベルの位置で発話を分割し、対話節単位を得た。各単位中に含まれる語数（CSJ の短単位に相当する単位）を、対話節単位の長さとして計測した。結果を、分割位置となった節境界ラベルの種類ごとに、表 3、4 に示す。

表 3　対話節単位の長さ（CSJ）

	平均値	中央値	四分位範囲	最大値	最頻値
/AB	11.3	9	11	77	3
/SB	19.8	14	14	63	14
/WB	13.0	11.5	7	35	11
/NB	6.1	3	4.5	39	2
/MB	1.3	1	0	7	1
/FB	1.1	1	0	2	1

表 4　対話節単位の長さ（Chiba）

	平均値	中央値	四分位範囲	最大値	最頻値
/AB	7.5	5	5	47	4
/SB	11.3	9.5	5.75	31	9
/WB	12.6	10	9	51	7
/NB	4.8	2	5	57	1
/MB	1.4	1	0	7	1
/FB	1.3	1	0	5	1

インタビュー形式の CSJ の方が、雑談形式である Chiba よりも 1 単位が長くなる傾向にある。これは、インタビューに答える際の語りが CSJ に多く含まれるためであると考えられる。/AB の最頻値は「3、4」であるが、大半が「そうですね」「そうなんだ」のような定型的な応答表現であった。これらの応答表現や、「違う」「わかんない」のような/AB 形式の短い発話は、会話連鎖上の第 2 位置の発話や、連鎖終結の第 3 部分 (坊農・高梨, 2009) において用られているものと考えられ、独話には見られない、対話に特有の現象であると言える。

4.6 独話節単位と対話節単位

最後に、対話節単位を CSJ の独話に付与された節単位（独話節単位）と比較する。2 種の対話データと独話データ（CSJ コアに含まれる学会講演・模擬講演、177 講演）について、発話分割点となった節境界の種類（ただし/FB と/MB を除く）と数を、表 5 に示す。なお、/AB は独話節単位で言う絶対境界、/SB は強境界、/WB は弱境界、/NB は述語以外の境界で終わる単位に、それぞれ相当する。また、独話節単位の長さを、表 6 に示す。

表 5　独話と対話における節単位末の種類と数

	CSJ（対話）		Chiba		CSJ（独話）	
/AB	160	(45.71%)	245	(42.31%)	9,561	(52.55%)
/SB	41	(11.71%)	50	(8.64%)	5,605	(30.81%)
/WB	50	(14.29%)	84	(14.51%)	1,534	(8.43%)
/NB	99	(28.29%)	200	(34.54%)	1,495	(8.22%)
合計	350	(100%)	579	(100%)	18,195	(100%)

表 6　独話節単位の長さ（CSJ 独話）

	平均値	中央値	四分位範囲	最大値	最頻値
/AB	28.09	24	22	147	17
/SB	21.53	18	19	148	13
/WB	22.34	18	19	115	1
/NB	13.48	9	19	111	2

CSJ の独話における節単位の認定基準 (丸山ほか, 2006) が対話節単位とは異なる[*2]ため厳密な比較にはならないが、それでもなお、独話節単位が対話節単位よりも長くなる傾向を見て取ることができる。

表 5 を見ると、/AB と/SB は独話で多く出現し、/WB と/NB は対話で多く出現していることが分かる。これは、文末表現で個々の発話単位を完結させながら複数の発話単位をつなげていく独話と、統語的には未完結な形式や発話の断片などでもターンを構成できる対話との差によるものであると考えられる。

また、表 3、4 および表 6 の節単位長を比較すると、/AB の長さにおいて、独話が対話を大きく上回っている。この差は、独話が学会発表やスピーチのように語りが主体である一方、対話には定型的な応答表現（4.5 節）が多く含まれることが原因であると考えられる。

これに対して、/SB の場合は、独話とインタビュー対話との間ではさほど大きな違いはない。これは、インタビューでの応答が説明を含む語りになった場合に長いターンを構成しやすく、かつそのような箇所で/SB が頻出しているためだと考えられる。

5　今後の課題

以上、対話の統語的な発話単位「対話節単位」を認定する基準について述べた。この統語的単位を用いた分析（係り受け構造、言い直しの範囲など）や、その機能に関する分析を、今後の課題としておく。

謝辞　本研究は科研費補助金基盤研究 (B)「対話における発話単位と機能の認定に関する研究」からの助成を受けています。

参考文献

坊農真弓・高梨克也 (編). (2009). 多人数インタラクションの分析手法. 東京: オーム社.

Den, Y., & Enomoto, M. (2007). A scientific approach to conversational informatics: Description, analysis, and modeling of human conversation. In T. Nishida (Ed.), *Conversational informatics: An engineering approach* (pp. 307–330). Hoboken, NJ: John Wiley & Sons.

伝康晴・小磯花絵・丸山岳彦・前川喜久雄・高梨克也・榎本美香・吉田奈央. (2008). 対話研究にふさわしい発話単位の認定に向けて. 人工知能学会研究会資料, *SIG-SLUD-A802*, 27–32.

北野浩章. (2005). 自然談話に見られる逸脱的な文の構築. 串田秀也・定延利之・伝康晴 (編), 文と発話 1: 活動としての文と発話 (pp. 91–121). ひつじ書房.

丸山岳彦・高梨克也・内元清貴. (2006). 節単位情報. 国立国語研究所報告書 124: 日本語話し言葉コーパスの構築法 (pp. 255–322).

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for organization of turn-taking for conversation. *Language*, *50*(4), 696–735.

吉田奈央・高梨克也・伝康晴. (2009). 対話におけるあいづち表現の認定とその問題点について. 言語処理学会第 15 回年次大会発表論文集 (pp. 430–433).

[*2] 例えば、独話では、主題要素が強境界の後方にも係り先を持つ場合、その強境界では発話を分割しない、という規則がある。