『日本語話し言葉コーパス』における話題導入表現の形態統語論的特徴と談話構造の分析

高梨 克也
科学技術振興機構さきがけ／京都大学学術情報メディアセンター
takanasi@ar.media.kyoto-u.ac.jp

1．はじめに

『日本語話し言葉コーパス』(CSJ) では，節単位認定に関する情報の一部として，「話題導入表現」タグが付与されている．しかし，認定基準によれば，これは統語的境界と談話構造上の境界とのミスマッチが生じているときにだけ便宜的に用いられるもので，話題導入としての談話機能を果たすすべての箇所が網羅的に認定されているわけではない．そこで，本稿では，CSJのうち，節単位と談話構造の両方が認定されている 40 講演を対象に，「話題導入表現」の形態統語論的特徴を解明するとともに，より機能的な観点から，「話題導入表現」と談話境界となる節単位との共通点を解明することを試みる．

2．談話の構造

談話とは1つ以上の文の集合であるが，これらの複数の文を「1 つの談話」としてまとめあげている原理の解明が重要である．これは，社会学的には，例えば，会話の各参与者の発話が当該の状況において行われている活動に指向していると表現される事態であり，また，認知科学的には，談話内の文の産出と理解にはトップダウンの制約が課せられていると表現される現象である[1]．

談話の構造（まとまり）が分かるためには，その境界または中心が特定できる必要がある．ただし，境界の表示と中心の予告が常に同一の文に含まれているとは限らない．また，談話構造の理解にはメタ談話的情報も重要な役割を果たすが（西條 1999），こうした情報は常に言語表現として明示されているとは限らない．さらに，まとまりをもつ「1 つの談話」はより小さなまとまりへと分割可能であるように見えることが多い．これを談話セグメントと呼ぶ．談話セグメントには，下位の談話目的がより上位の談話目的のための手段となっているという階層関係を仮定できるものの (Grosz&Sidner, 1986)，現実には，参与者にとっても分析者にとっても，こうした構造が常に明確に認定できるわけではない．

談話セグメントの境界や中心部分，階層関係を捉えることは自動要約（Mani, 2001）や会議ブラウジング（AMI）などの目的にとって重要な課題である．同時に，談話における発話単位の認定において，形態論的，韻律的特徴だけなく，機能的な観点を加えていく際にも，こうした談話構造の伝達に関わる特徴の同定が必要になる（伝他 2010）．そこで，本稿では，『日本語話し言葉コーパス』(CSJ) のアノテーション情報を利用することによって，談話セグメントの冒頭に置かれ，その内容を予告するものとしての「話題導入表現」のもつ特徴を分析する．

[1] 談話の「まとまり」を捉えるためには，これまでにも異なる立場からのさまざまな提案が行われてきたが，一致した見解が得られているとは言い難いのが現状である（石崎・高梨 2009）．

3．CSJ における2種類の談話情報

CSJ の大半の部分は学会講演や一般の協力者による体験談など（模擬講演）という独話からなる．このうち 199 講演には，話し言葉の文に相当する「節単位」（以下 CU）（高梨他 2004，丸山他 2006）や係り受け，重要文情報が，さらにそのうちの 40 講演には「談話セグメント」（以下 DS）（竹内他 2004）の情報が付与されている[2]．従って，この 40 講演については，発想の異なる次の2種類の談話情報が付与されていることになる[3]．

3.1 CU のコメント情報「話題導入表現」

CU は日本語述部の形態素列の規則性を利用した自動分割プログラムによってデフォルトの境界（文末などの[絶対境界]と並列節などの/強境界/）が与えられ（デフォルト境界にならない節末である<弱境界>も同時に認定される），このデフォルト境界の問題箇所を規則に従って適宜人手修正することによって認定されるが，この人手修正規則の一つとして「話題導入表現」がある[4]．

節境界の自動検出では，局所的な検出としては正しいものの，大域的な談話レベルの構造を考えた場合には望ましくないデフォルト境界が生じる場合がある．次の例では，2行目の「ワイ君はどういう人かと言うと」という表現はそれ以降の話題（ワイ君に関する説明）を導入する機能を果たしており，それに対して「凄く頭がよくて」「凄いおとなしいんですけど」「超面白い感じで」などが並列的に叙述される，という構造になっている．この場合，2と3が1つのCUとなったままだと，談話構造が不明確になるため，2の末尾が人手でCU境界に認定される．

1. 当時小学校六年生の時に一学期に隣りの席のワイ君のことが好きになったんですが/並列節ガ/
2. で<接続詞>ワイ君はどういう人かと<引用節>言うと<条件節ト>-；話題導入表現
3. 凄く頭が良くて/テ節/
4. で<接続詞>凄いおとなしいんですけど/並列節ケド/
5. 超面白い感じで/並列節デ/
6. で<接続詞>女子にも隠れて<テ節>人気があったけど/並列節ケド/
7. どっちかって<引用節>言うと<条件節ト>男子に人気があって<テ節>慕われてるという<トイウ節>感じで/並列節デ/...

[2] 談話構造情報を高い精度で大量のデータに付与するのは現状では極めて困難である．森本他(2004)には，こうした問題点についての報告がある（ただし，これは中間報告であるため，最終的な認定基準については竹内他(2004)の方を参照されたい）．

[3] CSJ の節単位と談話セグメントについては，坊農・高梨編(2009)のそれぞれ 2.1 節と 3.3 節にも解説がある．

[4] ボトムアップな観点から談話構造を認定していく際には，セグメント冒頭（近く）に生じる「話題導入表現」とセグメント後半ないし末尾に生じる評価表現やまとめ表現をセットにして扱うことが重要になる（高梨他 2003）．CSJ の CU のコメントの中にも，「直後がまとめ表現」があるが，紙面の都合で今回は扱えなかった．

言い換えれば，このように，形態統語論的な規則性に基づいて予測される境界と意味内容上の境界とのミスマッチが生じている箇所に対処するというのが「話題導入表現」認定の目的であるため，ここでは話題導入表現の談話機能が包括的に考慮されているわけではない．

3.2 DS のセグメント境界・談話目的

一方，談話構造認定では，談話セグメントの境界とこのセグメントの題名に相当する談話目的が付与される．従来，談話セグメントはいわゆる「談話主題（話題・キーワード）」の観点から認定されることが多かったのに対して，CSJ の談話構造認定の特徴は，単に談話主題 X だけでなく，「X の Y」という形式で表現可能な，説明タイプ Y を特定することが重要だと考えている点にある．ごく簡単にいえば，説明タイプ Y とは，話し手が X のどのような側面について，どのような観点から説明をしているかを表したものである．詳しくは 5.1 節で述べる．

3.3 トップダウン/ボトムアップの談話構造

以上のように，一方で，CU における「話題導入表現」は人手修正の結果として認定されることになる単位の末尾の形態論的特徴とその直後の部分との間の局所的な統語的・意味的関係から定義される「ボトムアップ」なもので，より大局的な談話構造を参照したものではない．他方，談話セグメントはより抽象的な「談話目的」という「トップダウン」の観点から認定されるものであり，作業時の入力データとして，発話単位としての CU とこれに付与された「話題導入表現」などのコメント情報が利用されているが，このコメント情報に従わなければならないわけではない．このように，CSJ における CU レベルと DS レベルでの 2 種類の談話情報は「泣き別れ」になったままの状態にある．そこで，以下では，CU 認定における「話題導入表現」と DS 認定におけるセグメント冒頭単位の形態論的，意味論的，語用論的特徴を分析する．

4．分析 1：節単位末尾の形態論的特徴

CSJ のうち，談話構造の付与された 40 講演（模擬 25，学会 15）を対象とする．これらの 40 講演に含まれる全 CU の合計は 4124 単位である．うち，「話題導入表現」コメントのある CU（以下「話題導入 CU」）が 121 で，全 CU の 2.9%である．一方，「談話目的」や「談話下位目的」の付与された談話境界になっている CU（以下「談話境界 CU」）は 630 で，全 CU の 15.3%にあたる．また，話題導入 CU と談話境界 CU の重複は 59 で，これは全話題導入 CU の 49%だが，全談話境界 CU の 9%にすぎない．このように，話題導入 CU は談話境界 CU になりやすいとは言えるものの，全 CU 数に占める頻度がそもそも多くないため，話題導入 CU だけを手がかりとして談話境界 CU を特定することは困難である．

4.1 談話境界 CU

全 CU を談話境界 CU と非境界 CU に二分し，各節単位末尾の形態的特徴を比較した．[表 1]には頻度の高いもののみを挙げる．%は各カテゴリの CU の全数（630，3494）のうちでの当該表現が含まれている CU の割合である．

[表 1] 談話境界 CU 末尾の形態論的特徴

	談話境界 CU		非境界 CU	
	n	%	n	%
[文末]	227	36.0%	1785	51.1%
/並列節ガ/	90	14.3%	225	6.4%
/テ節/	60	9.5%	246	7.0%
/並列節ケレドモ/	50	7.9%	144	4.1%
/並列節ケドモ/	28	4.4%	76	2.2%
/並列節ケド/	24	3.8%	157	4.5%
係助詞は	21	3.3%	13	0.4%
/条件節ト/	19	3.0%	49	1.4%

談話境界 CU では，[文末]の頻度がやや低い一方で，その分，/並列節ガ/，/並列節ケレドモ/，/並列節ケドモ/の頻度が若干多い．特に，/並列節ガ/は/並列節ケレドモ/や/並列節ケドモ/よりも，その差が顕著である．その他，全体に占める割合は低いものの，談話境界 CU では，係助詞「は」と/条件節ト/が多いようである．ただし，係助詞「は」を除けば，談話境界 CU を他の CU から明確に区別する末尾形態があるとは言い難い．

4.2 話題導入 CU

次に，デフォルト単位が人手修正によって切断されたことによって認定された CU のすべてを，話題導入 CU（121）とその他のもの（526）に二分し，末尾の形態を比較した．[表 2]には頻度の高いものや両者の差が顕著なもののみを挙げる．%は各カテゴリの CU の全数（121，526）のうちでの当該表現が含まれている CU の割合である．

[表 2] 話題導入 CU 末尾の形態論的特徴

	話題導入 CU		その他の人手修正 CU	
	n	%	n	%
+係助詞は	31	25.6%	3	0.6%
+<並列節デ>	28	23.1%	43	8.2%
+<条件節ト>	22	18.2%	4	0.8%
+<条件節タラ>	3	2.5%	1	0.2%
-<テ節>	12	9.9%	93	17.7%
-名詞	1	0.8%	79	15.0%
-<引用節>	0	0%	37	7.0%
-<理由節ノデ>	0	0%	27	5.1%

丸山他(2006)でも指摘されているように，話題導入 CU で顕著に多いのは，係助詞「は」，<並列節デ>，<条件節ト>，<条件節タラ>である（+ 印）．ただし，節単位途中も含め，これらの表現は定義上，デフォルト境界ではなく，従って CU の途中の位置にも生起しているはずであるため，これらの表現のうちのどの程度のものが話題導入表現になるかは分からない．逆に，話題導入 CU で少ないのは，<テ節>，名詞（その多くは「体言止」），<引用節>，<理由節ノデ>などである（- 印）．

5．分析 2：機能的成分の分析

5.1 機能的成分の認定

前節で見たように，話題導入 CU の末尾は定義上デフォルト境界ではなく，逆に，談話境界 CU の多くの末尾はデフォルト境界であるため，そもそも CU 末尾の形態だけを手掛かりとしたのでは，話題導入 CU と談話境界 CU に共通する性質は解明できない．そこで，今度は，話題導入 CU と談話境界 CU の両

者について，末尾形態だけでなく，その内部にどのような要素が含まれているかを機能的な観点から分析することにした[5].

高梨他(2004)では，「話題導入表現」としては，次のような「疑問詞＋発言動詞（メタ表現）」のパターンが典型的であることが指摘されている.

どういったことで始まったかって<引用節>言うと<P><条件節ト> - ;話題導入表現
それはどういうものかと<引用節>言うと<条件節ト> - ;話題導入表現

この指摘は，談話構造認定の立場と整合する．竹内他(2004)では，話し手が談話主題に相当するキーワードについて，聞き手が抱くだろう疑問を見越しながら説明を展開していくことによって，内容上一貫性のある節の連鎖パターンが形成されると考えられており，談話下位目的の認定作業においても，談話主題Xをどのような説明タイプYで説明しているかを示す<X,Y>の組を認定することが重要な方針となっている．説明タイプYは，Xに対する話し手の「評価」を，一貫性をもつ節連鎖として展開していく際の説明方法の類型であり，次のように分類されている（理論的背景については高梨他(2003)も参照）.

I. 中心的評価が相対的に明確なもの：利点・長所・欠点・問題点，程度・良さ・ひどさ・うれしさ，特徴・特色，解釈・意義，感想・印象・思い
II. 中心的評価が相対的に弱いもの：内容・状況・様子，種類・機能・形状・所属・効果
III. 出来事と時間軸の関係が特徴的なもの：経緯，帰結・結論・結果，変化，思い出・事件・経験
IV. 評価の中心が関係付けのもの：理由，きっかけ
V. 宣言的なもの：目標・目的，まとめ，分類
VI. 学会講演に特徴的なもの：定義・構成・対象・基準，図示・例示，手法・手順・方法，傾向・分布

そこで，こうした知見を踏まえつつ，話題導入や談話境界の機能にとって重要だと思われる次のような項目について，認定基準を作成した.

＊接続詞等
- **談話構造表示**：いわゆる談話標識の一部だが，当該部分が談話内で占める位置や他の部分との対比などを示す表現に限る（表現されている関係が不明確な「で」などは除外される）．その他，副詞的表現「まず」「次に」などや数詞を含む「一つ」「もう一つ」，「場合」「条件」「側」「方」なども含む.

＊形容語
- **評価語**：記述されている事態についての話し手の評価的態度の表現．多くは形容詞や副詞だが，話し手の評価や聞き手に評価的反応を引き起こす意図が明確な名詞（「別世界」「苦労」「野宿」など）や動詞（「自動で生成する」「落ちまくる」など），さらには「てしまう」などのモダリティ表現も含む[6].
- **強意語**：評価語によって表された評価を強めたり限定したりする表現で，多くは副詞.

＊名詞
- **側面語：**説明タイプのリスト（5.1 節）に対応．主題をフレームのような知識表現で表す場合にその属性となると考えられ，また「外の関係」の係り受けを形成しやすい名詞．形式名詞「の」「こと」を少し具体化した表現だと見なせる[7]．ただし，「場合」「条件」「側」「方」など，談話内での対比の表現に貢献するものは上記の「談話構造表示」に含めた.

＊機能表現
- **疑問表現：**疑問詞や Yes/No 疑問の場合の助詞「か」など.
- **と・/ていう：**直後に形式名詞「の」「こと」や側面語が続くことが多い.
- **の/こと＋は/が：**「○○のはXだ」のような疑似分裂文を形成するもの．「の」「こと」の部分がより具体的な名詞になると，当該名詞が「談話構造表示」や「側面語」の方に分類される.
- **がある/いる：**英語の There 構文に相当するもので，以降の主題となる要素を導入する.
- **の/ん＋です/だ＋が/けれども：**ただし，当該 CU が複数節からなる場合，挿入節や前半の背景的な節に含まれている場合は対象外．末尾形式の定義上，話題導入 CU では該当しない.
- **格助詞相当表現：**「として」「に関して」など，格助詞に相当する機能をもつ定型的表現で，「は」を伴うことも多い.

＊談話外行為：主に動詞（を含む述部）で表現される
- **試み：**当該の談話や述べられている出来事について，「今回の」という限定が当てはまるもの．「～すると」や「てみる」に言い換えると，直後にその帰結が述べられると予想されるようになる.
- **到達点：**そこまでで述べられてきた事態がある一定の状態に到達したことを表現し，直後にその次の事態が語られることが予期できるもの.

＊談話内行為
- **現場指示：**「これ」のようなスライドを指示する表現や「示す」のような指示に関連する行為を表す表現．基本的に学会講演でしか生起しない.
- **宣言・メタ：**「話し手がこの談話内のこの箇所で今行う」談話上の行為を明示化させるもの.

5.2 結果

以上の基準に基づいて，話題導入 CU と談話境界 CU へのアノテーションを行った結果を[表 3]に示す[8]．両カテゴリに含まれる 59 単位は重複して扱っている．学会講演と模擬講演とでは傾向が異なると考えられたため，区別した．%は各成分を含む CU が当該カテゴリの全 CU に占める割合である．「全体」で，5%以下の CU でしか生起しなかったものは除外した.

まず初めに言えるのは，それのみで話題導入/談話境界を特定できるような決定的な項目は見いだされなかったということである．話題導入/談話境界の認定においては，今後もこれらの複数の観点を同時に考慮していく必要がある.

[5] 「機能的」という用語は，下記の整理からも分かるように，これらの談話上重要な特徴と一般的な品詞論的カテゴリの間に緩やかな結びつきしかなく，ある機能が異なる品詞によって担われていることが多いことを考慮したものである.
[6] こうした評価表現の認定を話し手の意図や聞き手の理解から独立におこなうことには根本的な限界があると思われる（高梨他 2005）.
[7] 側面語は主題語（キーワード）よりは抽象的だが，「の」「こと」のような形式名詞よりは具体的だという意味で，記述レベルとメタレベルの中間に位置し，これらの両レベルを仲介するものであるとも考えられる.
[8] 各項目の認定基準の細部には明確化できていない直感的な点も含まれており，作業の再現性が保証できないため，今回の集計はあくまで試行的なものであると理解していただきたい.

[表 3] 話題導入 CU と談話境界 CU の機能的成分（*：「%」はパーセントではなく頻度）

			計*	接続詞等	形容語		名詞	機能表現						談話外行為		談話行為	
				談話構造表示	強意語	評価語	側面語	疑問表現	という	の/こと+は/が	が/も+ある/いる	んですが/けれども	格助詞相当表現	試み	到達点	現場指示	宣言メタ
話題導入CU (121)	学会 (47)	n	109	19	2	7	14	11	14	10	1	-	16	8	0	4	1
		%	2.3	40%	4%	15%	30%	23%	30%	21%	2%	-	34%	17%	0%	9%	2%
	模擬 (74)	n	165	20	16	33	17	13	25	18	6	-	9	4	0	-	1
		%	2.2	27%	22%	45%	23%	18%	34%	24%	8%	-	12%	5%	0%	-	1%
談話境界CU (630)	学会 (289)	n	553	139	7	31	98	30	28	13	7	53	36	47	6	43	29
		%	1.9	48%	2%	11%	34%	10%	10%	5%	2%	18%	13%	16%	2%	15%	10%
	模擬 (341)	n	679	137	52	155	36	12	56	61	30	74	11	12	39	-	8
		%	2.0	40%	15%	46%	11%	4%	16%	18%	9%	22%	3%	4%	11%	-	2%
全体(751)		n	1366	291	66	205	147	56	108	85	41	127	59	67	45	45	38
		%	2.0	42%	10%	30%	21%	8%	16%	12%	6%	18%	9%	10%	7%	7%	6%

次に，話題導入 CU と談話境界 CU に共通する項目だが，いずれについても「評価語」と「強意語」は模擬講演で多く生起しており，この点は理論的な予測通りである．ただ，学会講演では頻度が低く，課題が残る．逆に，話題導入 CU と談話境界 CU に共通して学会講演のみで多く見られたのは「試み」である．さらに，「側面語」も総じて多く見られたものの，模擬講演の談話境界 CU のみで頻度が低く，さらなる検討が必要である．

一方，「疑問表現」や「ていう」については，話題導入 CU では学会/模擬とも多く見られるものの，談話境界 CU では必ずしも頻度は高くないため，談話セグメントの冒頭文が聞き手の疑問を喚起する役割を果たすという理論的な予測は現時点では必ずしも裏づけられていない．

その他，談話の構造を明示するはずの「宣言・メタ」は，理論的な予測に反し，そもそも全般的な頻度が低かった．

5.3 構文論的パターン

「機能表現」については，各項目単独では必ずしも話題導入/談話境界の特徴になるとは言えないものの，これらの表現の連節によって，「どういう A かというと」「という B があるんですが」「なぜ B かということについてなんですが」といった特定の構文的パターンを形成していることも多い．こうした構文的パターンでは，A や B に「側面語」が入ることも多く，また C には評価語が入ることも可能で，さらに冒頭に「談話構造表示」が付加される場合もある．このように，今回採り上げた機能的成分は，個々の項目のみでは決定的な談話機能を持たないとしても，複数の項目の組み合わせを通じて，話題導入/談話境界としての重みを徐々に増大させてくことが可能なものであると考えられる．

6．今後の課題

今回の機能的成分の分析では，話題導入 CU や談話境界 CU でない CU にどの程度これらの成分が含まれているかという観点からの比較は行えなかったため，この点が緊急の課題となる．さらに，今回の対象データは独話のみであったため，今後はミーティングなどの会話データでどのような傾向が見られるかを引き続き検討していく必要がある．

謝辞

本研究は，JST 戦略的創造研究推進事業さきがけ「多人数インタラクション理解のための会話分析手法の開発」（研究代表者：高梨克也），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」（研究代表者：伝康晴）の一環として行われた．

参考文献

AMI: http://www.amiproject.org/

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

CSJ: http://www.ninjal.ac.jp/products-k/katsudo/seika/corpus/

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(2010)「対話研究にふさわしい発話単位の提案とその評価(2)～長い単位～」人工知能学会資料 SIG-SLUD-A903:13-18

Grosz, B.J. & Sidner,C.L. Attention, intention, and the structure of discourse. *Computational Linguistics*, 12 (3), 175-204.

石崎雅人・高梨克也(2009)「会話・対話におけるまとまりに関する一考察」人工知能学会資料 SIG-SLUD-A803:75-80

Mani, I. (2001) Automatic Summarization. John Benjamins. （奥村学他訳『自動要約』共立出版，2003）

丸山岳彦・高梨克也・内元清貴(2006)「節単位情報」『日本語話し言葉コーパスの構築法』国立国語研究所報告 124．255-322 http://www.ninjal.ac.jp/products-k/katsudo/seika/corpus/csj_report/05.pdf

森本郁代・竹内和広・高梨克也・井佐原均(2004)「二段階作業による『日本語話し言葉コーパス』の談話構造分析」『言語処理学会第 10 回年次大会発表論文集』389-392

西條美紀(1999)『談話におけるメタ言語の役割』風間書房

高梨克也・藤本英輝・河野恭之・竹内和広・井佐原均(2005)「会話連鎖を利用した態度情報と参与者間関係の特定方法」『言語処理学会第 11 回年次大会発表論文集』S6-4.

高梨克也・竹内和広・森本郁代・仲本康一郎・井佐原均(2003)「談話を語る／聞く動機とエピソード構造」『日本語用論学会第 6 回大会 Programs & Abstracts』76-79.

高梨克也・内元清貴・丸山岳彦(2004)「『日本語話し言葉コーパス』における節単位認定」（CSJ マニュアル）http://www.ninjal.ac.jp/products-k/katsudo/seika/corpus/public/manuals/clause.pdf

竹内和広・森本郁代・高梨克也・井佐原均(2004)「『日本語話し言葉コーパス』の談話境界情報について」（CSJ マニュアル）http://www.ninjal.ac.jp/products-k/katsudo/seika/corpus/public/manuals/discourse.pdf