話し言葉にみられる「で」の役割
－書き起こしデータからの定量的分析－

太田 公子，井佐原 均
独立行政法人 通信総合研究所
{kimiko,isahara}@crl.go.jp

1. はじめに

話し言葉と書き言葉，口語文と文語文の間では，接続詞の様相はかなり異なっており，さらに，同じ話し言葉でも，日常会話で現れる接続詞と，講演・講義の場面で用いられる接続詞とでは，相当の違いがみられるという [1]．それ故，接続詞としての機能や意味，用法の多様性が問題とされよう．ここでは，講演で用いられる接続詞（または接続表現）について議論するが，特に発話回数の多かった「で」における機能や役割について定量的に分析を行った．ここで扱う「で」は，接続助詞や連用中止形の「で」ではなく，複数の文の間をつなぐ接続詞の「で」である（たとえば，『飲んで騒ぐ』の「で」ではなく，『飲んだ．で，騒いだ』の「で」）．「で」というのは，文献 [1]（122頁）によれば，経時的に起こる事象を接続する接続詞で，「それで」という指示語系の接続表現から生じた表現とされている．前件に無理なくつながる事象を結びつけるという機能があり，一般的な表現である「それで」に対して，よりうちとけた会話的な表現であると位置付けされている．ここで用いた講演（[2]）が会話的であるとは考えにくいが，自発音声特有の様々な要因が「で」の頻数回の発話をもたらしたのかも知れない．講演の書き起こしデータ（話し言葉コーパス [2]）にみられる接続表現の一部を抽出した結果，「で」516回，「また」121回，「この」117回，「まず」91回，「そして」66回，「の」63回，「例えば」61回，「次に」52回であった．これは，200ミリ秒以上のポーズタグで挟まれた言葉を116講演分について抽出した結果である．ポーズタグで挟まれた言葉を選択した理由は，講演のような一方的な発話の場合は，講演中のポーズの代用として「それでは」「このように」「ただ」といった接続詞や指示詞，副詞などを用いているケースが目立ち，それらはポーズに挟まれて出現していたためである．116講演中最も多様された表現は「で」であり，ポーズを埋める働きとして用いられる接続表現であることがわかる．「で」の発話前後を観察したところ，前の部分では「～です」「～ます」「～であると」「～ですね」「～んですけど」といった終止的な表現，後ろ部分では「この」「その」などの指示語の他，「例えば」「それから」などの接続表現が用いられていた．これらの接続表現の使用には，個人差があり，116講演すべてにおいて「で」前後の発話が上述した形であるとは限らない．以上の予備調査を踏まえ，話し言葉にみられる「で」について具体的に調べていくことにする．

2. 「で」と他の接続表現

「で」の一般的な機能としては，前件に無理なくつながる事象を結び付けるものであることは既に述べた．しかし，これが意味する範囲は広い．文献 [3] によれば，「で」は前件の内容の同類や列挙を後に述べると言った添加型，さらに前件の内容から転じた別個の内容を後に述べると言った転換型の機能も有するとしている．それでは，話し言葉にみられる「で」はどのような型を有しているのか．他の接続表現と比較しながら述べる．

2-1. 分析手順

表 1に 116 講演中で発話回数の多かった接続表現とその回数を示す．これは，各講演の書き起こしデータに対し，形態素解析 [4] を施した結果から抽出したものである．「で」の回数が他と比べて多く，一講演平均約 30 回発話していることになる．もちろん，個人差はあり，一回も発話されない講演もあれば，122 回発話された講演もあった．

表 1: 116 講演中に頻出した接続表現とその回数

接続表現	回数	接続表現	回数
で	3404	そして	263
まず	513	更に	175
また	419	あるいは	168
例えば	271	つまり	117

次に，3404回の「で」のふるまいについて調べることにした．基本的な考え方は，「で」の後に出現する名詞，動詞，代名詞（使用する形態素解析結果に依存するが，

ここではこれら3つの品詞を扱った．以下これらをキーワードと呼ぶ）10個を抽出し，それらが「で」以前に出現した場合（出現しなかった場合は無視される），「で」以後のキーワードと「で」以前のそれとの形態素距離を求める，といったものである．接続詞は，文間の連接関係を明らかにする役割があり，後続の語句は先行の語句や文と関連をもっていることから，「で」の前後関係を調べる必要があると考えられる．また分析対象としてキーワードを選択したのは，たとえば，『こちらからは対数フレーム尤度を出力すると「で」ここで出力したこの対数フレーム尤度同士を比較して』の場合，「で」前後に共通するキーワードがあることから強い並列関係があるが，『Nグラムの制約を受けない状態になってます「で」具体的にどうしどうするかと言いますと』の場合は，共通するキーワードがないため，前の文を受けてはいるが，強い並列関係にないことが示唆されるためである．このように，キーワードは接続表現の役割を知る上で重要な手がかりとなり得る．分析手順は以下のようである：

① 書き起こしデータを整形する（ここでは200ミリ秒のポーズタグで改行した）．

② 形態素解析を行い，接続表現に相当する形態素を抽出する．

③ 「で」の後のキーワード W_{ij} と一致するキーワード W'_{ij} を発話開始から「で」前までの区間で検索し，W''_{ij} を求める（図1）．ここで，i は「で」の個数（$i=1,2,\ldots,3404$），j はある任意の「で」の後に出現するキーワードの順序（$j=1,2,\ldots,10$）である．たとえば，$W_{2,5}$ であれば，2個目に出現した「で」の第5番目のキーワードを示す．

④ ③の作業を幾つかのキーワード，$W_{i,2}$ や $W_{i,3}$ などに対しても同様に行う．（ここでは10個のキーワード，$W_{i,1}$ ～$W_{i,10}$ について行っている）．

⑤ 「で」から W_{ij} までの形態素数を形態素距離とし，$D_{b_{ij}}$ と表す．同様に，「で」から W''_{ij} までの形態素数を形態素距離とし，$D_{f_{ij}}$ と表す．

⑥ $n\times 20$ の行列を作成する（「で」の場合は n=3404）．

$$\begin{pmatrix} D_{b_{1,1}} & D_{f_{1,1}} & D_{b_{1,2}} & D_{f_{1,2}} & \cdots & D_{f_{1,10}} \\ D_{b_{2,1}} & D_{f_{2,1}} & D_{b_{2,2}} & D_{f_{2,2}} & \cdots & D_{f_{2,10}} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ D_{b_{n,1}} & D_{f_{n,1}} & D_{b_{n,2}} & D_{f_{n,2}} & \cdots & D_{f_{n,10}} \end{pmatrix}$$

⑦ この行列からユークリッド距離を求め，多次元尺度構成法（MDS）により布置する．

2-2. 分析結果

結果を図2に示す．同様の分析手順で行った「まず」「また」「例えば」「そして」「更に」「あるいは」「つまり」の結果も図3～図9に示す．横軸がI軸，縦軸がII軸である．布置された b と f は D_{bij}，D_{fij} の b，f を示し，数字は j を示す．「で」の分布によると，I軸方向は b と f の区別を，II軸方向は形態素距離の順を示している．I軸方向の正側に b が，負側に f がそれぞれ布置しているが，このように b と f が区別して布置された理由は，「で」以降のキーワード群が「で」の前でマッチするとき，「で」の前のある一部分に集中していることを示している．言い換えれば，「で」直後の話題は「で」の前のある部分の話題を受けていると考えられる．もし，受けている話題が，ある一部分に集中していなければ，図8の「あるいは」のように b と f が混在して分布することになる．また，II軸について，$b1$ がII軸方向で最も値が大きいことから，すべての i に対して「で」から最も近い W_{ij} が $j=1$ の場合であることがわかる．また，f についても同じである．次にII軸方向で値が大きいものは $b3$ である．これは i によって W_{ij} の位置が異なるため，$j=1$ の次の候補であるはずの $j=2$ と順序が入れ替わったと考えられる．しかし，$b*$ と $f*$ はI軸に対して対応関係にあり，$(b1, f1) > (b3, f3) > (b4, f4)$ …の関係があることから，「で」からの $D_{b_{ij}}$ 及び $D_{f_{ij}}$ が一定した間隔であることがわかる．これはすなわち，「で」の直後のキーワード（W_{i1}）は「で」の直前のキーワード（W''_{i1}）とマッチしており，「で」からの距離が離れれば，直前でマッチするキーワードの距離も「で」から離れることを示している．したがって，「で」前後の話題には共通性があり，特に「で」直前の話題を直後に再度話題にしていると考えられる．これは，「また」や「そして」にも共通して言えることであり，図4，図6の分布は「で」のそれと類似している．「例えば」（図5）に関しては，II軸に関して，上述したような関係が観察され

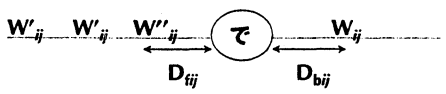

図1：「で」とキーワードとの関係図

W_{ij} はキーワード（名詞，動詞，代名詞），W'_{ij} は W_{ij} と一致したキーワード，W''_{ij} は幾つかの W'_{ij} のうち「で」に最も近いキーワードを示す．ここで i は「で」の個数（$i=1,2,\ldots,3404$），j は「で」の後に出現するキーワードの順序（$j=1,2,\ldots,10$）である．$D_{b_{ij}}$ は「で」から W_{ij} 区間の形態素距離と呼び，区間の形態素数に相当する．同様に $D_{f_{ij}}$ は「で」から W''_{ij} 間の形態素距離を示す．

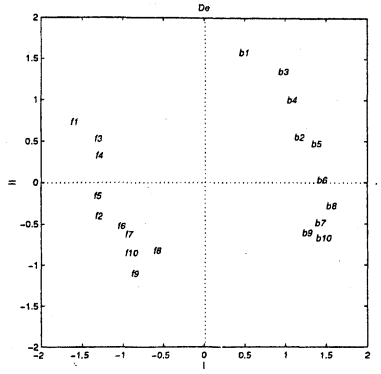

図 2:「で」の布置

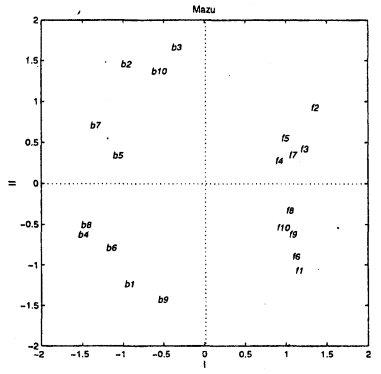

図 3:「まず」の布置

ない．II 軸方向で，j がランダムに点在していることから,「例えば」の直後のキーワードがその直前でマッチしている可能性が少なく,「例えば」の直前の話題は，その直後に話題に挙がるということは少ないことを示している.「例えば」と「つまり」(図 9) も布置は類似性がみられる．また,「あるいは」(図 8) は，上述した事柄を考慮すれば,「で」「また」「そして」の特徴と異なっていると判断できる.「また」(図 4) と「更に」(図 7) の布置は類似性がみられる．

2-3.「で」の役割

次に，布置の結果を詳細にみるための指標を提案することによって「で」と他の接続表現との関連性を明確にする．また，その指標を用いて各講演者が発話した「で」の役割について述べる．指標は以下のようにして求められる．

$$\Delta\theta_i = \cos^{-1}\frac{l_i^2 + l_{i+1}^2 - m_i^2}{2l_i l_{i+1}}, (i = 1, 2, \ldots, 9)$$

ここで $l_i^2 = x_i^2 + y_i^2$，$l_{i+1}^2 = x_{i+1}^2 + y_{i+1}^2$ であり，(x_i, y_i)

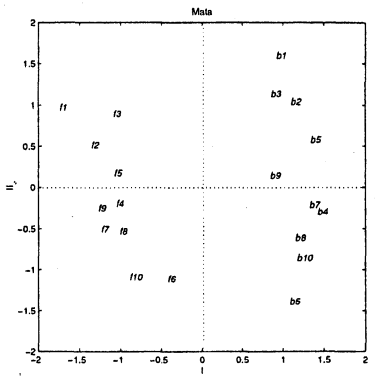

図 4:「また」の布置

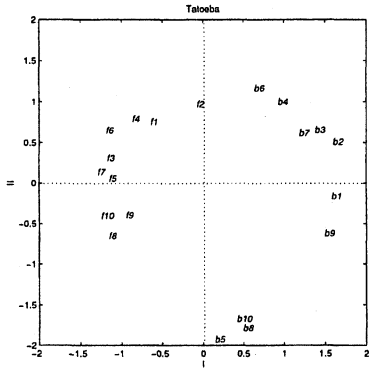

図 5:「例えば」の布置

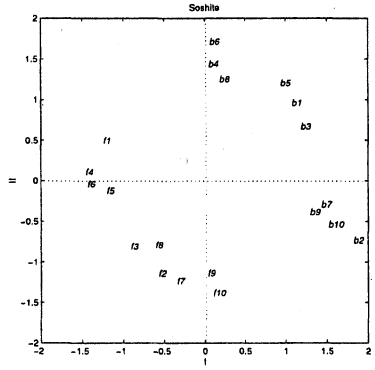

図 6:「そして」の布置

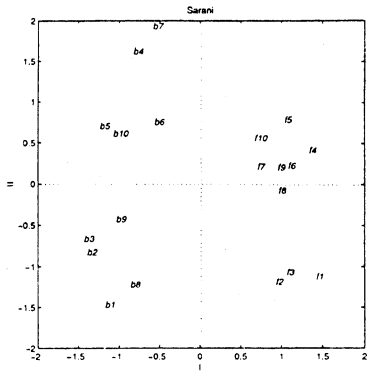

図 7: 「更に」の布置

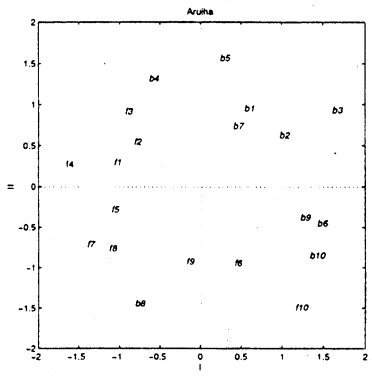

図 8: 「あるいは」の布置

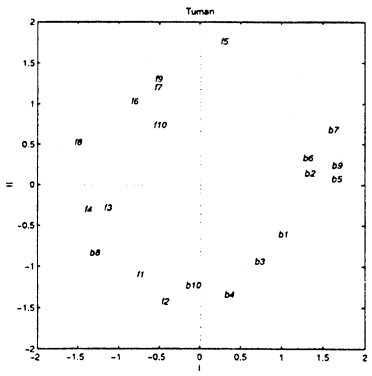

図 9: 「つまり」の布置

及び (x_{i+1}, y_{i+1}) は MDS によって得られた原点からの座標を指し，たとえば，$f1 = (x_1, y_1)$，$f2 = (x_2, y_2)$ となる．また，$\vec{m}_i = \vec{f}_i + \vec{f}_{i+1}$ であり，$m_i^2 = (x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2$ として算出される．このようにして算出された $\Delta\theta_i (i = 1, 2, \ldots, 9)$ は MDS によって布置された bi，fi の分散を表す．表 2に各接続表現の $\Delta\theta_i$ を示す（fi 側の値のみ）．値が大きければ話題の箇所が点在していることを示すが，値が小さければ接続表現前後の話題に共通点があることを示している．「で」は表に挙げた接続表現の中で最も小さい値を示した．次に，「また」「更に」となった．「そして」は「で」と同じような役割を示すと思われたが，ここで用いた話し言葉コーパスでは類似点がみられなかった．ここでは「そして」は話題を転換するための接続表現として用いられている場合が多く，「そして次に …」「そしてもう一点 …」のように発話されたケースが目立った．

表 2: 各接続表現の $\Delta\theta_i$

接続表現	$\Delta\theta_i$	接続表現	$\Delta\theta_i$
で	0.35	そして	1.15
まず	1.17	更に	0.78
また	0.67	あるいは	1.23
例えば	1.12	つまり	1.28

各講演者ごとに「で」の形態素距離行列から MDS を施し，$\Delta\theta_i$ を求めた結果，$\Delta\theta_i \geq 1.0$（話題転換型）は 16.7%，$\Delta\theta_i \leq 0.6$（話題並列型）は 36.7%，であった．また，「で」の発話回数と $\Delta\theta_i$ には相関はなかった．

3. おわりに

今回の分析結果から，以下に示すようなことがわかった．(1)「で」は話し言葉（講演スタイル）の中で主に話題並列の役割として使用されている．(2)「で」の使用に個人差はあるが，約 17%が話題転換の役割として，約 37%が話題並列の役割として使用している．残りの数十%はこれらの混合型である．(3) 他の接続表現にたとえて言うなら，「また」「更に」が「で」の機能や役割と類似している．

謝辞：通信総合研究所自然言語グループの方々並びに開放融合研究会の皆様に御指導頂いたことに感謝致します．

参考文献：[1] 鈴木一彦, 林巨樹, 研究資料日本文法第 4 巻, 明治書院 (1997). [2] 古井貞熙, 前川喜久雄, 井佐原均, “科学技術振興調整費開放的融合研究推進制度－大規模コーパスに基づく『話し言葉工学』の構築－”, 日本音響学会誌, 56(11) (2000). [3] 寺村秀夫他, ケーススタディ日本語の文章・談話, おうふう (2000). [4] 内元 清貴, 関根 聡 , 井佐原 均, “最大エントロピーモデルに基づく形態素解析—未知語の問題の解決策—”, 自然言語処理, 8(1), 127-141, (2001).