社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2005－NL－167（14）
2005－SLP－ 56（14）
2005／5／27

『日本語話し言葉コーパス』における自己修復部 (D タグ) の自動検出および修正に関する検討

下岡 和也[†]　**河原 達也**[†]　**内元 清貴**[‡]　**井佐原 均**[‡]

[†]**京都大学 情報学研究科**

〒 606-8501 京都市左京区吉田本町

[‡]**情報通信研究機構**

〒 619-0289 京都府相楽郡精華町光台 3-5

e-mail: shitaoka@ar.media.kyoto-u.ac.jp

あらまし　話し言葉においては，繰り返しや言い直しなどの自己修復部が数多く存在するが，書き起こしの整形過程においてこれらは削除・修正されるのが一般的である．『日本語話し言葉コーパス』(CSJ) においては，このような文節に対して D タグが付与されている．本研究ではまず，この D タグが付与されている自己修復部を自動検出する手法について検討する．具体的には，形態素や係り受けの情報を用いて機械学習を行い，実験的評価を示す．次に，このような自己修復部を，文整形・編集においてどのように処理すべきかについて検討する．当該文節に関する係り受け関係に基づいて場合分けを行うことで，削除すべき範囲を適切に同定できることを示す．

Automatic Detection and Correction of Self-Repairs in the Corpus of Spontaneous Japanese

Kazuya Shitaoka[†]　Tatsuya Kawahara[†]　Kiyotaka Uchimoto[‡]　Hitoshi Isahara[‡]

[†]School of Informatics, Kyoto University

Sakyo-ku, Kyoto 606-8501, Japan

[‡]National Institute of Information and Communications Technology

Hikaridai 3-5, Seika-cho, Soraku-gun, Kyoto 619-0289, Japan

e-mail: shitaoka@ar.media.kyoto-u.ac.jp

Abstract　In the transcripts of spontaneous speech, there are many self-repairs as well as fillers, and they are usually corrected by human editors. In the Corpus of Spontaneous Japanese (CSJ), a special tag (D-tag) is attached to the *bunsetsu* units of this kind of phenomenon. We present a method to detect such units based on a machine learning technique. Then, we investigate how to correct them, by classifying them based on the dependency structures.

1 はじめに

話し言葉における特徴の１つとして，フィラーや言い直しなどの言いよどみ (disfluency) の存在が挙げられる．この現象は，談話構造や話者のモデルを推定するのに重要であるとの指摘もあるが，従来の自然言語解析技術を適用する際には障害となる要因である．また，書き起こしを講演録や会議録の形で保存する場合にも，これらは削除・修正されるのが一般的である．フィラーに関しては，形態素情報や韻律的特徴から自動的な検出・削除が比較的容易であるのに対して，言い直しの検出や修正は困難であることから，これまであまり扱われていない．特に，大規模コーパスを用いて機械学習を行った研究事例はほとんどない．

本研究では,『日本語話し言葉コーパス』(CSJ) において，繰返しや言い直しなどの自己修復部に付与されている D タグを対象として，その自動検出及び修正を行うことを検討する．特に，形態素レベルの情報だけでなく，係り受けの情報に着目して，自己修復部の検出や修正対象の文節の同定の際に利用することを考える．

2 CSJ における自己修復部の係り受け構造

CSJ における文節間係り受けは，原則として『京都大学テキストコーパス』[1] の基準に準拠している．しかし，話し言葉では書き言葉に見られない現象が多く見られる．そこで，話し言葉特有の現象に対しては新たな基準を設けている [2]．ここでは，本研究の対象である言い直しや言い換えなどの自己修復部に対する係り受け構造付与について説明する．

CSJ では，言い直しや繰り返しなどの文節はその修復部に該当する文節に係ると定義されている．以下の例では,「山田」が「山田さん」に言い直されている．その際，文節「山田」はその修復部である文節「山田さんは」に係るものとし，文節「山田」には，自己修復部であることを意味するタグ「D」が付与される．

例) 山田 (D) ──┐
山田さんは──┐
言ってましたね

[1] http://www.kc.t.u-tokyo.ac.jp/nl-resource/corpus.html

3 自己修復部の処理に関する先行研究

自己修復部に関する最も代表的なモデルは RIM (Repair Interval Model)[3] である．RIM では，自己修復部を，被修復部 (ReParanDum Interval)，言いよどみ (DisFluency Interval)，修復部 (RePair Interval) の 3 つの区間に分割し，これらが必ず連続すると仮定する．これらを，“RPD，DF，RP” と表現すると，このモデルでは，まず，検出が比較的容易な DF の始端を決定し，DF の区間を求めて，その後，RPD と RP の区間を同定することで自己修復部を捕捉する．中断点の周辺には，パワーやピッチの急激な変化，あるいはポーズの存在といった音響的な特徴 [1] や,「えー」「あ」などのフィラー，あるいは「じゃなかった」「ごめん」などの手がかり表現といった言語的な特徴が見られるので，これらを検出に利用する．また，RP は RPD の繰り返しや修正表現であることから，パタンマッチングにより類似区間を検出することで，RPD や RP の区間を決定する．その後，RPD を RP で置換し，DF を削除して RP のみを残すことで自己修復部の修正を行う．しかし，このモデルでは RPD 内に必要な情報がある文に対応することができないので，新たな手法がいくつか提案されている．

船越らの手法 [4] では，コーパスの分析から，RPD 内の保持すべき語句は動詞に限られると仮定し，RIM では対応できない自己修復部を，“...RPD... 動詞 DF RP...” という形で捕捉するモデルが提案されている．ただし，対象としているのは対話コーパスであり，これに対して CSJ では，以下のように，上記の仮定にあてはまらない自己修復部も数多く観察されるので，適用は困難である．

例) [実空間] こういう みんなが 実際に [あのー] 顔を 合わせて集まるような [実空間]

藤井らの手法 [5] では，自己修復部を，文節を基本とする６単位に分割したモデルで捕捉し，修正する際には，各単位の削除に加えて適切な箇所への移動も考慮されている．この研究では本研究と同様に CSJ を対象としているが，扱う自己修復部は D タグではなく，自己修復部にフィラーや言いよどみの存在を仮定しているため，これらが存在しない以下のような自己修復部を捕捉できない．

例) 綺麗な 海だと 白い [砂] 純白の [砂で]

このように，D タグが付与されている自己修復部は多様であるので，本研究では，これを捕捉するための特定のモデル化を行うのではなく，まず，任意の文節に対して D タグが付与されるかどうか判定することを考える．

4 形態素・係り受け情報と機械学習を用いた自己修復部の検出

本研究では，任意の文節に対して D タグが付与されるか否かの判定を，形態素や係り受けの情報を用いて機械学習により行う．ここでは，SVM に基づく YamCha[6] を用いた．

自己修復部を検出する際には，RPD に該当する箇所と RP に該当する箇所の類似度が有用な情報となる．D タグが付与されている文節は RIM における RPD に該当するが，CSJ における係り受け付与においては，RP に該当する文節に係ると定義されているため，この情報を用いるには，係り先の文節内の単語との類似度を素性とする必要がある．

そこで本研究では，以下に述べる素性を考えて，これらの中から，最も精度がよくなるものを事後的に選択して用いた．なお，素性 (10) を用いない場合は F 値が大きく低下し，この有効性が顕著であった．

(1) 直後にポーズがあるかどうか
(2) フィラー/言いよどみが含まれるかどうか
(3) 文節内の形態素数
(4) 文節内の先頭/末尾の単語の品詞
(5) 文中での位置
(6) 直後の文節内の単語と形態素レベルで完全に一致する割合
(7) 直後の文節内の単語と部分一致する割合 (内容語に限定)
(8) 係り先の文節内の単語と形態素レベルで完全に一致する割合
(9) 係り先の文節内の単語と品詞レベルで一致する割合
(10) 係り先の文節内の単語と部分一致する割合 (内容語に限定)
(11) 係っている文節の個数
(12) 係り先との距離

また，YamCha における多項式カーネルの次数は 3，解析方向は Left-to-Right とした．

表 1: SVM の学習手法

	D タグ (正例)	D タグ以外 (負例)
手法 1	全て	全て
手法 2	部分一致・部分一致せず	全て
手法 3	部分一致	全て
	部分一致せず	全て
手法 4	部分一致	部分一致
	部分一致せず	部分一致せず

本研究では，D タグが付与されている文節を 1 つのクラスとして扱うだけでなく，表層的な情報で判別が可能なものとそうでないものの 2 つのクラスに分類することを考える．ここでは，係り先の文節内の単語と部分一致しているかどうかで，D タグが付与されている文節を以下の 2 クラスに分類する．

・表層的な情報で判別可能
　(例) そういう [風な] 風に 考えられるんじゃないかと
・表層的な情報で判別不可能
　(例) [ちょっと穴は] んー 溝は 作れないかもしれない

さらに，D タグが付与されていない文節についても，同様に，係り先の文節内の単語と部分一致しているかどうかで，2 つのクラスに分類することもあわせて考える．したがって，ここでは，D タグ検出のための分類器の構成として，表 1 に示すような 4 つの場合を考える．手法 2 以外は 2 クラス分類となり，手法 3 は 3 クラス分類となるが，YamCha における多値クラス識別手法として，Pairwise 法を用いた．

5 CSJ における自己修復部 (D タグ) の検出実験

CSJ のコアに含まれる 187 講演を用いて，D タグの検出実験を行った．20 講演をテストデータ，残りを学習データとして用いた．ここでは，CSJ に人手で付与されている係り受けのタグを用いた場合と，係り受け解析を自動で行った場合を比較した．自動で係り受け解析を行う際には，著者らが以前提案した手法 [7] を用いる．

各条件における実験結果を，それぞれ表 2，表 3 に示す．なお，手法 1〜4 は表 1 に示したものである．以降では，CSJ において D タグが付与されている文節の係り先を (RIM における) 修復部と呼ぶ．

表 2: D タグの検出精度 (人手による係り受けタグ)

	再現率	適合率	F 値
手法 1	50.3% (146/290)	69.2% (146/211)	58.3
手法 2	50.7% (147/290)	73.9% (147/199)	60.1
手法 3	50.7% (147/290)	75.4% (147/195)	60.6
手法 4	50.3% (146/290)	72.3% (146/202)	59.4

表 3: D タグの検出精度 (自動係り受け解析)

	再現率	適合率	F 値
手法 1	29.3% (85/290)	56.7% (85/150)	38.6
手法 2	25.9% (75/290)	57.7% (75/130)	35.7
手法 3	26.2% (76/290)	54.7% (76/139)	35.4
手法 4	27.2% (79/290)	47.9% (79/165)	34.7

まず，条件の違いに関して考察する．係り受け解析を自動で行った場合に大きく精度が下がっている．これは，4 節で述べたように，最も有効な素性が係り先の文節内の単語と部分一致する割合であるにもかかわらず，D タグが付与されるべき文節に対する解析精度が 45.7%(133/290) と低く，この情報が得られないためと考えられる．

次に，手法の違いに関して考察する．人手による係り受けタグを用いた場合は，D タグのクラス分類を行うことで適合率が上昇しているが，自動解析を行った場合は，そのような改善が見られない．これも，D タグが付与されるべき文節の係り先の同定率が低いことが原因と考えられる．

表 4〜表 7 に，各クラスについての個別の精度を示す．人手による係り受けタグを用いる場合 (表 4,5) には，D タグのクラス分類によって，表層的な情報で判別可能なものに対する適合率が大きく上昇している．一方，自動解析の場合 (表 6,7) には，D タグが付与されるべき文節の係り受け解析精度が低く，本来は表層的な情報で判別可能な文節 (184 個) のおよそ半数 (87 個) が，そうではない文節のクラスに分類され，正しい分類器が適用されていない．

表 4〜表 7 より，表層的な情報で判別可能な箇所については一定の精度で検出できていることがわかる．修復部の特定を含めた正しい係り受け情報が得

表 4: 手法 1(人手タグ) におけるクラス毎の検出精度

	再現率	適合率	F 値
表層的な情報で判別可能	70.7% (130/184)	75.6% (130/172)	73.0
表層的な情報では判別不可能	15.1% (16/106)	41.0% (16/39)	22.1

表 5: 手法 3(人手タグ) におけるクラス毎の検出精度

	再現率	適合率	F 値
表層的な情報で判別可能	72.3% (133/184)	82.6% (133/161)	77.1
表層的な情報では判別不可能	13.2% (14/106)	41.2% (14/34)	20.0

表 6: 手法 1(自動解析) におけるクラス毎の検出精度

	再現率	適合率	F 値
表層的な情報で判別可能	71.3% (69/97)	66.4% (69/104)	68.7
表層的な情報では判別不可能	8.3% (16/193)	34.8% (16/46)	13.4

表 7: 手法 3(自動解析) におけるクラス毎の検出精度

	再現率	適合率	F 値
表層的な情報で判別可能	66.0% (64/97)	71.9% (64/89)	68.8
表層的な情報では判別不可能	5.7% (11/193)	26.8% (11/41)	9.4

られている場合には F 値が 77.1 であり，これらを自動で解析した場合でも F 値 68.8 となった．その反面，表層的な情報では判別不可能な箇所についてはほとんど検出できていない．これは，用いている素性が表層的なものであることの限界と考えられる．したがって，D タグが付与されている文節と修復部との類似性を表現する別の素性が必要であると考えられる．例えば，これらの文節の文法的な働きや内容語の意味的素性が同等であると考えられることから，このような情報の利用が必要である．

また実際には，D タグが付与されるべき文節に関する係り受け精度が低いことにより，表層的な情報で判別可能な箇所も正しく処理できていない場合が多い．CSJ においては，これらの文節は修復部に係るとしているが，文法的には正しくないと考えられるので．修復部を同定するための係り受け解析を行う際に，文字列が部分一致する割合といった情報を用いる必要があると考えられる．

6 係り受け情報を用いた自己修復部の修正

次に，書き起こしの整形のための自己修復部の修正処理について検討する．自己修復部を修正する際には，単純にDタグが付与されている文節を削除すればよいわけではない．その文節に係る文節がある場合には，削除してよい範囲を適切に特定する必要がある．

本研究では，Dタグが付与されている文節に係ってくる文節があるかどうか，また修復部にその他の文節が係っているかどうかで，場合分けを行った上で，修正方法について検討する．以下では，それぞれの場合について述べる．

(1) Dタグの文節に係ってくる文節がない

この場合は，Dタグの文節のみを削除しても，文法的にも意味的にも問題は起こらない．この場合の例を表8に示す．最初の例では,「ペットボトルの」には何も係っていないため，これを削除する．テストデータ中の全Dタグ290箇所において，この場合に該当するのは133箇所であった．

(2) Dタグの文節に係る文節があるが，修復部に係る文節はない

これは，Dタグの文節のみを修復部で言い直している場合である．Dタグの文節と修復部は，文法的にも意味的にも同等の働きをしていると考えられ，Dタグの文節に係っている文節が修復部にも係っていると考えるのが妥当である．したがって，Dタグの文節のみを削除する．この場合の例を表9に示す．最初の例では,「そういう」は「風な」に係っているが，これを「風に」に係ると考えても問題ないため,「風な」のみを削除する．テストデータ中の全Dタグ290箇所において，この場合に該当するのは77箇所であった．

(3) Dタグの文節に係る文節があり，修復部にも他の文節が係っている

これは，Dタグの文節のみを言い直しているのではなく，その文節に係っている箇所も含めて言い直している場合である．この場合は，言い直している範囲を特定する必要がある．この場合の例を表10に示す．最初の例では,「誰それを」「病院」が「連れ」に係り,「病院に」が「連れてくみたいな」に係っている．ここでは,「連れ」に加えて「病院」も削除する必要があるが,「連れ」に係っている「誰それを」は残すべきである．また，4番目の例では,「話題」に係っている「その」や「有益な」は残して,「話題」のみを削除すべきである．

表8: Dタグの文節に係ってくる文節がない場合の例

1. すると 今 [ペットボトルの] 五百ミリリットルの ペットボトル
2. 私が もし [ああゆ] ああいう 風に なったら
3. ええー [それ] そこの 課題に 到達するまでに
4. あの [ここに] いす あの 高島平に 住むように なってから
5. 英語を あの [話した] あのー 書いたり 読んだり する
6. だから まー [演出] 一種の ディレクタールールですね

表9: 修復部に他の文節が係っていない場合の例

1. そういう [風な] 風に 考えられるんじゃないかと
2. 興味を [持っていまして] 持つように なりまして
3. 誤り傾向を [考慮した] 考慮する 為の 誤り訂正モデルを
4. 正と 負の [指令で] えー 指令を 持つものと いたします
5. これは あの その 部分の [ソナグラム] 波形 を 示して
6. ここで 検討した えー [内容について] 手法について

表10: 修復部に他の文節が係っている場合の例

1. 誰それを 病院 [連れ] 病院に 連れてくみたいな
2. 全然 インターラクションに [入ってきてない] あのー システムの 中に 入ってきてないんで
3. えー 二人で [アクセス] 二人なり 三人なり 四人なりで アクセスするんですね
4. えー その 有益な [話題] 最新の トピックスへの 到達の
5. ルールを 制御する 為の [ルール] もう 一段 上に ある ルールっていうのを
6. 初めて 会った 者同士の [間に] あ 電子的に 出会った 二人の 間に

この場合は，Dタグの文節および修復部に係っている文節集合から類似した文節のペアを抽出し，Dタグの文節に係っている方をDタグの文節とともに削除するといった処理が必要となる．この処理には，藤井らの手法[5]が適用できると考えられる．テストデータ中の全Dタグ290箇所において，この場合に該当するのは80箇所であった．

ただし，テストデータ中の全Dタグ290箇所において，文節単位の削除が困難な場合が11箇所あった．それらの例を表11に示す．1と2は,「〜について」「〜に対する」といった表現が機能的表現として扱われている場合であり，3と4は，述語が言い直されている場合である．これらの箇所については，上記の枠組みでは扱えない．

表 11: 文節単位での削除が困難な場合の例

1. 共通した [見学については] ついて
2. さまざまな あ [話者に対する] 対して え 高い
3. 傾向が 我が 社にも [ありました] ました
4. ウェブコーパスから [作成した] した モデルです

7 まとめ

本研究では，CSJ において，繰り返しや言い直しなどの自己修復箇所であることを示す D タグが付与されている文節を検出する手法を提案した．これは，D タグが付与されるか否かの判定を，形態素や係り受けの情報を用いて学習した SVM により行うものである．CSJ のコアを用いて評価を行った結果，表層的な情報で判別可能な箇所についてはある程度高い精度 (F 値で 7 割程度) を得ることができた．

また，D タグが付与されている文節に関する係り受け関係で場合分けを行うことで，実際に文編集を行う際に，削除すべき範囲を同定する方法に関して検討を行った．評価用データを用いて調査した結果，およそ 7 割の箇所について，削除すべき範囲を同定できることがわかった．

今後の課題としては，D タグ検出の際に，文字列の表層的な情報だけではなく，文節の文法的な働きや内容語の意味的素性についての情報を用いることや，係り受け解析を行う際に，文字列が部分一致する割合などの情報を用いることなどが挙げられる．

参考文献

[1] F.M.Quimbo, T.Kawahara, and S.Doshita. Prosodic analysis of fillers and self-repair in Japanese speech. In *Proc. ICSLP*, pp. 3313–3316, 1998.

[2] 内元清貴, 丸山岳彦, 高梨克也, 井佐原均. 『日本語話し言葉コーパス』における係り受け構造付与. 平成１５年度国立国語研究所公開研究発表会予稿集, 2003.

[3] C.Nakatani and J.Hirschberg. A speech-first model for repair identification and correction. In *Proc. ACL*, 1993.

[4] 船越考太郎, 徳永健伸, 田中穂積. 音声対話システムにおける日本語自己修復の処理. 自然言語処理学会誌, Vol.10, No.4, 2003.

[5] 藤井はつ音, 岡本紘幸, 斎藤博昭. 日本語話し言葉における自己修復の統計モデル. 言語処理学会第 10 回年次大会発表論文集, 2004.

[6] T.Kudo and Y.Matsumoto. Chunking with support vector machines. In *Proc. NA-ACL*, 2001.

[7] K.Shitaoka, K.Uchimoto, T.Kawahara, and H.Isahara. Dependency structure analysis and sentence boundary detection in spontaneous Japanese. In *Proc. COLING*, 2004.