自 然 言 語 処 理 100−10
（1994．3．18）

大規模コーパスに基づいた日本語自己修復文の分析

佐川雄二　大西昇　杉江昇
名古屋大学工学部

対話など話し言葉には，書き言葉にはあまり存在しない言い誤りをはじめとする不適格性が多く含まれるため，従来の文法を始めとする自然言語理解のための手法をそのまま適用することはほとんど不可能である．話し言葉に現れる不適格性の中で，最も頻度の高いものは話者自身による言い直しである自己修復 (self-repair) である．筆者らは以前，繰り返しと未知語を手がかりとして自己修復文を構文解析するシステムを提案した．本研究では，大規模コーパスを用いた分析により，機械的な方法で解析可能と思われる自己修復文がどの程度存在するかを検討する．その結果繰り返し，未知語および孤立語を手がかりとして約 85%の自己修復文が機械的に解析できる可能性が高いことが判明した．

Large Corpus-based Analysis of Japanese Self-Repaired Utterances

Yuji Sagawa　Noboru Ohnishi　Noboru Sugie
School of Engineering, Nagoya University
Furo-cho, Chikusa-ku, Nagoya 464-01,Japan

Self-repair is the most frequent phenomenon that makes computational treatment of spontaneous speech difficult. In this paper, we show the results of large corpus-based analysis of self-repaired Japanese utterances. The aim of our analysis is to find effective clues to eliminate self-repairs by computer. From the result, about 85% of self-repairs can be eliminated with repetitions, unknown words and isolated words as clues.

1 はじめに

対話理解，特に音声による対話の計算機による理解の問題は，マルチモーダルインターフェイスを実現する上での重要な問題である．対話理解の問題が，従来のキーボード入力を主体とした自然言語理解の問題と最も異なるのが，誤りを始めとする不適格性に対するロバストネスを確保しなければならないという点である．

キーボード入力においては，ユーザは発話に際して十分な時間をかけて準備が行なえる．しかし音声でリアルタイムに対話する場合はそうはいかない．長い沈黙は対話の効率を低下させるだけでなく，ユーザ自身の思考の流れも低下させるおそれがある．

我々の日常の会話を注意深く聞いていると実に多くの不適格性を見い出すことができる．しかしそれらの多くは話者はおろか，聞き手でさえ気づかない場合がある．また不適格性に気づいたとしても，何の問題もなく話者の意図した意味を推論できる場合が多い．このような，不適格性を感じながらも正しく意味を理解できるような発話を，「容認可能な不適格性を持つ発話」(permissibly-ill-formed utterance) と呼ぶこととする．

対話理解においてこのような容認可能な不適格性を持つ発話を理解する能力は重要である．なぜならその実現なくしては，ユーザの発話は大きく制限され，その制限を守ることはおそらくキーボードから入力するよりも注意深い発話の準備が必要だからである．例えば，本研究で対象とする自己修復は，Blackmer and Mitton[1] の報告では，実に平均 4.8 秒に一度の頻度で起こっている．このように瀕出する現象が扱えないのでは，音声入力にする意味がない．

本研究は，今述べたように，おそらく最も高い頻度で現れる不適格性だと思われる自己修復の計算機による理解を目指したものである．日本語において次に瀕出すると思われる容認可能な不適格性は，助詞の省略と倒置であるが，これらに関しては山本ら [2] が計算機による解析手法を検討している．

自己修復を含む文 (自己修復文) の計算機による解析手法については，まず Hindle[3] が提案したが，中断位置が前もって検出されているという前提を持つものだった．その前提の妥当性として Hindle は中断点に abrupt cutoff が存在している点をあげているが，十分な根拠はない．つづいて Langer[4] が自己修復文の解析手法を提案したが，そこでも中断位置が検出済みという前提が存在した．Langer の場合は，中断点に挿入される「えー」とか「あの」などの編集表現 (editing expression) をその検出の手がかりとしているが，編集表現は中断点に必ず存在するわけでもなく，自己修復以外にも言い淀み (hesitation) などでも現れるため，この前提は正しくない．

近年になって Shriberg[5] が自己修復解析のパターンマッチング法を提案した．この手法は我々の手法と類似しているが，彼らの分析に使われたコーパス [6] は，ボタンを押してから発話するなど，やや制限された対話である．また，Nakatani and Hirschberg[7] は，韻律情報を中心的に使用する speech-first 法を提案しているが，韻律イベントの抽出自体難しい問題であり，言語的な手がかりと相補的に使用する方が，韻律情報のみを使用するより確実である．

また上記の研究はすべて英語に関するものである (Langer は一部ドイツ語に関する考察を行なっている)．自己修復が言語に依存する側面を持つかどうかは不明であり，日本語に対して彼らの手法が有効かどうかはわからない．

これに対し，我々は日本語の自己修復文を構文解析するシステムをインプリメントした [8][9]．本システムは，未知語と繰り返し表現に基づいて中断位置とどの部分を修復部で置き換えるかを決定し，自己修復文をほとんど意味の変わらない適格文に変換して，構文解析を行なう．108 の自己修復文に対し，約 70%の解析率を持つことが示された．

本研究ではさらに多くの自己修復例を分析することにより，さらに解析率の高い自己修復文解析システムの構築可能性を探る．分析には，(株) エイ・ティ・アール自動翻訳電話研究所において作成された ADD 中の約 1,000 の自己修復文を用いた．筆者らの知る限り，対話理解の観点から行なわれたこのような規模の日本語自己修復文の分析は存在しない．

その結果繰り返し，未知語および孤立語を手がかりとして約 85%の自己修復文が解析できる可能性が高いことが判明した．また，以前提案したシステムで解析可能な自己修復も全体の約 64%と前回の評価実験とほぼ変わらない結果となった．

2 自己修復

自己修復 [10] は，話者自身による発話の言い直しである．例えば次の (例 1) がそうである．

(例 1) 普通の人よりも，この，一年間で普通の人よりも発育がいい

話者は，誤りや不適切な表現の検出後，発話を中断して，言い直しを開始する．この中断は次の (例 2) のよ

うに単語の途中でも許されるため，自己修復文を許す文法の記述は難しく，計算機による処理を難しくしている．

(例 2) この検診で，けつえ，あの，血尿があると言われた時には

なお，中断と修復の開始との間にはさまれることのある表現 (上の 2 つの例では「この」,「あの」) を，編集表現 (editing expression) と呼ぶ．

3 自己修復文の理解

自己修復文において，言い直された部分 (被修復部とよぶ．(例 1) では「普通の人よりも」，(例 2) では「けつえ」) は，文の意味を理解する上で必要のない部分である[1]．したがって自己修復文を理解する上で，解かなければならない問題は，どの部分が被修復部であるかを決定することである．それが正確に決定できれば，後は被修復部を削除し，その後に編集表現があればそれも削除すれば (例 1)，(例 2) はそれぞれ (例 1')，(例 2') のような適格文に変換できるので，これをシステムにかければ通常の文と同じように処理できる．

(例 1') 一年間で普通の人よりも発育がいい

(例 2') この検診で，血尿があると言われた時には

4 被修復部決定のための手がかり

被修復部を機械的に決定するための手がかりとして，本研究では繰り返し，未知語，孤立語の 3 つを考える．以下，それらの手がかりがどのようにして自己修復にともない出現するかという話し手側からの考察と，それらの手がかりがどのように被修復部を決定する際に使われるかという聞き手側からの考察を行なう．

4.1 繰り返し

自己修復にともなってよく観察される現象として，繰り返しがあげられる．

自己修復の多くは，中断直前の語句と同じ種類の語句を用いて行なわれることが知られている [10]．例えば (例 1) では，「普通の人よりも」という連用句が，

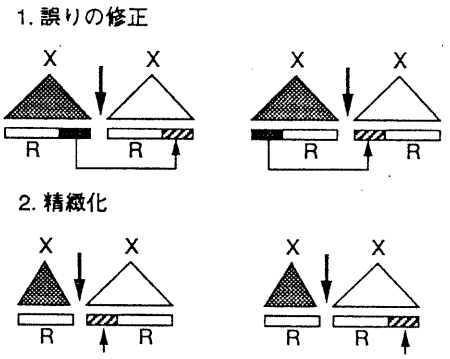

図 1: 自己修復にともなう繰り返し

「一年間で普通の人よりも」と同じ連用句で修復されている．

自己修復には，不適格な表現を訂正するために使われるものと，すでに言った語句に対して新たに情報を追加するもの (精緻化とよぶ) の 2 種類がある．各々の場合においてどのように繰り返しが現れるかを図 1 に示す．

訂正の場合，訂正する部分に前からかかる部分がある場合，その部分も修復部分に含めて言い直す場合がある (図左上)．この場合その部分が繰り返されることになる．前からかかる部分を含まない場合は繰り返しは現れない．必ずしも句 X を繰り返す必要はないからである．

一方，訂正する部分に後ろからかかる部分がある場合は，その部分はほとんどの場合修復部分に含まれる (図右上)．これは，中断直前の部分木の中で，訂正する部分を含む最小のものが句 X であるため，最低でもこの句を修復しなければならないためである．

一方，精緻化の場合，ある部分を前後どちらから修飾する語句を挿入するかで 2 つの場合がある．どちらの場合も新たに修飾部分が追加される語句が繰り返されるが，後ろに挿入する場合は必ずしも被修飾句を繰り返さない場合が存在するのに対し，前に挿入する場合は，ほとんどの場合被修飾句は繰り返される．これは，後ろに挿入する場合，句 X の修復という形でなく，中断直前の発話につながる発話として計画できるためである．

以上のようなプロセスで自己修復は生成されると思われるが，では，聞き手の側から，被修復部を見つけると言う観点から見た場合，繰り返しを含む自己修復

[1] 言い間違い方によって話者の心的状態が推測されたり，聞き手に何らかの感情を引き起こすことがあり，広い意味では聞き手は被修復部も含めて文を理解しているが，本研究では話者が伝えようとした意味を理解することを目的としている．

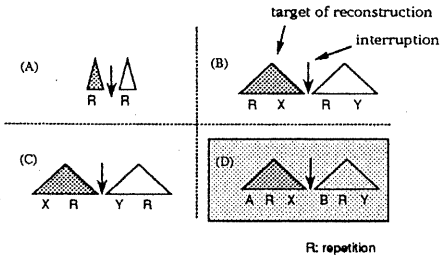

図 2: 可能な中断前後の構造
Possible structures around interruption

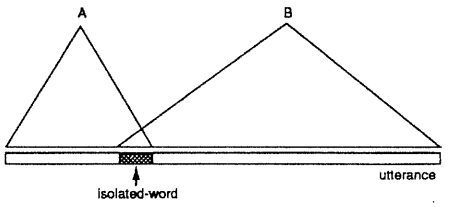

図 3: 孤立語の検出
Detection of an isolated word

はどのような特徴を持つのだろうか.

繰り返しを含む自己修復の中断前後の構造としては, 図2のような4種類が考えられる. A は単純な繰り返しなので, 簡単に被修復部を見つけることができる. 残りは繰り返しの間に語句が存在する場合で, この場合中断点がどこにあるかによって被修復部が変わってくる.

B, C については, 繰り返し部分の位置だけから, 中断位置および再構成対象を決定することができる. D についても繰り返しを含む連続する同種類の語句ということで決定できるが, 我々の予想ではこれは非常にレアなケースであると思われる. もしこの予想が正しければ, 繰り返しの存在する自己修復の機械的な理解は比較的容易である.

また, 繰り返しを伴う自己修復において, 話者はしばしば全く同一の語句ではなく, 意味的に同じ語句を使って繰り返すことがある. 例えば (例 3) がそうである.

(例 3) この先生には, えー, アメリカ, 米国における産業用車両の柔軟なガイダンス

こうした場合には, 繰り返し自体を見つけることに問題が残る可能性がある.

4.1.1 未知語

繰り返しを伴う自己修復は, 被修復部が完全に発話された後での修復であるが, その検出が完全に発話されない早い段階で行なわれた場合, 話者はしばしば単語の途中で中断し, 修復を始める.

この場合, 同じ単語を言い直す場合と, 別の単語を言い直す場合がある. 前者は, 図2の A と見ることもできるが, 本研究では, 単語内同一単語修復として別カテゴリとする. 次の (例 4) がそうである.

(例 4) てん, 展示担当の方とお話をしたいんですけれども

別の単語に言い直した場合, 途中まで発話された単語は, もともと発話しようとした単語と別の単語となって残ってしまう. 例えば, (例 2) では, 話者はおそらく最初「血液」と発話しようとして,「けつえ」まで言いかけたところで「血尿」と言い直したのだと思われる. その結果「けつえ」という文字列が残ってしまっている.

この場合のように, 単語内中断は未知語として残る場合がある. 未知語の検出は容易であるので, 未知語を含む自己修復では, その語句を被修復部として検出することができる.

4.1.2 孤立語

単語内中断で, 別の単語に言い直しかつ被修復部が未知語でなく, 字面上別の単語と区別のつかなくなった場合, その残された文字列が孤立語となるケースが多い. 孤立語 (isolated word) とはその単語が, その前後の語句のどちらとも構文的に接続できないものをいう.

つまり, 図3のような2つの構文木のいずれもができない場合である. このような部分解析を各語に関して行なうことは, 効率上の問題を引き起こすことが予想されるが, 我々の予想では, 比較的早い時点で失敗するものが多く, 実現可能と思われる.

5 分析

前章で述べた手がかりを用いた被修復部の検出の有効性を検証するため, 大規模データベースに基づく日

本語自己修復文の分析を行なった．分析は筆者ら自身が行なった．

5.1 コーパス

分析に用いたコーパスは，(株) エイ・ティ・アール自動翻訳電話研究所で作成されたADD[11]である．ADDの詳細に関しては，[11]を参照されたい．

ADDでは，言い淀み，言い直しが括弧でマークされているので，これを抜きだし，自己修復だけを取り出して分析した．

5.2 分析結果

分析結果を表1に示す．全体で1,082の自己修復が存在した．

5.3 その他のカテゴリ

ここでは，繰り返し，未知語，孤立語のいずれも存在しないものについていくつかのカテゴリわけを行なった結果を述べる．

5.3.1 活用誤りの語幹なし修復

動詞句や形容詞句などの活用の誤りの修復の場合，(例5)のように語幹を伴って修復する場合もあるが，(例6)のように伴わない場合も多い．

(例5) 発表時間が非常に短いこともあります，ありましてですね

(例6) 是非持って来て頂い，きたく思うんでございますが

この場合は，接続する語幹のない語尾部分の存在で，被修復部の決定は可能かと思われるが，本研究では対象外とする．

5.3.2 発話のやり直し

中断後，(例7)のようにまったく別の発話を開始している場合．

(例7) え，あの私，あ岡本さんでいらっしゃいますか，お忙しいところどうも申し訳ございません

この場合は，前の方から単語を徐々に削除しながら構文解析して，成功した時点からが，言い直した部分であると決定できる可能性があるが，本研究では対象外とする．

5.3.3 適格文へのすりかわり

全体として，別の意味の適格文として，解釈可能なもの．次の(例8)がそうである．

(例8) きょう，協賛する学会会員の

これは「今日，協賛する学会会員の」と解釈することも可能である．このような場合，韻律情報が誤った解釈を防いでいるはずで，それが実現できれば，このカテゴリのものは正しく解析できるはずである．しかし，実現は今後の課題である．

5.3.4 単語分離・接続修復

次の(例9)のように単語の途中から修復を開始している場合．

(例9) もう準じ備にですね，えー，かなり使っておりますんでね

この場合，被修復部を決めることができない．

5.3.5 異カテゴリ修復

(例10)のように，異なるカテゴリ，もしくは類似度が機械的に判定できるであろう範囲を下回る語句で修復している場合．

(例10) 一週間ずっと，付いておられる，行けると思うんでね

5.3.6 あいまいな修復

次の(例11)のように，修復のカテゴリが一意に決定できないもの．

(例11) 公式なアポイントを，に，え，数日中にとることができるかということと

この場合，「公式なアポイントを，2，3日中に,,,」と言いかけて数日中と言いかえたと思われるが，「を」を「に」に言い替えたとも考えられる．後者の場合間違った方に修復しているので，前者しかないと思われるが，どちらも正しいような場合もないとはいえず，処理の複雑さからいって，頻度が低い場合は対応しなくても良いと判断できる．

6 考察

分布の傾向は，以前の分析[9]とほぼ同じであり，以前の分析の妥当性が示された．今回の分析から前回提

案したシステムを評価すると，解析率は約64%であり，前回の評価とほぼ同等である．

繰り返し，未知語，孤立語を含む自己修復は，全体で85.8%であり，手がかりとしての有効性を示している．

繰り返しを含む自己修復で特徴的な結果は，やはり図2のDの構造を持つものが非常に少ない点である．これは聞き手の側にとっては，処理が簡単になると言う意味で利点であり，機械による自己修復理解の実現にも明るい結果である．

また，同一カテゴリを用いた繰り返しは，ほとんどが単純な繰り返し(図2, A)においてのみ現れている．隣接した同一カテゴリの語句は検出しやすいが，離れている場合，どのカテゴリで括るかで，可能な組合せが増大するので，このようなものが少ないことはやはり望ましい．

未知語については，それほど比率は高くならなかった．これは字面で見た場合，日本語には同音意義語が多いので，未知語ではなく，別の単語として見える場合が多いためだと思われる．ただしこのようなものの多くは，孤立語となっている．

孤立語の検出については，2,3語前語程度の部分解析の時点で失敗するものが多かった．これは処理の効率上望ましい．

3つの手がかりが存在しないもののうち，最も多いものは適格文へのすりかわりであった．しかしこの中には，(例8)のように字面ではわからないが，人間が聞いた場合自己修復とわかるものがほとんどで[2]，おそらく韻律情報などを使えばうまく処理できると思われる．

全体として，自己修復は比較的単純な構造のものが大部分を占めていることがわかる．これは自己修復は，時間的制約の下で行なわれる[12]ため，話者がなるべく単純な方法で修復しようとしているためと考えられる．逆にその結果，聞き手にとっても理解しやすくなって，これほど頻出するにも関わらず，さほどコミュニュケーションの障害となっていないのだと思われる．

7 むすび

本研究では，筆者らが以前提案した日本語自己修復文解析システムの有効性と，拡張の指針を得るべく行なった，大規模コーパスに基づいた日本語自己修復文の分析結果について述べた．

[2]実際分析したものはATRのマーキングで言い直しとされているものであるからすべてそうであるといえる．

その結果繰り返し，未知語および孤立語を手がかりとして約85%の自己修復文が機械的に解析できる可能性が高いことが判明した．

今後の課題としては，孤立語を検出する機能を実現して評価することと，韻律情報をいかに現在のシステムに取り込むかがあげられる．

謝辞

(株)エイ・ティ・アール自動翻訳電話研究所作成の対話データベースを使用致しました．作成に当たられた方々に感謝致します．

参考文献

[1] Blackmer, E. R. and Mitton, J. L.: Theories of Monitoring and the Timing of Repairs in Spontaneous Speech, *Cognition 39*, pp. 173–194 (1991).

[2] 山本幹雄, 小林聡, 中川聖一：音声対話文における助詞落ち・倒置の分析と解析手法, 情報処理学会論文誌, Vol. 33, No. 11, pp. 1322–1330 (1992).

[3] Hindle, D.: Deterministic Parsing of Syntactic Non-fluencies, *Proceedings of the 21st Annual Conference of the ACL*, pp. 123–128 (1983).

[4] Langer, H.: Syntactic Normalization of Spontaneous Speech, *COLING 90*, pp. 180–183 (1990).

[5] Shriberg, E., Bear, J. and Dowding, J.: Automatic Detection and Correction of Repairs in Human-Computer Dialog, *Proceedings of the DARPA Speech and Natural Language Workshop*, pp. 419–424 (1992).

[6] MADCOW, : Multi-site Data Collection for a Spoken Language Corpus, *Proceedings of the DARPA Speech and Natural Language Workshop*, pp. 7–14 (1992).

[7] Nakatani, K. and Hirschberg, J.: A Speech-First Model for Repair Detection and Correction, *Proceedings of the 31st Annual Meeting of ACL*, pp. 46–53 (1993).

[8] Sagawa, Y., Ohnishi, N. and Sugie, N.: Repairing Self-Repairs in Japanese, *Proceedings of*

Natural Language Processing Pacific Rim Symposium (NLPRS '93), pp. 191–198, Fukuoka (1993).

[9] 佐川雄二, 大西昇, 杉江昇：自己修復を含む日本語不適格文の分析とその計算機による理解手法に関する考察, 情報処理学会論文誌, Vol. 35, No. 1, pp. 46–52 (1994).

[10] Levelt, W. J. M.: *Speaking: From Intention to Articulation*, Chapter 12, pp. 458–499, The MIT Press, Cambridge, MA (1988).

[11] 江原, 井ノ上, 幸山, 長谷川, 庄山, 森：ATR 対話データベースの内容, Technical Report TR-I-0186, (株) エイ・ティ・アール自動翻訳電話研究所 (1990).

[12] Carletta, J., Caley, R. and Israd, S.: A System Architecture for Simulating Time-Constrained Language Production, Research Paper RP-43, Human Communication Research Centre, University of Edinburgh (1993).

表 1: 分析結果 (括弧内は全自己修復数に対する比率)

繰り返し	A 接続	同一単語 (列)	141(13.0%)
		同一カテゴリ	108(10.0%)
	B 接続	同一単語 (列)	96(8.9%)
		同一カテゴリ	2(0.2%)
	C 接続	同一単語 (列)	136(12.6%)
		同一カテゴリ	3(0.3%)
	D 接続	同一単語 (列)	4(0.4%)
		同一カテゴリ	0(0%)
単語内同一単語修復			105(9.7%)
未知語			98(9.1%)
孤立語			235(21.7%)
小計			928(85.8%)
その他	活用誤りの語幹なし修復		23(2.1%)
	発話のやり直し		6(0.6%)
	適格文へのすりかわり		111(10.3%)
	単語分離・接続修復		4(0.4%)
	異カテゴリ修復		5(0.5%)
	あいまいな修復		4(0.4%)
計			1082