自動抽出の対象としての日本語同語反復表現に関する考察

滝澤　修　　井佐原　均
{taki, isahara}@crl.go.jp

郵政省通信総合研究所

1．はじめに

日本語には、一文中に同じ語が繰り返され、そのような状態になっていることに起因して自然言語処理における何らかの特別扱い（別処理）が必要な表現がある。例えば「約束は約束だ」は、自分自身によって自分を定義するトートロジーとしての言外の意味をもつので、深い意味理解のためには別処理が不可欠であるし、あるいは「建物という建物が倒壊した」や「走るに走れない」のような慣用的な表現は、文字通りの自然言語解析では誤った結果が得られてしまうために、例えば機械翻訳などではこれらの表現用のパターンを用意しておいて処理する必要がある。このような表現のことを「同語反復表現」と呼ぶ（注1）。

同語反復表現は、「遅れれば遅れるほど」（反復語：動詞「遅れる」）と「高ければ高いほど」（反復語：形容詞「高い」）のように、同じパターン（この例の場合は「～ば～ほど」）における反復語をある程度自由に取り替え可能で、しかもどの反復語の場合でも言語処理上は同じ表現として扱うことが可能と思われる。従って同語反復表現は、「手を染める」のような一語性慣用表現（注2）や連鎖型共起表現（注3）よりは、反復語の取り替えが可能な点で、自由度が大きいといえる。一方、「走るに走れない」や「寝るに寝られない」などの同語反復表現は、「～に～ない」というパターンを持った離散型共起表現（注4）の一種とみなせる。しかしこのパターンは例えば「家に帰れない」にもマッチしてしまうが、これは対象とする同語反復表現ではない。つまり同語反復表現は、一般的な離散型共起表現よりは、同じ語が繰り返されているという制約がある分だけ、自由度が小さいといえる。

慣用表現（連鎖型共起表現）や離散型共起表現については、N-gram統計を用いて大規模コーパスから自動的に抽出する手法の研究が盛んに行われている[1][2]。しかし同語反復表現のように、連鎖型共起表現とみなすには自由度が大きく、かつ離散型共起表現とみなすには自由度が小さい表現は、従来の自然言語処理では、自動抽出の対象としてはあまり注目されていなかった。しかしながら同語反復表現は前述のように、自然言語処理上は別処理を施す必要があるものが多い。そのことから、同語反復表現を自動抽出することの重要性は高いといえる。

そこで筆者らは、同語反復表現を自動抽出する手法の検討を進めている[3][4][5]。自然言語処理において慣用表現として扱う範囲は、システムが持つ辞書や文法規則に依存する。それと同様に同語反復表現についても、自然言語処理に有益かどうかという観点から、自動抽出すべき対象の範囲を決める必要がある。そこで本稿では、自然言語処理の観点から、同語反復表現として分類すべき表現の範囲や種類を考察した結果について述べる。また、同じ語が繰り返されている文字列を新聞記事コーパスから自動抽出した結果を吟味し、抽出手法の課題についても検討する。

2．同語反復表現の範囲と種類

前節で述べたように、日本語同語反復表現の自動抽出の目的は、自然言語処理において特別扱いすべき表現を抽出し、処理を効率化・高性能化することにある。そのためには、自然言語処理に有益かどうかという観点から、自動抽出すべき対象を決める必要がある。

自動抽出すべき対象の決定のためには、大きく以下の2つの事項を決定する必要があるものと思われる。

（1）特別扱いが必要か不要かの決定
　　→自動抽出に必要
（2）特別扱いが必要な表現のタイプ分け
　　→自動抽出後の処理に必要

本稿では(1)を同語反復表現の「範囲」の決定、(2)を「種類」の決定と呼ぶことにする。それぞれについて以下の項で検討する。

2.1 同語反復表現の範囲

抽出機構の適用を想定している自然言語処理システムがもつ辞書情報や文法規則に依存するため、同語反復表現の範囲を明確に決めることは難しい。また一語性慣用表現などは、機械的に抽出し、出現頻度の多さを基準にある程度客観的にその範囲を決定可能であるが、同語反復表現はあまり多く

出現しないため、出現頻度を範囲決定の基準にしにくい。そのため、範囲は主観的に決めざるを得ない。本項では、同じ語が繰り返されているが同語反復表現とはみなしにくい（つまり自然言語処理上は別処理をする必要が無いと思われる）ものを除外していく方法を用いて、範囲について主観的に考察していくことにする。

●冗語およびオノマトペ的繰り返し表現

言い誤りや言い淀みによって同じ語が反復される場合は、冗語（pleonasm）として扱うべきである（注5）。次に、同じ語が連接して現れる「繰り返し繰り返し」のような重複表現も、強調のための冗語として扱うことができる。また「ますます」（益々）や「いやいや」（嫌々）のようなオノマトペ的な繰り返し表現は、自然言語処理において一形態素として辞書に登録しておくことで対応が可能である。以上の表現は同語反復表現とはみなさないとして差しつかえないものと思われる。

●連鎖型共起表現とみなすべき表現

「するようにした」「しようとする」や２重否定「ないことはない」などは、同じ語が繰り返されているものの、修辞的な意味をもたず、頻繁に出現する表現であるので、連鎖型共起表現として扱うのが適当であり、同語反復表現からは除外するべきと思われる。

●反復語の周囲の語の性質や、構文構造など

	同語反復表現か
「私は私、彼は彼だ」	○
「彼は彼の考えを持っている」	?
「彼は彼の妻に電話した」	×

図１　「～は～」の形式の表現と、同語反復表現との関係

図１に示すように、同じ「～は～」の形式をもつが後反復語（後に現れる反復語）に後続する語の性質や構文構造などの違いがある表現の場合、同語反復表現とみなせるものとそうでないものとがある。図１の「彼は彼の考えを持っている」の場合、「彼」が「彼の考え」を持っているのであるから、構造的には同語反復表現ではない。しかし「彼」が「彼の考え」を持っていることは自明であり、そのことにあえて言及していることは、Griceのいわゆる「会話の公理」に一見違反することであり、修辞的な効果を狙っていると考えられる。このような表現を同語反復表現に含めるべきかという問題がある。また「送るだけ送って下さい」は、動詞「送る」が繰り返されている同語反復表現とみなせるが、「送れるだけ送って下さい」のように、「送る」に可能の助動詞が付加した場合には、「送る」を反復語とする同語反復表現とは見なしにくくなると考えられる。しかし「美しいものは美しい」の場合、前反復語（先に現れる反復語）が反復語間語列（前反復語と後反復語との間にはさまれた語列）の「もの」を修飾する構造になっているものの、トートロジー的な意味を持つので、同語反復表現として扱うのが適当とも考えられる（注6）。同様に「落ちる場合は落ちる」も微妙である。このように同語反復表現の範囲の決定には、反復語間語列や、後反復語に後続する語の性質、さらには構文構造などを考慮する必要がある。

●反復語が指す対象

反復語として同じ単語を使っているが、反復語が指す対象が異なっているものも同語反復表現に含めるべきかという問題がある。例えば、以下の例文[6]における下線部が同語反復表現とみなされるかどうかを考えてみよう。

「世界はなおしばらくは国益と国益がぶつかり合いながら新秩序への模索が続いていくだろう」

この例の場合、前者の「国益」を持つ国と、後者の「国益」を持つ国とは、明らかに異なる。つまり両者は異なった「国益」を指している。従ってこの例は、自然言語処理（機械翻訳など）の観点から考えると、例えば「剣と盾がぶつかり合いながら」などの言い回しと同様に扱って差しつかえないものと思われる（注7）。従って、この例は同語反復表現とはみなすべきでないといえる。

但し厳密に考えてみると、どの同語反復表現でも前反復語と後反復語とが全く同じ対象を指していることはない。例えば同語反復表現と考えて差しつかえない「死を死として受け止める」の場合、前反復語の「死」は、「死」というインデックスを指し、後反復語は「死」のもつ属性を指している、と考えることができる。同様に「約束は約束だ」の場合も、前反復語の「約束」は、おそらく聞き手と発話者との間で以前に交わされたであろう特定の約束を指し、後反復語は「守るべきものという属性を持っている一般の約束」を指すものと解釈できる。このように、反復語が指す対象の異同を同語反復表現の範囲の決定に用いることは、慎重を要する。

●反復語の相違

反復語が異なっていても類似した意味を持つ場合に、トートロジー的な意味を持つことがある。例えば「男はオスだ」の場合、形式上は同語反復表現ではないが、同語反復表現である「男は男だ」と、意味処理上は類似した扱いが可能と思われる。

また「走りに走る」や「乗りに乗る」は、前反

復語の「走り」「乗り」は名詞で後反復語の「走る」「乗る」は動詞なので、同じ語が繰り返されているという同語反復表現の条件から外れている。しかし同表現は慣用的パターンをもち、直訳不能な点で、自然言語処理上は別扱いすべき、すなわち同語反復表現としての資格を備えている表現のように思われる。

これらの表現を同語反復表現として扱うかどうか、検討を要する。

2.2 同語反復表現の種類

同語反復表現の範囲の決定が困難である理由の一つが、「同語反復表現は○○の条件を満たす表現だ」という内包的定義が難しいことにあると思われる。そこで本項では、「『RはR』『RというR』…などが同語反復表現だ」という外延的定義をするために、同語反復表現の種類（グループ）について考察する。同語反復表現は、いくつものグループ化された表現の集合体と考えるべきであるが、そのグループは必ずしも排他的関係にはなく、さまざまな観点からのグループ分けが混在して構成されていると考えられる。

●トートロジー的同語反復表現と慣用的同語反復表現

1節で述べたように同語反復表現は、比喩と同様な「説明の省略」（つまり効率性）の機能をもつ「トートロジー的同語反復表現」と、定型表現的な「慣用的同語反復表現」の少なくとも2つに分けることができると考えられる。前者は深い意味解析などの対象としては重要な表現だが、例えば機械翻訳においては逐語訳が可能であるので、さほど重要ではないと思われる。一方後者は、文字通りの解析では誤った結果が得られてしまうために、機械翻訳などでも専用のパターンを用意しておいて処理する必要があり、重要な表現といえる。両タイプの境界を明確に決めることは難しい。例えば「歳が歳だから」は、文構造上はトートロジー的だが、慣用的な意味（例の場合「年をとっているから」という解釈が可能）をもつことから、慣用的同語反復表現とみなすことも可能である。

●同一の同語反復表現の場合に生じる分類の必要性

同じ反復語間語列かつ同じ反復語でも、異なった意味をもつと考えられる例もある。例えば「やることはやる」の場合、以下の2通りの異なった解釈が可能である。

(1) 「一応やってみる(が、多分できないだろう)」
(2) 「やらなければならないことは、やる義務があるのだ」

3．新聞記事コーパスに基づいた検討

前節での検討結果より、同語反復表現は少なくとも、「同じ語が繰り返され、反復語の間に別の形態素（助詞など）が1つ以上存在している表現」とすることができる。この条件に「反復語は動詞、形容詞、名詞」と、「反復語間の形態素数は5個以下」の2つの条件[4]を加え、朝日新聞社説2年分(1993, 1994)[7][6]から同語反復表現の候補を自動抽出した。

自動抽出の結果得られた、1993年分640文と1994年分689文を視察によって検討したところ、大部分が、同語反復表現とはみなせないという結果になった。そして結果をさらに詳しく検討し、以下の知見を得た。

●反復語間語列に読点が存在している場合は一般に並列表現であり、同語反復表現ではない。
→対策として、反復語間語列に読点が存在していないという条件を自動抽出に用いることが考えられる。しかし「彼は、やっぱり彼だ。」のように、明らかな同語反復表現でも反復語間語列に読点が存在することはありえるので、このような表現を洩らす問題を解決する必要がある。

●「する」「これ」「以内」などの普通名詞以外が反復語として抽出された場合は、同語反復表現ではない場合が多い。
→対策として、反復語が普通名詞という条件を自動抽出に用いることが考えられる。しかし指示代名詞による同語反復表現「それはそれで」などを洩らす問題を解決する必要がある。

●「しようとする」のような、同語反復表現とはみなさないとした連鎖型共起表現が多く抽出された。
→対策として、連鎖型共起表現の自動抽出をした後に、同語反復表現の自動抽出をするという処理の順序を設定することが考えられる。

4．まとめ

本稿では、同語反復表現として分類すべき表現の範囲や種類を考察した。また同じ語が繰り返されている文字列を新聞記事コーパスから自動抽出した結果を吟味し、手法の課題について検討した。

日本語同語反復表現の自動抽出の研究はまだあまりなされていない。その理由としては、少なくとも以下の2つが考えられる。

(1) 出現頻度があまり多くなく、しかも形式が多様であるため、自動抽出手法の一般化が図りにくく（図る必要性が少ない）、研究対象とする価値があまりない（ほかにすべきことが沢山ある）と考えられている。

(2) 同語反復表現は、自然言語処理上は共起表現あるいは慣用表現の一種とみなせば済むと考えられている。すなわち同語反復表現というカテゴリを設けること自体が妥当でなく、その範囲を決定することは無意味である。

これらの考え方に対して筆者らは、同語反復表現の自動抽出研究は従来たまたま見落とされていただけと考え、同研究の重要性を主張する立場をとっている。今後は本稿で指摘した課題の検討を進めるのと同時に、本研究の重要性についても更に客観的に検討する予定である。

【謝辞】新聞記事コーパスからの自動抽出実験に際しては、(株)リコーの簡易日本語解析系ＱＪＰ[8]の形態素解析機能を用いた（注8）。使用を許諾して下さった同社と、開発者である同社の亀田雅之氏に感謝致します。

（注1）佐山らは、「同語反復表現」を"tautology"の訳語として用いている[9]。しかしtautologyは、自分自身によって自分を定義する、論理における一形態のことであり、言語表現としてのトートロジー(rhetorical tautology)に限定される。そのため例えば「走りに走る」のようにトートロジーとはみなせない表現は同語反復表現ではないことになる。それに対して筆者らは、同語反復表現をもっと広く「同じ語が繰り返されている定型的な言語表現」ととらえ、トートロジーを同語反復表現の一部とみなしている。

（注2）一語性慣用表現[10]とは、慣用的な意味を持ち、他の語の挿入や語の交換をすると慣用的な意味を失ってしまうような定型表現である。「手を染める」などはその例である。この例の場合、他の語が挿入した「手を念入りに染める」や、語が交換された「足を染める」などは、慣用的な意味を失ってしまう。一語性慣用表現は、自然言語処理上は１語として扱うのが妥当とされる。

（注3）連鎖型共起表現[2]とは、ひとまとまりとして共起しやすい表現パターンのことである。「することになっている」などはその例である。一語性慣用表現は連鎖型共起表現の一部とみなすことができるが、連鎖型共起表現は慣用的な意味を持たなくてもよい。

（注4）離散型共起表現とは、離れた位置に共起する表現の組のことである[2]。例えば「その内容は～というもの」や「あやうく～するところだった」などである。この場合、"～"の部分に挿入できる語彙はある程度任意となる。

（注5）一般的な冗語の例としては、"a wrong mistake"や「過熱しすぎ」などがある。

（注6）「美しいものは美しいものだ」の場合、「美しいもの」という名詞句による同語（句）反復表現とみなせる。

（注7）文脈処理では、前者の「国益」を持つ国と、後者の「国益」を持つ国とをそれぞれ同定する必要があるが、それは同語反復表現の処理とは切り離して考えることができる。

（注8）形態素解析機能をもつ日本語解析器としてはJUMANがある[11]。本研究ではJUMANでなくQJPを用いた。その理由としては以下の２つが挙げられる。

- 同語反復表現の自動抽出は、自然言語処理における一作業にすぎない。そのため抽出機構はできるだけ軽いほうが良く、システムの軽さを特長とするQJPが適している。
- JUMANは複合名詞などを分割するのに対しQJPは分割しない。例えば「金融政策」をJUMANは「金融」「政策」の２語として扱うが、QJPは「金融政策」の１語として扱う。同語反復表現の自動抽出処理では、分割しないほうが好都合である場合が多い。

【参考文献】

[1] 新納,井佐原:「疑似Nグラムを用いた助詞的定型表現の自動抽出」,情処論,Vol.36,No.1,pp.32-40,1995.
[2] 池原,白井,川岡:「大規模日本語コーパスからの連鎖型および離散型共起表現の自動抽出法」,信学技報,NLC95-3,1995.
[3] 滝澤,井佐原:「計算機によるトートロジーの意味理解」, 言処年会,pp.161-164,1995.
[4] 滝澤,井佐原:「品詞の並びに関するヒューリスティックスを用いた日本語同語反復表現の検出」,情処研報,95-NL-110-3,1995.
[5] 滝澤,井佐原:「日本語同語反復表現の検出手法について」,言処年会,pp.305-308,1996.
[6] 朝日新聞記事1994年版CD-ROM, 日外アソシエーツ.
[7] 朝日新聞記事1993年版CD-ROM, 日外アソシエーツ.
[8] 亀田:「軽量・高速な日本語解析ツール『簡易日本語解析系Q_JP』」,言処年会,pp.349-352,1995.
[9] 佐山,阿部:「日本語同語反復表現の意味解釈」,心理學研究,Vol.65,No.1,pp.25-33,1994.
[10] 首藤,吉村,武内,津田:「日本語の慣用表現について」, 情処研報,88-NL-66-1,1988.
[11] 松本,黒橋,山地,妙木,長尾:「日本語形態素解析システムJUMAN Ver.3.11 使用説明書」,1996.