ロシア語の報道文体におけるパラフレーズの問題
―「大統領」「首相」を表わす語句に着目した分析―[1]

世 利 彰 規

0. 導入

報道のジャンルのテクストで，しばしばある特定の対象を意味するのに多様な表現が用いられる。本稿ではこの言いかえを「パラフレーズ」と呼ぶ。本稿で「情報伝達を第一とする報道テクストでパラフレーズは，どのような条件下で，どれほど頻繁に現われるのか」という問題をあつかう。

1. 立場と目的

本稿における言語研究は，人文科学としての言語研究に属する。[2]

ロシア語をはじめとするヨーロッパの言語で書かれた新聞などの報道文を読むときに気がつくことがある。それは，同じ対象に二度以上言及する場合，できるだけ同じ言い回しを用いることを避け，ちがった言い回しを用いる傾向があることである。[3] この言いかえは外国語としてロシア語を学ぶ者が報道テクストの読解に取り組む際に大きな障害となる。この言いかえについての傾向や仕組みについて，客観的・定量的な観点から取り組むことは意義のあることだと考えられる。

しかし，実際の報道テクストにおいて言い換えられる対象は多く存在する。その中でもとりわけ多く言及されるのが，政治の話題についてのテクスト中でロシア大統領や首相の動向や活動である。これらの対象は頻繁に言及され，様々な表現によって言いかえられる。そこで本稿において，ロシア語の報道文における限定された人物（「大統領」「首相」）に

---

[1] 本稿は 2013 年度言語処理学会（於名古屋大学）での発表「ロシア語の報道文体におけるパラフレーズのメカニズム」の会議レポートに訂正を加え，加筆したものである。

[2] 近代からはじまる言語研究は，それが人文科学の分野であっても，科学的であるという立場に立っている。科学的であるためには客観的な数値による議論が必要である。技術的な進歩によって，電子化した生の言語データを WWW から得ることは比較的容易となっている。しかし，数値によって結論を出す場合，基礎的であっても統計学的な手法が要求される。本研究が人文科学の立場にありながらも，統計による数値を用いた分析を採用した理由もそこにある。このことについては日本語学からの指摘がある（水谷静夫『曲り角の日本語』岩波新書，2011 年，156 頁）。

[3] このことについては文体論の立場から指摘がある（外山滋比古『外山滋比古著作集 8　風の音』みすず書房，2002 年，135 頁）。

対する言いかえ（「パラフレーズ」）がどのような頻度で現れるのか分析する。また，パラフレーズはテクスト内のどの要素と関係して生じるのかについて考察する。

本稿の構成は次のようになる。まず，ロシア語の報道文において本当に全く同じ語は繰り返されない傾向にあるのか，という問題提起が正しいものなのかどうか統計的手法を用いて調査する。そのような傾向があることを統計的に示したのちに，テクストの大きさ，間に置かれる文の数や語数などの間隔の要素がパラフレーズの出現と相関を持っていることを統計によって実証する。

2. 先行研究

この「同一対象を異なる表現で言いかえる」という現象は，テクスト言語学という分野での研究がなされている。

パラフレーズに関連する問題を初めて取り扱ったのは Harweg 1979 である。[4] 彼はパレフレーズを言語学上の基本単位である「統合 paradigmatic」と「範列 syntagmatic」の観点から分類を行った。Harweg によると，統語的な，順序に関係する側面に着目した場合は「範列的代入 syntagmatic substitution」と呼ばれる。一方，交替可能な表現そのもの，表現のヴァリエーションについて言及する場合，「統合的代入 paradigmatic substitution」と称される。

ロシア語におけるパラフレーズの範列関係に着目して論じた研究としては，Бытева 2008 がある。[5]

このように，パラフレーズされる表現の統合的性質，すなわち言いかえにはどのような表現のヴァリエーションが見られるかについて文法・文体論の領域での研究がなされてきた。しかし，あるテクストの中でどのようなパラフレーズ表現がどの順序で現れるか，どの頻度で用いられるか，などについて考察したものは少ないように思われる。

3. 問題提起と作業仮説

3.1. 問題提起

本稿が扱う問題を提起する。ヨーロッパの言語においてニュースや新聞といった報道に関わるテクストは文体論的に「文語的」とみなされ，それらの内にはある特徴を見ることができる。その特徴とは，同じものを表現するときに，なるべく同じ言葉を繰り返さない，同じことを言うにも違った表現を探し出して使う，というものである。これらは必然的な

[4] von Roland Harweg, *Pronomina und Textkonstitution* (München: Fink, 1979).

[5] *Бытева Т.И.* Очерки по русской перифрастике. М., 2008.

規則ではなく，あくまでも傾向とされる。

このような表現の交替は文学的なテクストにおいては装飾的な役割を果たすため問題はない。しかし，報道に関わるテクストは情報の伝達を第一の目的とする。その場合，同じことを言うのにちがった表現を用いるというパラフレーズは情報伝達の妨げとなる。

例）

*При этом **Президент** отметил, что РФ выступает за такой вариант решения проблемы, который избавил бы Сирию от непрекращающейся гражданской войны. "Мы не хотим, чтобы оппозиция, придя к власти, начала борьбу с властью, которая перейдет в оппозицию", — сказал **глава государства**. Речь идет о том, чтобы "сначала договорились, как они будут жить дальше, а потом менять существующий порядок вещей, а не наоборот - сначала все разогнать и уничтожить, а потом решать", — пояснил **Владимир Путин**.*

(http://www.1tv.ru/news/polit/222478)

これに際して，大統領が語ったところによると，ロシアは，とどまることを知らない市民戦争からシリアを救うため，問題解決のこのような方法に賛成する。「私たちは，反政府軍が，権力を握り，手に渡った権力をもって戦争を始めたりしないよう望む」と国家元首は語った。重要なのは，「まずこれからどうやって彼らが生きていくつもりなのかを話し合うこと，次に，事態に関する現体制を変えることで，逆にすべてを追い払い，破壊してから決めるのではない」とプーチン氏は語った。

また，パラフレーズとは矛盾する言語学上の概念が存在する。それは「言語の経済性」と呼ばれる。その概念は次のように説明される。人間は言語によって通信を行う場合，活動を最小限にしようとする傾向がある。そしてその傾向は言語の体系にも影響する。ある言語の語彙に完全に意味の重なる同義語は存在しない。なぜならそのような同義語は人間の記憶の負担となるためである。

したがって，一見同じ意味を持つように見える語句も，異なる意味を含んでいたり，用法の面において相違点が存在していたりする。

「言語の経済性」はある言語が内包する語彙や表現に関する概念であった。しかし，これはテクストのレベルでも当てはまる。同じ物事を言い表すのに，いちいち違う言葉を使うとする。その場合，そのたびに類推や想起といった作業を行わなければならず，記憶領域に無駄な負担がかかる。これは情報の伝達を妨げ，「言語の経済性」に矛盾すると言える。

しかし，「言語の経済性」の概念を提起したアンドレ・マルチネは次のようにも述べている：

> 文字どおりにとると，こういう断定はつぎのような意味を含みそうだ：通信にははっきりした寄与をもたらさないものは，言語のなかになにも存続できないだろうし，また言表の要素はそれぞれどういう機能をみたすかに厳密に比例する生産努力を要求する，と。じつは，このことはすべて，現実条件を抜きにした議論では正当化されるにしても，通信活動のおこなわれる事情と相容れるものではない。〈…〉通信のじっさいのうえでの必要から，それゆえ，言語形は，いつでもまたどんな面でも大幅に冗長 redondant であることが要求されるのである。[6]

マルチネによると，「言語の経済性」は何より優先される絶対的な規則ではなく，対立する「冗長性」と共存するものであると考えることもできそうである。これはこの問題が文法に関わるはっきりしたものというよりも文体的なゆらぎやすいものであることを意味している。以上の論点からつぎの問題を提起することができる。

問題提起：①本当に報道テクストの中で同じ言い回しの繰り返しは避けられる傾向にあるのか。[7] そのような傾向が存在する場合，②それはテクストのどのような性質と関係するのか？

3.2. 作業仮説

報道ジャンルのテクスト内で同じ表現の繰り返しが避けられる場合，次のような仮説を立てることができる。単調さを避けるためにパラフレーズが用いられるとしたら，一度用いられた表現の印象が残らないよう二つの表現と表現との間に一定の間隔が置かれるはずである。つまり，たとえ同じ対象を指示するのに二度同じ表現を繰り返したとしても，繰り返された二つの表現の間に十分な間隔が置かれていれば，単調な印象が持たれることは少ないということになる。ここでの「間隔」として，表現と表現の間に置かれる語数，文の数を挙げることができる。これは直観的に当然と思われることであるが，実際に「間隔」と言いかえの関連性を客観的・定量的に分析した研究の例はない。直観的に考えられることに対し，統計による客観的な結果を示すことも意義を持つと考えて許されるだろう。

ここから，問題提起に対して，次のような作業仮説を立てることができる：

「報道文体においてパラフレーズされるか否かは，表現と表現との間隔の大きさが関係する」

---

[6] アンドレ・マルチネ（三宅徳嘉訳）『一般言語学要理』岩波書店，1972 年，250-251 頁。

[7] 今回は，分析対象を報道テクストに限定した。従来のロシア語研究の領域では文学テクストを分析したり，文学作品から用例をとったりすることが多かった。しかし，文学における言語は，情報伝達と冗長性以外の要因の影響も考えられる。本稿での対象は情報伝達の問題とするため，文学テクスト以外のジャンルである報道テクストの分析を行った。

表現と表現の間に置かれる間隔として考えられるのは 1）語数，2）文の数の 2 つが考えられる。同じ言い回しが使われる場合，間に置かれる語数，同じ対象の指示の回数について調べる必要がある。さらに「パラフレーズの出現が，語数，文の数と相関を持つかどうか」についても考察する。

上の仮説の実証のために下記のような作業を行う：「大統領」「首相」を意味する語句について，パラフレーズされる頻度および同一表現の繰り返される頻度と，テクスト内で「大統領」「首相」の間に置かれる語数や文の数という間隔とが相関関係にあるかどうかを調べる。

4. パラフレーズの分類

4.1. どの表現を違うものと見なすのか

分析の際，もっとも難しい問題となるのが，どの表現を同じとみなし，どれをパラフレーズされた違ったものとみなすかという指標を決定することである。あまりに大まかな指標を設定すると，分析の制度が損なわれる。一方で，あまりにも表現を細かく分類してしまうと，パラフレーズの頻度が不自然に大きくなってしまう。そこで，ここでは表現を構成する名詞を中心に考え，定語がついているかどうかを判断基準の一つとする。глава のついた表現，президент や лидер, премьер など外来語由来の名詞のついた表現，個人名のついた表現の 3 つに分類し，それらに一致・不一致定語がついているかいないか，父称を用いているかいないかで「長い形」短い形」の 2 通りに分けた。また российский президент Владимир Путин, премьер-министр のような付語も定語とみなす。

首相／大統領	政府首班／国家元首	個人名
президент (Росии)	глава государства	Дмитрий Медведев
российский президент	глава правительства	Владимир Путин
российский лидер	глава российского государства	Медведев
премьер-министр (России)	глава российйкого правительства	Путин
российский премьер-министр		
российский премьер		

上の表のように，глава のついた役職名とそうでない役職名，および個人名称とをそれぞれリストとしたテクストから検索し，計算する。曲用変化も正規表現を用いることで正しく検索することができる。

4.2. パラフレーズの判定方法

二回以上同じ表現での指示の場合，同一表現の繰り返しとする。それ以外の場合をパラフレーズとする。

同一表現の繰り返しの例

‹...› *Направленные на это договорённости достигнуты по итогам переговоров, которые сегодня в Дели провёл* ***Владимир Путин****. В программу официального визита пришлось вносить некоторые изменения из-за обстановки в индийской столице.*

Визит ***Владимира Путина*** *в Дели проходил на фоне масштабных волнений, которые же вторую неделю продолжаются в Дели.* (http://www.1tv.ru/news/print/222753)

〈…〉この方針へと向けられた決定は，**プーチン氏**によってデリーで行われた交渉の結果なされた。公式訪問の予定はインドのいくつかの都市での滞在のため，ある程度の変更がなされることとなった。

プーチン氏のデリー訪問は，既に二週もの間デリーにおいて続いている大規模な混乱を背景に行われた。

パラフレーズの例

Во время пресс-конференции ***Владимиру Путину*** *задавали вопросы и об отношениях с другими странами, в том числе с соседями России.* ***Глава государства****, в частности, заявил, что видит позитивные сигналы от новых грузинских властей, направленные на урегулирование отношений.*

(http://www.1tv.ru/news/economic/222494)

記者会見の際に，**プーチン氏**はロシアの隣国を含めた他国との関係についても疑問を提起した。**国家元首**は，特に，新たなグルジア政権から発せられた関係調整への肯定的なシグナルを見出していると語った。

5. テクスト処理の概要

5.1. データと取得方法

まず，分析するデータの取得方法について述べる。今回「大統領」「首相」の 2 つに分

析対象が定まっているので，機械的に処理することができると考えた。そこで単純なプログラムによってデータの取得をおこなう。実装する処理は次の通りである：1）パラフレーズされた表現として文字列をテクストの中から探し出し，2）直前に出てきた文字列と同一のものか異なるのか判定し，3）比較ののべ回数を記録し，4）各表現の出現回数を数え上げる。以上の処理によってあるテクストの中で対象が指示される回数とパラフレーズの使用回数の比率を求めるための数値を得ることができる。使用するプログラムの処理について詳しくは次節で説明する。

次に対象とするデータについて述べる。取得元 http://www.1tv.ru/ から 2012 年 10 月から 2012 年 12 月までの期間に取得する。[8] このソースは言語の規範性においては十分信頼に足ると見なすことができる。ページから自動的にテクストを取得する方法には様々なものが考えられるが，今回分析対象のキーワードの含まれるページを正確に見つけることができなかったため，手作業で 204175 語からなる 2314 テクストを作成し，機械による処理にかけた。中には違う見出しで同じ記事を載せているページもあり，そのような場合も除外した。

調査対象は「大統領」「首相」の 2 人に限定する。ロシア連邦では，2000 年から現在まで事実上，ヴラジーミル・プーチンとドミートリー・メドヴェージェフのタンデム体制で政治が行われている。したがって，役職名と個人名は一対一に対応する。この一対一の対応は偶然のものであるが，この対応のため，分析対象とするデータを単純にできる。

タンデム体制によって考慮しなければならないこともある。2008 年 5 月にプーチンが大統領の職を退いたのち，2008 年 6 月から 2012 年 5 月まで首相として務めており，その間メドヴェージェフが大統領として国政を主導した。この体制のため，2000 年から 2008 年および 2012 年から現在までの期間と 2008 年から 2012 年までの期間において個人名称と「首相」「大統領」の役職名が交替して用いられる。機械的に処理を行う場合，この組み合わせの交替に注意する必要がある。

5.2. キリル文字からラテンアルファベットへの変換

キリル文字からラテンアルファベットへの換字には次の変換規則を採用した。ただし，硬音記号 Ъ および ъ については特定のラテンアルファベットへの置換は行わず，空文字とした。[9]

---

[8] 分析するテクストは，「プーチン」「メドヴェージェフ」のタグがついた一覧から表題を除いた部分から作成した。

[9] キリル文字からラテン文字へは多くの翻字方法がある。しかし，それらの表は欧文文字コード（ANSI コード）表に定義されていない文字（š, ž, è など）が用いられる場合がある。今回の翻字の目的はプログラムにおける処理を容易にするためのものである。したがって，今回は文字化けによ

変換表

А → *A*	*а* → *a*	*Р* → *R*	*р* → *r*
Б → *B*	*б* → *b*	*С* → *S*	*с* → *s*
В → *V*	*в* → *v*	*Т* → *T*	*т* → *t*
Г → *G*	*г* → *g*	*У* → *U*	*у* → *u*
Д → *D*	*д* → *d*	*Ф* → *F*	*ф* → *f*
Е → *Je*	*е* → *je*	*Х* → *X*	*х* → *x*
Ё → *Jo*	*ё* → *jo*	*Ц* → *C*	*ц* → *c*
Ж → *Zh*	*ж* → *zh*	*Ч* → *Ch*	*ч* → *ch*
З → *Z*	*з* → *z*	*Ш* → *Sh*	*ш* → *sh*
И → *I*	*и* → *i*	*Щ* → *Shch*	*щ* → *shch*
Й → *J*	*й* → *j*	*Ъ* → なし	*ъ* → なし
К → *K*	*к* → *k*	*Ы* → *Y*	*ы* → *y*
Л → *L*	*л* → *l*	*Ь* → '	*ь* → '
М → *M*	*м* → *m*	*Э* → *E*	*э* → *e*
Н → *N*	*н* → *n*	*Ю* → *Ju*	*ю* → *ju*
О → *O*	*о* → *o*	*Я* → *Ja*	*я* → *ja*
П → *P*	*п* → *p*		

例）

変換前：O спорте сегодня говорили и на совещании Дмитрия Медведева с вице-премьерами.

(http://www.1tv.ru/news/social/220697)

本日行われたメドヴェージェフ氏と副首相らの会談においても話題となったのはスポーツについてだった。

変換後：O sportje sjegodnja govorili i na sovjeshchanii Dmitrija Mjedvjedjeva s vicje-prjem'jerami.

5.3. 文字列処理

以下のテクスト処理はスクリプト言語である Python を使っておこなう。

大文字・小文字の区別，コンマやハイフンなどの記号，Prjemjer-ministr Vladimir Putin と

る処理のミスを防ぐため, ANSI コード内に存在するラテンアルファベットでそれぞれのキリル文字を一意に置き換えるための独自の翻字表を用意した。

いった役職名＋個人名称の付語による表現は次のように処理する：1）まず前処理として文字列中の文字全てを小文字に統一する。2）そして，記号「.」「,」「?」「-」「'」「"」「:」「;」「(」「)」「«」「»」を空文字に置換して消去する。[10] 例えば，「vicje-prjem'jer」中の記号「-」は変換後「vicjeprjemjer」のように，空文字によって置換される形で消去・連結される。記号を消去する理由は，それらがテクスト中の単語数を数える上で障害となるためである。3）そして全ての役職名と個人名の組み合わせの変換表を基に，役職名＋個人名称の付語による同格表現中の曲用語尾を消去し，[11] 単語と単語の間のスペースを連結して一単語とする。この処理はテクスト内の文をすべて配列にして，語数を数えるとき，ひとまとまりと見なすようにするためである。

例）

Премьер-министр Дмитрий Медведев дал большое интервью пяти российским телеканалам: Первому, России, НТВ, Рен-ТВ и "Дождю". (http://www.1tv.ru/news/crime/221563)

ラテンアルファベット翻字後（囲み線は次の処理で扱われる要素を指す）：*Prjem'jer-ministr Dmitrij Mjedvjedjev dal bol'shoje intjerv'ju pjati rossijskim tjeljekanalam: Pjervomu, Rossii, NTV, Rjen-TV i "Dozhdju".*

1）の処理後：*prjem'jer-ministr dmitrij mjedvjedjev dal bol'shoje intjerv'ju pjati rossijskim tjeljekanalam: pjervomu, rossii, ntv, rjen-tv i "dozhdju".*

2）の処理後：*prjemjerministr dmitrij mjedvjedjev dal bolshoje intjervju pjati rossijskim tjeljekanalam pjervomu rossii ntv rjentv i dozhdju*

3）の処理後：*prjemjerministrdmitrijmjedvjedjev dal bolshoje intjervju pjati rossijskim tjeljekanalam pjervomu rossii ntv rjentv i dozhdju*

一連の処理を行うことで，文中の語数を正確に数え，各語をリストとして適切に格納でき

[10] 仮説の第二の部分の文の個数を調べるために作られるリストではピリオド「.」は消去されない。

[11] この処理には正規表現を用いた。たとえば，正規表現 “prjezidjent(a|u|om|je)?¥srossii”は，prejezidjent という文字列の後に a か u か om か je という文字列が 0 回あるいは 1 回続き，半角スペース（¥s）の後に rossii という文字列が後続するような場合を見つけることを意味する。上に挙げた場合を一続きの文字列“prjezidjentrossii”に置き換える。python には正規表現を扱うため用意された re モジュールを使用した。

る。

得られたデータは R が処理しやすいようにデータフレームの形でテクストファイルとして保存する。具体的には次のような形となる：

	text_number	length	total_comparison	paraphrase	same_words
1	1000	112	4	4	0
2	1001	115	2	2	0
3	1002	139	5	5	0
4	1003	111	3	3	0
5	1004	100	3	3	0
...	...	...	...	...	...

上の text_number はファイルを作成するときに設定した，一意に定められたファイル名を指す。length はそれぞれのテクストに含まれる句読点を除いた語数を表わす。「大統領」「国家元首」など同じ人物を指す表現が出現した場合，同じ表現が続けて現れるかどうか比較を行う。total_comparison は比較の総合回数, paraphrase は違う表現が用いられた回数, same_words は同じ表現がつづけて用いられた回数を示す。例えば，最初の 1000 という名前のファイルは 112 語から成り，「大統領」を表わす表現は 4 回出現し，そのうち，言いかえが用いられたのは 4 回, 同じ表現がつづけて用いられたのは 0 回, ということになる。

今回の分析において，テクスト処理を python，統計処理を R に分担させるという形をとった。[12] Python によって出力したテーブル形式のテクストファイルを統計処理用ソフトウェアの R にデータフレームとして読み込ませるというバッチ型の処理である。語数を求めるためにリストの長さを求める関数を用いる。文の数はリスト内のピリオドの数を数えることで求める。

6. 統計分析

統計手法は大きく分けて「記述統計学」「推測統計学」の 2 つに分類される。例えば無数に存在する言語に関するデータをとることは現実的に不可能であり, 調査によって得られたデータは有限のものであることが多い。「記述統計学」は得られたデータの範囲内における傾向や分析結果に注目する。それに対して「推測統計学」は得られた有限のデータ

[12] 用いるのは, Python2.7.3 および R3.0.0 である。いずれも Windows 64bit 環境の下で処理を行った。R は python の機能と同じくスクリプト言語の体裁をとっているため共通した面が多い。同じ統計処理は Excel でも行うことができ，その場合 CUI コマンドを覚える必要がない。しかし関数による計算をくり返す必要があり，かえって煩雑となってしまう。これらの理由から R を今回使用する。

データフレームは，行列に似た形のデータだが，行列とは異なり，量的変数と質的変数といった異なるタイプのデータを同時に取り扱うことができる。

から，現実の言語に対して分析を行い，結果を導き出すことが目的となる。

本稿では，まず標本から得られた記述統計学的な結果を提示する。そのあと，限られた言語データから言語の全体像を明らかにする推測統計学的な方法によって分析をおこなう。

用いるデータは 2012 年 10 月から 2012 年 12 月まで集めた 2,314 テクスト 204,175 語である。これらに対して上で述べたような処理を施し，得られた数値に推測統計学的な分析を施して，仮説の検証を行なう。

6.0. 標本から得られたデータの数値要約

下の表は，標本から得られたデータ数値をテーブル形式にしたものである。5.3.の終りで述べたように，データフレーム形式にした全データ行 2,314 行をそれぞれ列ごとに処理したものである：

プーチン：

	語数	指示回数	言いかえ回数	同一表現回数
合計	118498 語	2583 回	2183 回	400 回
平均	91.50425 語	1.994595 回	1.685714 回	0.3088803 回
中央値	77 語	2 回	1 回	0 回
最大値	483 語	19 回	15 回	5 回
最小値	15 語	0 回[13]	0 回	0 回
最頻値	56 語	1 回	1 回	0 回

まず上の表の行が何を表わしているのかについて説明する。「語数」は 1 つのテクストごとの語数を表わす。「指示回数」は「ヴラジーミル・プーチン」「ドミートリー・メドヴェージェフ」が 1 つのテクスト内で指示された回数，「言いかえ回数」はそれらがパラフレーズされた回数，「同一表現回数」は同じ言い回しによって指示された回数を指す。

表の列は標本による数値要約を表わす：「合計」は変数ごとの合計値，「平均」は「合計」を変数の個数で割ったものである。中央値は変数をソートしたときに中央に位置する値のことで，この場合データ数は 2,314 個と偶数なので中心に位置する値の 2 つの平均値を中央値とする。最頻値はある変数のうち最も頻繁に観測される値を指す。最大値は変数の中で最も大きな値，最小値は最も小さな値を指す。上の表の個々のセルの値は，表の行にお

[13] この表では得られた生の標本データについて数値要約を行ったため，0 という値も存在する。しかし，次に行う処理においてはこのような 0 の値の含まれる観測値は除外した。

けるデータに表の列で表された処理を施したものである。

次に示すのはメドヴェージェフについてのデータである。

メドヴェージェフ：

	語数	指示回数	言いかえ回数	同一表現回数
合計	85677 語	2008 回	1743 回	265 回
平均	83.91479 語	1.966699 回	1.70715 回	0.2595495 回
中央値	74 語	2 回	1 回	0 回
最大値	385 語	14 回	10 回	6 回
最小値	22 語	0 回	0 回	0 回
最頻値	78 語	1 回	1 回	0 回

上に挙げたデータのうち重要となるのは平均, 中央値, 最頻値である。それらによると，個々のテクストは 80 語から 90 語から成っている。その中で「ヴラジーミル・プーチン」「ドミートリー・メドヴェージェフ」が指示される回数は 2 回であることが多い。その場合，パラフレーズと同一表現はおよそ 17:2 の割合で用いられる。以上が今回得られた標本データによる数値の要約である。

6.1. 推測統計学の手法による分析 I

統計的手法として，仮説検定法を用いる。この検定方法の特徴は，収集されたデータがたまたま有意な値を持つ確率である p 値を考慮する点にある。実証的な研究を行う場合，言語資料の量的な側面に着目する必要がある。しかし，集められた言語資料は，無数に生み出される（母集団）のほんの限られた一部分でしかない。そこには，偶然分析に用いられたサンプルに仮説にとって有意な結果が含まれているのではないかという反論の余地がある。このような反論に対処するために母集団とサンプルとの誤差の範囲を示す p 値がある。p 値とは母集団の部分をなすサンプルと母集団とがどれほど食い違っているかを示す値である。p 値がある一定の水準（有意水準）より小さければ，分析結果の有意性が示される。

そのため, 分析結果が仮説について有意であることを証明するために仮説検定を採用して調査を行う。

仮説検定の大まかな流れは次のようになる：[14]

[14] 言語学に対する統計分析の応用については，石川慎一郎ほか『言語研究のための統計入門』くろしお出版，2010 年 を参考とした。

1. 仮説が有意かどうか判断するための有意水準を求める。本研究では 5%とする。
2. 帰無仮説と対立仮説を設定する。帰無仮説を棄却し，対立仮説が有意であることを示すことが分析の目的となる。帰無仮説は作業仮説を否定するものである。
3. 帰無仮説を正しいと仮定した上で，検体統計量を計算する。
4. 検定統計量が従う標本分布に照らして，手元の標本で計算された値が，どの程度の確率で出てくるのかを評価する。この確率値を p 値と呼ぶ。現実には有意な差が存在しないのにも関わらず，誤差や偶然によってたまたま有意であるかのような差が生じる確率である。この p 値が一定値（有意水準）以下である必要がある。
5. p 値が有意水準を下回っていた場合，対立仮説が採用される。そうでない場合は，帰無仮説が採用され，定期した作業仮説の優位性を示すことはできなくなる。

まず，本当に報道テクスト内で同じ対象をなるべくパラフレーズして指示する傾向があるのかを計量的に示す。そのことにより，本稿の提起した問題が成り立つことになる。

6.1.1. 準備

今回行う分析の目的は，パラフレーズ・同一表現の出現頻度を比較して，パラフレーズの出現頻度の方が多いことを示し，提起された問題が有効であることを示すことである。最も単純な統計分析として集計と比率がある。しかし，その方法だと，調査対象の母集団全体のうち偶然有意な結果をもつデータを採取してしまったという誤差の要素を考慮することができない。[15] 標本抽出に伴う誤差を考慮する必要がある。この誤差を調べることによって「実際の全体の数にどれくらい近く推定することができるのか」を知ることができる。この推定された全体の数は，ここでは母平均となり，採集された例から得られた結果は標本平均となる。

ここでの調査の目的は「同じ対象を指示する場合，パラフレーズされる頻度の方が同じ表現が繰り返される場合より多い」ことを示すことである。

6.1.2. 分析 I

問題提起である「報道文では『大統領』『首相』を指示する際，同じ語句の使用は避けられる傾向にある」ことを母平均の区間推定によって示すことである。この傾向を示すことによって，本稿の問題提起が妥当であることを確認する。手順においては，まず標本平均を求め，その値が推定される母平均と等しいかどうか確かめるために，t 検定にかける。

[15] このような誤差は調査対象の母集団を調べつくすことのできるケースにおいては考慮する必要はない。しかし，現代の言語の状態を対象とする今回のような場合，母集団の様子を網羅的に調べることは不可能である。そのため，誤差の要素を考慮する必要が出てくる。

一般に，母集団の分散が分からない場合，*t* 検定は以下の検定統計量を持つ：

$$t = \frac{\bar{X} - \mu}{\hat{\sigma} / \sqrt{n}}$$

この式は標準正規分布にしたがう検定統計量

$$Z = \frac{\bar{X} - \mu}{\sqrt{\sigma^2 / n}}$$

を変形したものである。ここで $\bar{X}$ は得られた標本からの平均，ギリシア文字 μ は母平均を意味する。σ^2 は母分散といい，母集団におけるデータのばらつき具合を表わす。偏差は平均からそれぞれのサンプルの値がどのくらい離れているのか表し，正負両方の符号がつく。分散は符号を正に統一するため偏差の二乗をとり，平均で割ったものである。n はデータの数であるサンプルサイズを意味する。$\sqrt{(\sigma^2/n)}$ は平均値と同じ単位にするために σ^2 の平方根をとったものであり，母標準偏差を表わす。上の検定統計量は，標本平均から推定した母平均を引いたものを母標準偏差で割ったものである。

上の標準正規分布と *t* 検定のための検定統計量には一つの違いがある。それは，母集団の分散の度合いが分からないということである。そのため，*t* 検定の σ には誤差を含むことを示す「キャップ」（^）が上についている。

t 検定の実施にあたって，R に用意されているメソッド t.test( ) を用いる。

手順 1.

有意水準は 5%に定める。

手順 2.

今回の分析において母分散は未知であり，目的とするのが，母平均の値そのものなので，*t* 検定において次のような帰無仮説と対立仮説を設定する。

- 帰無仮説 H_0：母平均と標本平均との間には差がない。
- 対立仮説 H_1：母平均と標本平均との間には差がある。

一つ注意しておかなければならないことがある。それは，仮説の成立には帰無仮説の棄却が**できない**ことを示す必要がある，ということである。通常の仮説検定の流れでは帰無仮説は棄却される。棄却するべき帰無仮説と対立仮説とを入れ替えた理由として，分析に使用する R の関数 t.test( ) がとる mu オプション（この mu オプションは帰無仮説として母平均 μ を設定するものである。この引数に値を設定しないと帰無仮説がデフォルトとしての値 mu = 0 として検定が行われてしまう）が帰無仮説に関係するものであること

に起因する。この検定での目的は標本平均と母平均の値が等しいことを示すことである。そこで，mu オプションに棄却するべき帰無仮説の値として標本平均の値を設定して比較を行う。もし等しければ出力される *p* 値は有意水準を上回るはずである。そうなると，結果的に帰無仮説を棄却することはできなくなるが，検定の目的とする「標本平均と母平均は等しい」ことを統計的に示すことができる。

このように仮説を設定した理由は，mu オプションは帰無仮説として母平均を設定することにある。ここでの検定の目的は「母平均と標本平均とが等しい」ことを示すことにある。しかし，それは mu オプションで指定される帰無仮説で表される。そこで本節の検定そのものもオプションに合わせたほうが複雑さを避けるため，目的とする命題を帰無仮説として設定した。[16]

手順 3.

帰無仮説 H_0 を正しいと仮定した上で検定統計量を計算する。サンプルから得られた平均と母平均とがどれほど離れているか求めるために，t 検定を R の関数 t.test( ) によって行う。t.test( ) は引数に原データを指定する。Mu オプションには標本平均の値を指定する。標本平均と母平均の値が等しければ，出力される *p* 値は有意水準 0.05 を上回る。推定の蓋然性を表わす信頼度はデフォルトの 95%とする。

手順 4.

得られた出力結果から *p* 値が手順 1 で定めた有意水準 0.05 を上回るかどうかを確かめる。手順 2 で設定した有意水準 5%より *p* 値が小さい場合（$p < \alpha$ のとき）に帰無仮説を棄却し，対立仮説を採用する。そうでない場合は対立仮説を棄却する。

手順 5.

計算の結果，次のような結果が得られた。R による出力を以下に示す：

この作業を，パラフレーズの現れる場合と同じ表現が繰り返される場合の二度に分けて行う。そうして母平均と等しいと推測された標本平均の値を比較する。

得られた結果は以下のようになる。以下の出力はパラフレーズについての分析を指す：

[16] 本来は推定において帰無仮説は棄却されるべきものであり，この分析で得られた結果は，統計学的には不完全なものである。ただし，ここでの分析は本論考における問題が成り立つかどうかを示すための副次的なものである。そのため，このような本来的でない使い方であっても採用することにした。

```
t = 0, df = 2315, p-value = 1
alternative hypothesis: true mean is not equal to 1.695164
95 percent confidence interval:
  1.636241 1.754088
sample estimates:
mean of x
  1.695164
```

以下の表示は同一表現についての分析を指す。

```
t = 0, df = 2315, p-value = 1
alternative hypothesis: true mean is not equal to 0.287133
95 percent confidence interval:
  0.2611668 0.3130991
sample estimates:
mean of x
  0.287133
```

検定の結果得られた p 値は 1（p-value = 1）であり，これは有意水準 0.05 より大きい最大値である。したがって，標本平均と母平均が等しいという帰無仮説は棄却できず，仮に 100 回試行をおこなっても一致する可能性がある。この検定において帰無仮説は本来棄却されるべきものであり，正しい手順を踏んでいない。しかし，この検定は本論考の問題設定が意味をもつかどうか確かめるという副次的なものであるため，このような使用方法を採用した。

結果として母平均と等しい標本平均の値は，パラフレーズの方が同じ表現の繰り返しよりも多く用いられるという傾向を示している。したがって，報道テクストの中で同じ言い回しの繰り返しは避けられる傾向にあることが明らかになった。

6.1.3. 分析Ｉのまとめ

母平均を求めた結果，パラフレーズが現れる場合が同じ表現が繰り返される場合よりも多いことが明らかになった。このことから，本稿で定期した「同じ対象を指示するのに，同じ表現の繰り返しは避けられる傾向にある」ということが明らかになった。

この節では，「同じ対象がテクスト内で何度か言及される場合，同じ表現の繰り返しよりもパラフレーズが好まれる傾向にある」ことを統計的に示した。この結果は本稿の問題提起が有効であることを示すものである。

6.2. 推測統計学の手法による分析Ⅱ

6.2.1. 準備

ここでの分析の目的は，「パラフレーズの出現とパラフレーズされた表現と表現の間隔の相関性が存在することを示す」ことである。表現と表現の間隔として「語数」「文の数」といった要素が考えられる。

今回は上の分析とは違い，「パラフレーズの出現頻度」と「語の数」や「文の数」といった二つ以上の変数との関係・相関を問題とする。語数，文の数が増加するに従って，同じ対象が言いかえられる回数も増えるのではないかという仮説を立てる。これは，2 つの変数の間の「相関」を問題としている。そのため，統計的仮説検定の手法の中の「相関分析」をおこなうこととする。

今回おこなう相関分析は 2 つの変数の間の「相互関係」についてのみ問題にする。相関分析は，ある変数 A のもう片方の変数 B への「影響」や「原因」に関しては何も主張することができない。相関分析が主張するのはあくまでも 2 変数間に相関関係があるという事実である。このことに関しては言語研究と統計に関する文献の中で触れられている。[17]

相関分析には設定できる変数が 2 つだけなので，「語数」「文の数」の２つの場合に分けて 2 度検定を行う。

分析Ⅱに際して，前の節で説明した文字列処理に加えて以下の処理を行う。語数はリストの長さから求めるとする。文の数はテクスト内のピリオドを数えることによって導き出す。

以下のテクストを例に説明する。対象とする表現を глава государства とし，一度現れてからの語数，ピリオドの数を問題とする。

例）

‹...› ***<u>Глава государства</u>*** *уже в восьмой раз общается с журналистами в подобном формате. Встреча вызвала огромный интерес.*

На сегодняшнее мероприятие были аккредитованы свыше 1200 представителей средств массовой информации из Москвы и практически всех регионов страны.

В течение 4,5 часов ***<u>глава государства</u>*** *отвечал на вопросы о борьбе с коррупцией, судебной системе, выборах губернаторов, уголовных делах против оппозиционеров и чиновников. Одной из главных тем стала реакция Госдумы на принятый в США "акт Магнитского" и особенно запрет*

[17] 石川ほか『言語研究のための統計入門』，90 頁; Stefan Th. Gries, *Statistics for Linguistics with R* (Hague: Mouton, 2009), p. 237

американским гражданам усыновлять сирот из России.（http://www.1tv.ru/news/polit/222504）

〈…〉**国家元首**はすでに 8 回このような形での記者会見をおこなっている。会見には大きな注目が集まった。

本日の催しのために，モスクワおよび実質上すべての地方の報道各社から 1200 人以上の代表者が派遣された。

四時間半にもわたる会見の中で**国家元首**は汚職との戦い，司法機関，知事の選出，敵対候補や役人に対する刑事事件についての質疑に答弁した。それらの主要なものの一つには，アメリカ合衆国で採択された「マグニツキー法」に対するロシア国会の反応，とりわけアメリカ国民によるロシアからの養子縁組の禁止があった。

表現と表現のあいだに，ピリオドが 2 つ続けて打たれた場合，1 文とみなす。すなわち，上の例では，***Глава государства*** ‹...› . *На сегодняшнее* ‹...› *страны*. *В течение 4,5 часов* ***глава государства*** ‹...› と二つピリオドが打たれているため，1 文が表現と表現の間に置かれていることになる。

有意水準は 5%に定める。帰無仮説を「母集団において相関が 0 である」とする。今回は，帰無仮説を正しいと仮定した上で検定統計量を計算する。R に備えられている関数 cor.test( )を用いて無相関検定を行う。

6.2.2 分析Ⅱ

ここでの作業仮説は，「報道文体においてパラフレーズされるか否かは，表現と表現との間隔の大きさが関係する」である。この仮説を検証するためには，表現と表現との間隔と言いかえの出現との間に相関関係が存在することを示せばよい。そのため，ここでは相関分析の内の無相関検定を使用する。無相関検定とは 2 つの確率変数の間に相関関係があるか否かを判定する。検定に使用する相関係数そのものは相関の強さを表わす。しかし，無相関検定は相関関係の有無に関するものであって，相関関係の強弱や因果関係について言及するものでない。無相関検定は得られた相関係数が統計的に有意かどうかについて判断するものである。相関係数には，ピアソンの積率相関係数（Pearson's product-moment correlation）を使用する。統計学における「相関係数」は通常，この係数を指す。[18] 無相関検定の検定統計量は次の式から得られる：

[18] 相関係数には，ピアソンの積率相関係数のほかに，スピアマンの順位相関係数（Spearman's rank correlation coefficient），ケンドールの順位相関係数（Kendall tau rank correlation coefficient）がある。ピアソンの積率相関係数と異なり，これらの相関係数はノンパラメトリックである。言語研究のための統計についての参考書籍のうち，これらの相関係数について説明したものに，Claudia Meindel, *Methodik für Linguisten* (Tübingen: Narr, 2011), pp. 222-231 がある。

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

ここで r は標本相関係数，n は分析のため実際に集めてきたデータの大きさ，つまりサンプルサイズを指す。変換された t は，帰無仮説のもとで自由度 df = n – 2 の t 分布にしたがう。t 分布とは統計学で用いられる確率分布の一つで，形状はつりがね状で正規分布に似ている。t 分布の自由度はこの分布の形状を決定する。

検定統計量を求めるには，相関係数が必要となる。今回使用するピアソンの積率相関係数 r_{xy} は次のように求める：

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

データが二つの数値で組をなす列として $\{(x_1, y_1),(x_2, y_2),(x_3, y_3), \ldots(x_n, y_n)\}$ で与えられる。$\bar{x}$, $\bar{y}$ はそれぞれ x_i, y_i の標本平均を意味する。n は検定にも用いたサンプルのサイズを表わす。

R には無相関検定を実行するための関数 cor.test( ) が用意されている。この関数を用いて分析をおこなう。有意水準は分析 I と同様に，0.05 とする。

引数として渡す表現と表現の間の語数や文の数のデータは要素の合計／カウント回数の合計で得られる比率とする。例えば，語数について，あるテクストの中で 5 回パラフレーズが用いられ，それぞれの間に 47 語, 4 語, 11 語, 25 語, 10 語が置かれるとする。その場合比率は（47+4+11+25+10）／ 5＝19 となる。別のテクストでは 3 回パラフレーズが用いられ，それぞれの間に 9 語, 25 語, 66 語が置かれる。比率は，（9+25+66）／ 3＝33 となる。このように，それぞれのテクストにおいて表現の出現回数がまちまちであるため，そのままでは処理を通すことのできるデータフレームを作ることができない。以上の理由よりデータフレームの変数部分に比率を使用する。

また，表 1 であげた表現が一度も現れず，比較回数が 0 回のテクストは分析対象から除外した。

語数との相関

「ある表現がパラフレーズされるか否かは表現と表現の間の語数，文の数と有意な関係を持つ」という作業仮説に従って帰無仮説と対立仮説を設定する。

- 帰無仮説 H_0：パラフレーズの出現回数はパラフレーズされた表現と表現の間の語数

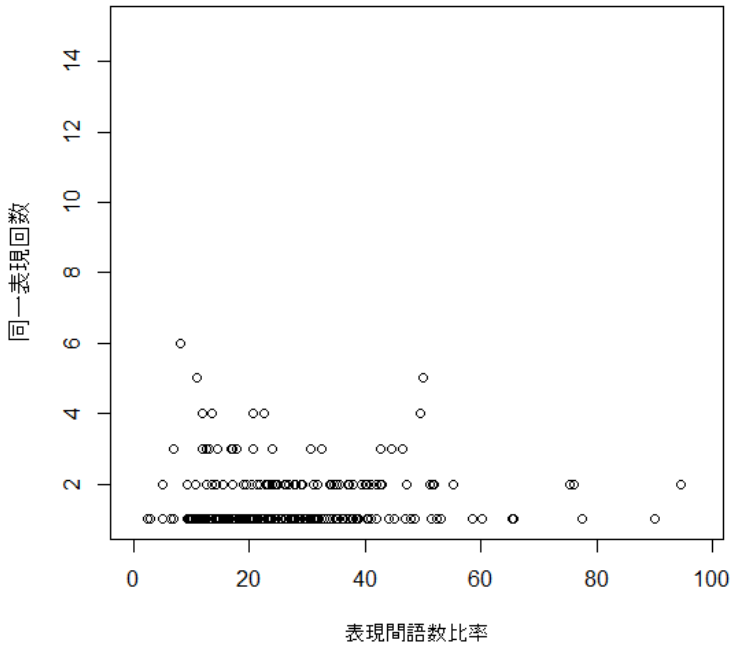
は無関係である。

- 対立仮説 H_1：パラフレーズの出現回数はパラフレーズされた表現と表現の間の語数と有意な関係を持つ。

分析結果の出力は以下のようになる。表現間におかれた語の数の比率と同じ表現が続く回数との相関を示す。R の出力結果の他に，標本データによる散布図を視覚的な理解のため挙げる。この散布図はそれぞれのテクスト中の同一表現の出現する回数と表現間の語数の比率との関係をプロットしたものである：

```
data:  表現間語数比率 and 同一表現回数
t = 7.4392, df = 1274, p-value = 1.856e-13
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1508464 0.2560462
sample estimates:
      cor
0.2040353
```

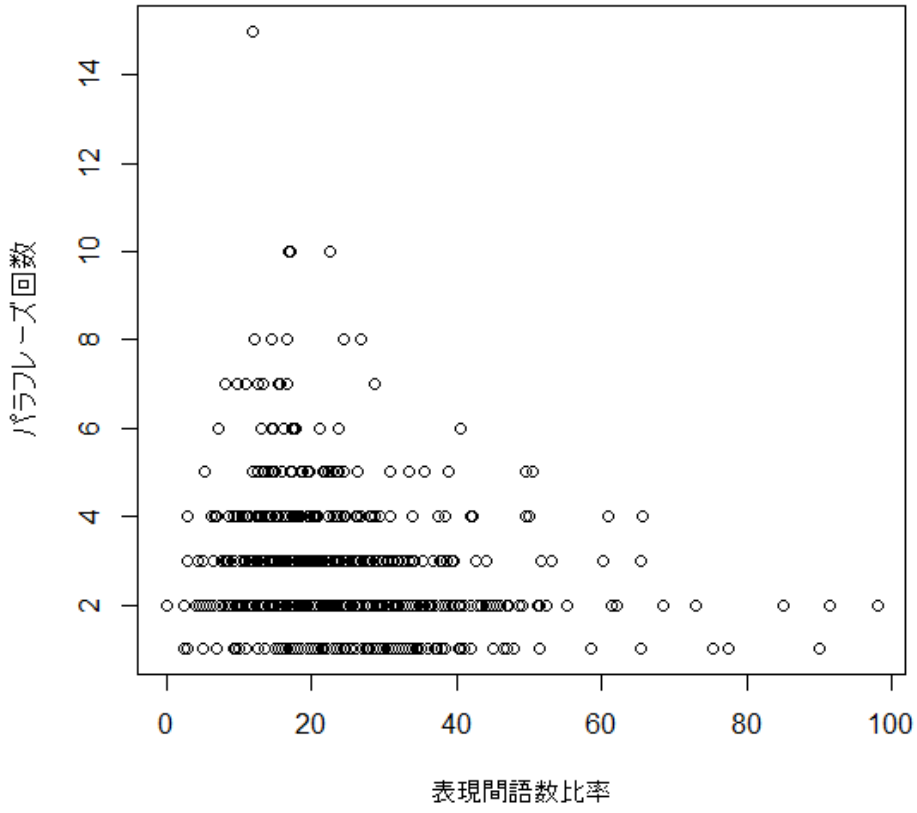
散布図：



次に示すのは，表現間に置かれた語の数の比率とパラフレーズの出現回数との相関である。散布図はそれぞれのテクストにおいて，言いかえの出現する回数と表現間の語数の比率との関係をプロットしたものである：

```
data:  表現間語数比率 and パラフレーズ回数
t = -5.9259, df = 1274, p-value = 3.993e-09
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.2167107 -0.1098904
sample estimates:
        cor
-0.1637806
```

散布図：



上の結果で浮動小数点の形をとった *p* 値（分析結果において下線で示す）がそれぞれ 1.856e-13，3.993e-09 となっている。これはそれぞれ 1.856×10^{-13}，3.993×10^{-9} を意味する。また，t は検定統計量によって実際に得られた t の値，df は t 分布の自由度を示す。cor 下の実数値が相関係数をあらわす。これらの *p* 値は上で定めた有意水準 0.05 を下回るので帰無仮説は棄却され，表現間の語数の比率とパラフレーズの回数，同一表現の出現回数は相関関係にあることが明らかになった。しかし，cor 以下に表示される得られた相関係数は-0.2 以上 0.2 以下あるいはその周辺にあるので，相関関係はひじょうにゆるやかである。

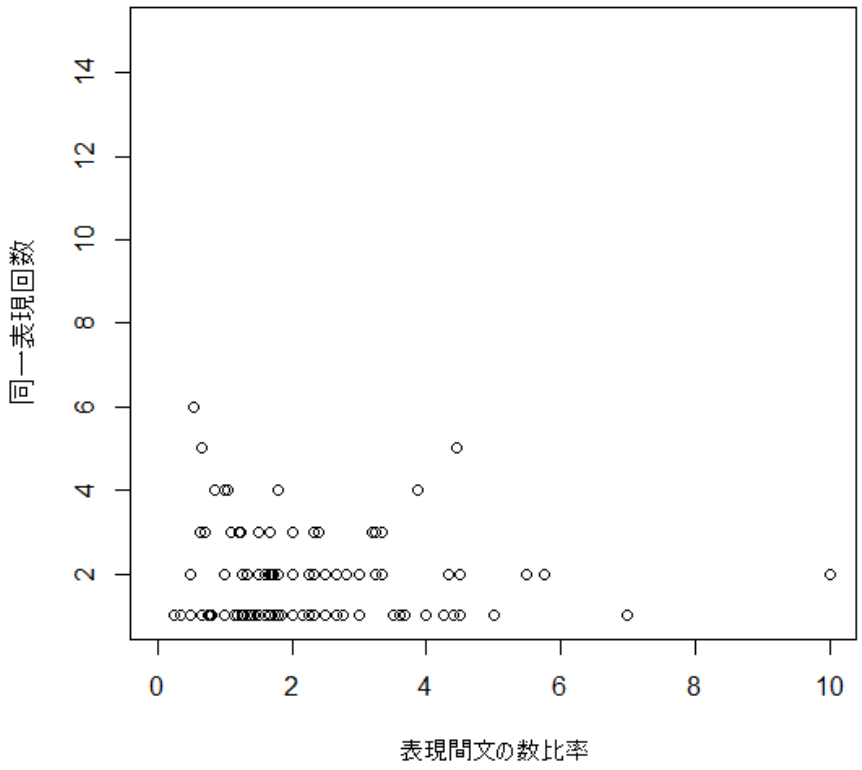
文の数との相関

- 帰無仮説 H_0：パラフレーズの出現回数はパラフレーズされた表現と表現の間の文の数は無関係である。
- 対立仮説 H_1：パラフレーズの出現回数はパラフレーズされた表現と表現の間の文の数と有意な関係を持つ。

分析結果の出力は以下のとおりである。まず表現間に置かれた文の数の比率と同じ言い回しが続いた場合との相関を示す：

```
data:  表現間文の数比率 and 同一表現回数
t = 6.4841, df = 1274, p-value = 1.274e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1250847 0.2313444
sample estimates:
        cor
0.1787357
```

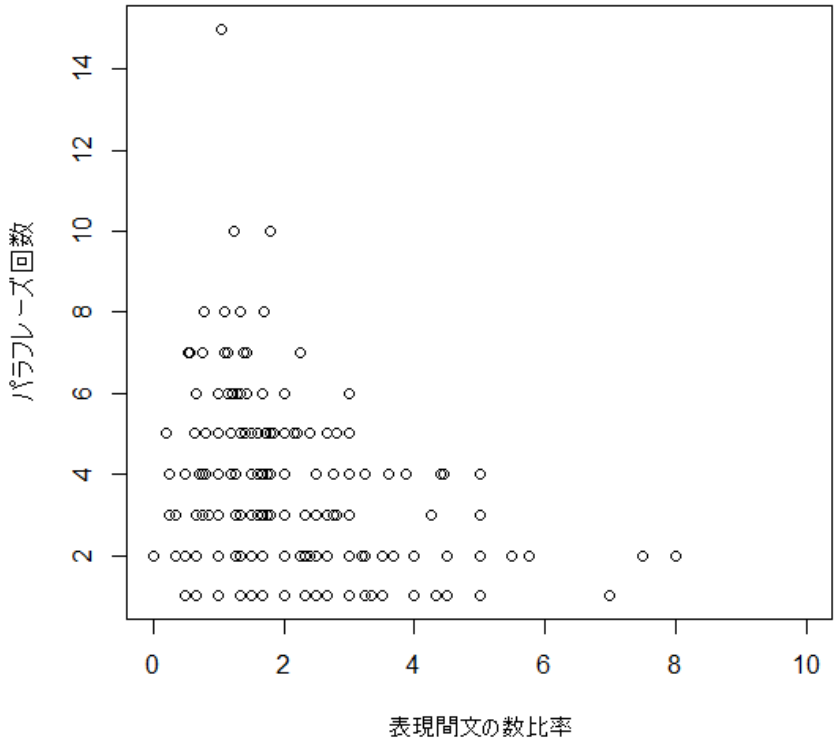
以下に挙げる散布図はそれぞれのテクストごとに，同一表現の出現する回数と表現間の文の数の比率との関係をプロットしたものである：



次に示すのが，表現間に置かれた文の数の比率と言いかえられた表現の出現回数との相関である：

```
data:  表現間文の数比率 and パラフレーズ回数
t = -6.6075, df = 1274, p-value = 5.73e-11
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.2345614 -0.1284317
sample estimates:
        cor
-0.1820266
```

以下に挙げる散布図はそれぞれのテクストごとの，言いかえの用いられる回数と表現間の文の数の比率との関係をプロットしたものである：



上の結果で浮動小数点の形をとった p 値が 1.274e-10，5.73e-11 となっている。これはそれぞれ 1.274×10^{-10}，5.73×10^{-11} を意味する。これらの値は p 値は上で定めた有意水準 0.05 を下回るので帰無仮説は棄却され，表現間の文の数の比率とパラフレーズの回数，同一表現の出現回数は相関関係にあることが明らかになった。しかし，cor 以下の相関係数は-0.2 以上 0.2 以下の範囲に収まるので，非常に弱いものである。

以上の結果から，言いかえと表現間の間隔の間には相関関係があることが示された。しかし，ここで注意しなければならないことがある：この検定の結果は表現間の語数および文の数とパラフレーズの回数との間に相関関係があることを示すにすぎない。表現間の語数や文の数の比率と言いかえとの間に特に強い相関関係があるのではなく，単に相関が存在することを無相関分析は示しているにすぎない。

6.2.3. 分析Ⅱのまとめ

この節では，「『大統領』『首相』という同一人物を二度以上指示する場合，パラフレーズが用いられるか，同じ表現が繰り返されるかどうかは表現と表現の間隔の大きさである

『語数』『文の数』と相関関係にある」ことを統計的に示した。統計処理として，無相関検定を使用した。

7. 結果のまとめおよび課題

7.1. 結果のまとめ

本研究ではロシア語の報道文体においてテクストの中で一度用いられた表現が繰り返されないことをパラフレーズと名づけ，その傾向が本当に存在するのか，そしてその傾向はテクスト内のどの要素と関係して引き起こされるのか，という問題について取り扱った。具体的には，大統領と首相の役職名，および個人名がテクスト内で繰り返される頻度を対象とした。パラフレーズの使用は単調さを避けることを目的と仮定し，単調さの印象は表現と表現の間に起因するとした。そこで，パラフレーズの出現頻度に関係するのは表現と表現の間隔という仮説を提起し，1）語数，2）文の数，の二つを表現間の隔たりの候補として考察した。すなわち，同じ表現が繰り返し用いられる場合，間に語数，文の数が関係すると仮定し，検証を行った。

結論として次のことを述べることができる。ロシア語の報道テクストの中で同じ表現の繰り返しは避けられる傾向にある。本研究では同じ対象を再び指示する場合，言いかえが用いられるか，同じ表現が繰り返されるのかは，表現間の語数，文の数といった表現と表現との間隔との間に相関関係が存在することを明らかにした。これに際して，無相関検定を用いたが，この検定は単に相関の有無があることを示すのに過ぎない。

具体的な相関の強さについての値についても触れる。表現間の語数と同一表現の相関は0.204,035,3，パラフレーズの出現との相関は-0.1637806 であった。表現間に置かれた文の数と同一表現の出現比率との相関は 0.1787357 で言いかえとの相関は-0.1820266 であった。これらの値は絶対値 0.2 の周辺にある。そのため，相関関係は非常にゆるやかなものである。このことは，言いかえが文法の規範的な問題というよりは文体的な細かなレベルの問題であることを意味している。語数，文の数と言いかえのあいだの相関に特に強い意味がある，表現の間に置かれる語数や文の数がパラフレーズの出現の原因となる，といったことを示すためには別の検定方法をとる必要がある。

7.2. 今後の課題

今回問題とした報道文体における言いかえの問題はロシア語だけではなく，英語やドイツ語など他のヨーロッパの言語においても見られるものである。ロシア語独自の特徴をあきらかにするには他のヨーロッパの言語と比較対照するという方法も考えられる。

今回は「言いかえ」と表現と表現との間の間隔との相関関係が存在するのか，という問題について取り扱った。「言いかえ」は一度言及された対象を違った表現によって指し示すということであり，役割としては代名詞などの照応詞と共通した部分を持つ。そのため，「言いかえ」は代名詞などの照応詞の指す対象の特定という問題にも関係してくる。このようなトピックを扱うには，今回の分析で採用した統計的手法やテクスト処理手法は貧弱である。そのため，自然言語処理の技術，とりわけ機械学習の手法を使うことが考えられる。

Проблема перифразы в газетно-журнальном стиле русского языка

СЭРИ Акинори

Главная цель данной работы — показать существование в русском языке тенденции к перифразе в газетно-журнальном стиле и объяснить элементы, которые связаны с появлением перифраз. В публицистическом стиле русского языка, особенно в газетно-журнальном стиле, редко повторяется тот же самый оборот для указания на одного и того же персонажа. Вместо этого название предмета заменяется описанием его существенных признаков или указанием на его характерные черты. В стилистике такой оборот называется «перифраза».

В отношении этого явления может быть поставлен следующий вопрос: действительно ли существует такая тенденция в газетно-журнальном стиле? Явление перифразы противоречит лингвистическому принципу языковой экономии. При установлении этой тенденции можно поставить следующий вопрос: какие элементы в тексте связаны с этим феноменом? На этот вопрос выдвигается следующая гипотеза: качества слов и предложений имеют тесную связь с появлением перифраз. Для разрешения данных вопросов произведен анализ на основе фактов.

Объекты анализа ограничиваются словами и фразами, которые обозначают понятия «президент» и «премьер-министр». Понятие «президент» обозначается в тексте следующими названиями: «глава государства», «российский лидер», а премьер-министр - «глава правительства». Кроме того, для указания на эти объекты употребляются собственные имена. В данной работе объекты анализа разделяются на три категории: официальные названия со словом «глава», названия без этого слова и имена собственные. Анализируемые тексты собраны из Интернета.

Существование тенденции к перифразе является предварительным условием следующего вопроса. Для того чтобы разрешить второй вопрос, доказано существование связи между появлением перифраз и расстоянием между выражениями — словами и предложениями, которые существуют между двумя указанными объектами.

При анализе использован язык программирования. Сначала кириллица в тексте транскрибирована латинским алфавитом, потом все предложения преобразованы в массивы типа последовательности символов. После этих обработок произведен анализ и получена нужная информация.

В данной работе употребляется статистический метод — проверка статистических гипотез. Этот метод эффективно работает при предположительном анализе естественных языков, у которых есть генеральная совокупность с бесчисленным количеством. При этом надо иметь в виду погрешности (различие между генеральной совокупностью и выборкой — всеми объектами и множеством случаев). Элемент погрешности в контексте статистики выражается с помощью *P*-значения — вероятностного значения, которое решает статистическую значимость данного результата. При этом статистическом анализе использован язык программирования R, чтобы упростить сложную обработку.