2H5-6

# 言語表現と統計グラフの相互変換に関する基礎検討

Primary Study of Mutual Transformation of Linguistic Expression and Statistical Chart

小泉 尚之[*1] 松下 光範[*2] 松田 昌史[*2] 馬野 元秀[*1]
Naoyuki Koizumi Mitsunori Matsushita Masafumi Matsuda Motohide Umano

[*1]大阪府立大学 大学院理学系研究科 情報数理科学専攻
Department of Mathematics and Information Sciences
Graduate School of Science, Osaka Prefecture University

[*2]日本電信電話株式会社 NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories, NTT Corp.

We have many kinds of data of time series such as stock prices and daily temperature, which are understood via their linguistic expressions. We, therefore, proposed a method to express a global trend and local features and oscillation of time series in a linguistic expression. This method, however, is based on some certain utterance patterns and a user-defined vocabulary. In this paper, we try to show the utterance patterns are used in linguistic expressions for graphs. We conduct an experiment that participants transform some typical type of artificial graphs into linguistic expressions. We classify utterance patterns into five groups. Next we extract a vocabulary for each group and consider relations between utterance patterns and graphical shapes or oscillation levels. Their result will lead to an approach to acquire rules for generating linguistic expressions.

## 1. はじめに

近年、我々は多くの情報を簡単に手に入れることができるようになった。しかし、得られた情報は文章、図表、画像、音声など、様々なメディアで表現され、その構成や形式も様々である。そのため、これらの雑多な情報を利用する際に、利用者には様々なスキルが要求され、スキルの習熟の差による情報格差が生じている。そこで、様々な形式で表現された情報を利用者の要求に適した表現に変換して提供する技術が求められている[1]。

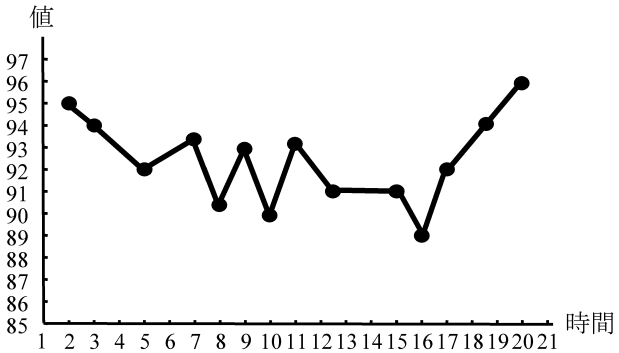
そのような技術のひとつとして、我々は言語表現と統計グラフの相互変換についての研究を進めている。このような技術の実現には、グラフを表現する語彙や着目点の選択基準、発話パターンなど、文章とグラフの対応関係を明らかにする必要がある。本稿では被験者実験を通じて、グラフを言葉で表現する際の発話パターンと使用される語彙についての分析を行う。さらに、グラフの形状と発話パターンとの対応関係から導かれる仮説についての考察を行い、その結果に基づいて先行研究の考察を行う。

## 2. 先行研究との関連

我々は、一般に人がグラフを言葉で表現する際に、最もグラフの理解に貢献している要素から順に表現する、という仮説に基づき、時系列データを全体的傾向、局所的特徴、振動の特徴の 3 つの要素を組み合わせて言葉で表現する方法を提案した[2]。その方法の概略と出力結果（図 1）を以下に示す。

**(1)** 時系列データを前期、中期、後期の 3 つのファジィ的な期間に分割し代表値を求め、前期から中期、中期から後期の値の変化量を言葉とその一致度で表現し、それらを組み合わせて全体的傾向を言葉とその一致度で表現する。

連絡先: 小泉 尚之，大阪府立大学 大学院理学系研究科 情報数理科学専攻，〒599–8531 大阪府堺市中区学園町 1 – 1，TEL: 072–254–9675，FAX: 072–254–9930 E-mail: koizumi@marron.cias.osakafu-u.ac.jp




全体的にはやや下に凸で、後期に中ぐらい増加している部分と少し減少している部分があり、中期に中ぐらい振動している。

図 1: 時系列データと言葉への変換結果の例

**(2)** 各期間の代表値からファジィ推論によって全体的傾向を表す時系列データを生成し、元のデータとの違いが顕著である部分を局所的特徴と考え、その差を言葉とその一致度で表現する。

**(3)** 全体的傾向を表す時系列データと元のデータの標準偏差の差と振動回数を組み合わせて各期間の振動の特徴を言葉とその一致度で表現する。

**(4)** これら 3 つの要素を組み合わせて時系列データを言葉とその一致度で表現する。

ここで、最もグラフの理解に貢献している要素が全体的傾向であり、次いで局所的特徴、振動傾向である、という我々の提案と実際に人がグラフを言葉で表現した文章とを比較・考察を行うと共に、わかりやすい表現で出力するためにこれら 3 つの要素の組合せ方の一般的な規則についての考察を行う必要があると考えられる。

## 3. 被験者実験

本実験は、グラフを言葉で表現する際に用いられる語彙や発話パターンの分析のために、グラフを説明した文章の収集を目的としている。また、分野に依存しない一般的な語彙や発話パターンの分析のために、刺激となるグラフは実際の統計グラフではなく、計算機で作成した人工的なグラフとした。実験では、グラフを文章で説明させる「説明課題」とその文章から元のグラフを推測させる「解読課題」を実施した。

### 3.1 説明課題

各参加者に対し、81 種類のグラフからランダムに 6 種類を選び、記入用紙にそれぞれ文章で説明させた。各グラフと記入用紙にはあらかじめ共通の問題番号が印字されており、グラフと文章が一意に対応付けできる。各参加者に対し、6 種類のグラフと対応する記入用紙の 6 組を同時に配付した。

回答時間は合計で 40 分とし、1 問当たりの回答時間は統制しなかった。また、回答順序に制限はなく、時間の許す限り前の問題に戻って文章の追加・変更を自由に行えるようにした。ただし、文章の収集を目的としているため、説明する際には図や記号の使用を禁止した。なお、回答はすべて筆記で行い、筆記具は実験者が用意した。

### 3.2 解読課題

説明課題で得られた記入用紙を説明課題後の休憩時間を利用して 2 部コピーしておいた。各参加者に対し、説明課題で得られた記入用紙からランダムに 12 種類を選び、選択用紙から元のグラフを推測させた。ただし、選ばれた文章には推測を行う参加者自身が書いた文章が含まれないように操作した。説明課題と同様に、記入用紙と選択用紙にはあらかじめ問題番号が印字されており、文章と選択したグラフが一意に対応付けできるようにした。各参加者に対し、12 種類の文章と対応する選択用紙の 12 組を同時に配付した。

選択用紙には、正解のグラフと共に、8 種類のグラフがランダムに描かれており、合計 9 種類のグラフの中から文章に合致するグラフを選択させた。ここで、選択する際にはポイントの自由分配とし、1 問につき 100 ポイントとして 9 種類のグラフの中から文章に合致すると判断したグラフにポイントを自由に割り振らせた。そして、正しいグラフに割り振ったポイントを得点とし、ポイントを割り振った参加者とその文章を書いた参加者の両者に得点を加算した。

回答時間は合計 25 分とし、1 問当たりの回答時間は統制しなかったが、次の問題に着手した後に、前の問題には戻れないように統制した。

### 3.3 参加者

実験は大阪府立大学の学生 41 名（男性 17 名、女性 24 名）を 3 つの集団に分けて行った。3 つの集団はそれぞれ別々の日に実験を行い、1 日目は 14 名（男性 6 名、女性 8 名）、2 日目は 14 名（男性 5 名、女性 9 名）、3 日目は 13 名（男性 6 名、女性 7 名）であった。参加者の平均年齢は 19.2（最小 18、最大 24）歳であり、在籍学部、文・理系について調べた。日本語に関する実験であるため、参加者は日本人に限定した。また、参加者への実験報酬と課題の成績を連動させた。

### 3.4 実験刺激

実験には計算機で作成した 81 種類のグラフを実験刺激として用いた。一辺の長さが 300 の正方形を縦横に 3 等分、全体として 9 等分した後、各列から 1 マス選択し、選ばれた 3 マスの中心点をつなぎ、スプライン関数を用いて平滑化した（図 2 上）。この 27 種類のグラフに対し、ランダムに上下に最大 20（6.7%）の振動を加えたもの（図 2 左下）と、ランダムに上下に最大 50（16.7%）の振動を加えたもの（図 2 右下）を合わせた 81 種類のグラフを実験刺激とした。

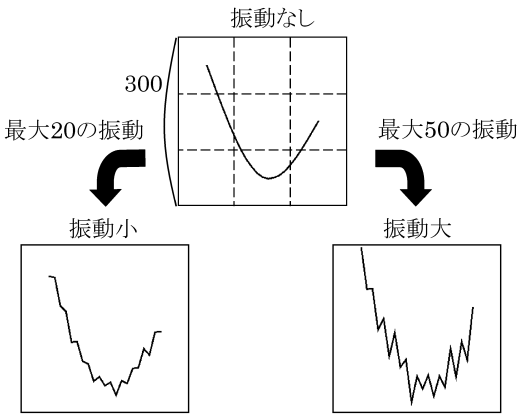



図 2: 参加者に提示するグラフの例
ただし、実際の実験で使用したグラフには罫線は引かれていない

### 3.5 匿名性の保持と一貫性の保証

実験では参加者に自由な表現を促すために匿名性を保持した。匿名性を保持すると同時に問題の一貫性を保つために、参加者には実験開始前に ID カードを配付した。ID が書かれた封筒を用意し、この封筒に問題用紙を入れて配付・回収することで問題用紙からは ID を特定することはできず、匿名性が保持されている。また、各課題では ID によって回答する問題番号をあらかじめ一意に決めておくことで問題の一貫性を保持した。

### 3.6 手続き

参加者に各課題について解説した後、説明課題を 40 分で行った。20 分の休憩を挟み、もう一度解読課題の解説をした後、解読課題を 25 分で行い、最後に実験に関する事後アンケートを行った。また、実験中は質問以外の発言を禁止した。

## 4. 結果と考察

実験によって得られた 246 文章について、発話パターンと語彙の分析、発話パターンとグラフ形状との関係の考察を行う。

### 4.1 実験結果

説明課題の結果、246 文章の総文字数は 30577 字であり、1 文章の平均文字数は 124.3 字であった。また、形態素解析エンジン Chasen[3] を用いて単語に分解した結果、単語は 989 種類で総単語数は 20736 個であり、1 文章の平均単語数は 84.3 語であった。

解読課題の結果、100 ポイントすべてを正解に割り振られていなかった回答は 19 個であり、それらの文章の平均文字数は 95.5 字、単語は 267 種類、平均単語数は 65.7 語であった。

### 4.2 発話パターン

得られた文章を吟味すると、まず「全体的な形に言及している」文章と「左から右に向かって順に言及している」文章に分類できた。次に、この両方を言及している文章については、言及する順番が重要であると考えて分類を行った。また、主にグラフの特徴的な一部分のみしか言及していない文章もあった。

表 1: 発話パターンの分類結果

| 発話パターン | 文章数 | 割合 (%) | 平均文字数 | 平均単語数 |
|---|---|---|---|---|
| パターン 1 | 127 | 51.63 | 115.5 | 78.4 |
| パターン 2 | 58 | 23.58 | 129.4 | 86.5 |
| パターン 3 | 22 | 8.94 | 132.6 | 90.2 |
| パターン 4 | 17 | 6.91 | 146.8 | 98.8 |
| パターン 5 | 22 | 8.94 | 129.1 | 95.3 |

表 2: 各パターンの語彙と出現回数（一部抜粋）

| パターン 1 | 回数 |
|---|---|
| グラフ | 183 |
| ギザギザ | 69 |
| 上 | 65 |
| 点 | 60 |
| 山 | 59 |
| 方 | 57 |
| よう | 52 |
| 右 | 49 |
| 左 | 49 |
| 左端 | 46 |
| 下 | 44 |
| 右端 | 44 |
| 谷 | 42 |
| 形 | 39 |

| パターン 2 | 回数 |
|---|---|
| グラフ | 82 |
| 山 | 51 |
| 方 | 38 |
| 右 | 33 |
| 左 | 31 |
| 上 | 29 |
| ギザギザ | 28 |
| 線 | 22 |
| 下 | 21 |
| 谷 | 22 |
| 左端 | 18 |
| 位置 | 17 |
| 右端 | 16 |
| 部分 | 16 |

この結果から、得られた文章を人手により以下の 5 種類の発話パターンに分類した（表 1）。

**パターン 1** 全体的な形について言及した後にいくつかの特徴的な部分について言及する

**パターン 2** 左から右に向かって順に特徴的な部分について言及する

**パターン 3** 全体的な形について言及した後に左から右に向かって順に特徴的な部分についても言及する

**パターン 4** 左から右に向かって順に特徴的な部分について言及した後に全体的な形についても言及する

**パターン 5** 主にいくつかの特徴的な部分について言及する

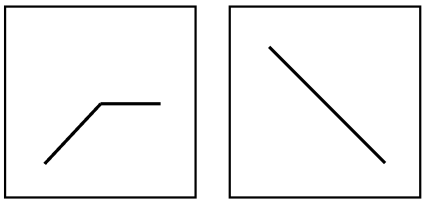
### (1) 語彙

各パターンごとに名詞（複合名詞）と動詞（複合動詞）を抽出した。ただし、パターン 3、パターン 4、パターン 5 はパターン 1 とパターン 2 のサブセットであるため、ここではパターン 1 とパターン 2 についての結果を示す（表 2）。

### (2) グラフ形状

各発話パターンとグラフの形状との関係を調べた。実験で用いたグラフはその作り方から 9 種類の形状に分類することができる。

グラフは縦横に 3 等分されたマスの各列から 1 マスずつ選択していたので、グラフを前半と後半に分けて見ると、それぞれ「一定」、「増加」、「減少」の 3 種類の傾向があることがわかる。そのため、これらを組み合わせた 9 種類をグラフ形状とし、例えば、前半が「増加」で後半が「一定」の場合は「増加一定」とラベル付けを行う（図 3 左）。ただし、前半と後半が同じ傾向の場合にはその傾向をラベルとする（図 3 右）。



増加一定　　減少

図 3: グラフ形状の例

表 3: グラフ形状ごとの発話パターンの出現割合 (%)
ここでは各パターンを P1–P5 と表記している

| 形状 | 振動 | P1 | P2 | P3 | P4 | P5 |
|---|---|---|---|---|---|---|
| 一定 | なし | 100 | 0 | 0 | 0 | 0 |
| | 小 | 71.42 | 14.29 | 0 | 0 | 14.29 |
| | 大 | 28.57 | 0 | 0 | 0 | 71.43 |
| | **計** | **66.67** | **4.76** | **0** | **0** | **28.57** |
| 増加 | なし | 100 | 0 | 0 | 0 | 0 |
| | 小 | 66.67 | 33.33 | 0 | 0 | 0 |
| | 大 | 50 | 25 | 0 | 0 | 25 |
| | **計** | **70** | **20** | **0** | **0** | **10** |
| 減少 | なし | 100 | 0 | 0 | 0 | 0 |
| | 小 | 33.33 | 66.67 | 0 | 0 | 0 |
| | 大 | 66.67 | 33.33 | 0 | 0 | 0 |
| | **計** | **66.67** | **33.33** | **0** | **0** | **0** |
| 一定増加 | なし | 33.33 | 16.67 | 16.67 | 33.33 | 0 |
| | 小 | 37.5 | 50 | 0 | 0 | 12.5 |
| | 大 | 14.29 | 71.42 | 0 | 0 | 14.29 |
| | **計** | **28.57** | **47.63** | **4.76** | **9.52** | **9.52** |
| 一定減少 | なし | 44.45 | 33.33 | 22.22 | 0 | 0 |
| | 小 | 25 | 75 | 0 | 0 | 0 |
| | 大 | 20 | 60 | 0 | 20 | 0 |
| | **計** | **31.82** | **54.55** | **9.09** | **4.54** | **0** |
| 増加一定 | なし | 44.44 | 44.44 | 11.12 | 0 | 0 |
| | 小 | 11.11 | 55.56 | 0 | 11.11 | 22.22 |
| | 大 | 66.67 | 0 | 0 | 0 | 33.33 |
| | **計** | **37.5** | **37.5** | **4.17** | **4.17** | **16.66** |
| 減少一定 | なし | 50 | 16.67 | 16.67 | 16.66 | 0 |
| | 小 | 11.11 | 88.89 | 0 | 0 | 0 |
| | 大 | 44.45 | 0 | 11.11 | 11.11 | 33.33 |
| | **計** | **33.34** | **37.5** | **8.33** | **8.33** | **12.5** |
| 増加減少 | なし | 61.54 | 0 | 30.77 | 7.69 | 0 |
| | 小 | 66.67 | 0 | 0 | 25 | 8.33 |
| | 大 | 30 | 30 | 10 | 10 | 20 |
| | **計** | **54.28** | **8.57** | **14.29** | **14.29** | **8.57** |
| 減少増加 | なし | 71.43 | 0 | 28.57 | 0 | 0 |
| | 小 | 58.82 | 15.69 | 9.81 | 11.76 | 3.92 |
| | 大 | 73.33 | 6.67 | 13.33 | 0 | 6.67 |
| | **計** | **63.75** | **11.25** | **13.75** | **7.5** | **3.75** |
| **総計** | | **51.63** | **23.58** | **8.94** | **6.91** | **8.94** |

この 9 種類のグラフ形状にそれぞれ振動なし、振動小、振動

表 4: カテゴリと発話パターンの出現割合 (%)

| カテゴリ | 振動 | P1 | P2 | P3 | P4 | P5 |
|---|---|---|---|---|---|---|
| 一定増加減少 | なし | 100 | 0 | 0 | 0 | 0 |
| | 小 | 61.54 | 30.77 | 0 | 0 | 7.69 |
| | 大 | 42.86 | 14.28 | 0 | 0 | 42.86 |
| | **計** | **67.5** | **15** | **0** | **0** | **17.5** |
| 一定増加 一定減少 増加一定 減少一定 | なし | 43.33 | 30 | 16.67 | 10 | 0 |
| | 小 | 20.59 | 67.65 | 0 | 2.94 | 8.82 |
| | 大 | 37.04 | 29.63 | 3.7 | 7.41 | 22.22 |
| | **計** | **32.97** | **43.96** | **6.59** | **6.59** | **9.89** |
| 増加減少 減少増加 | なし | 66.67 | 0 | 29.63 | 3.7 | 0 |
| | 小 | 60.32 | 12.7 | 7.94 | 14.29 | 4.76 |
| | 大 | 56 | 16 | 12 | 4 | 12 |
| | **計** | **60.87** | **10.43** | **13.91** | **9.57** | **5.22** |

表 5: 一定を含むグラフ形状の出現割合 (%)

| カテゴリ | 振動 | P1 | P2 | P3 | P4 | P5 |
|---|---|---|---|---|---|---|
| 一定増加 一定減少 | なし | 40 | 26.67 | 20 | 13.33 | 0 |
| | 小 | 31.25 | 62.5 | 0 | 0 | 6.25 |
| | 大 | 16.67 | 66.67 | 0 | 8.33 | 8.33 |
| | **計** | **30.23** | **51.16** | **6.98** | **6.89** | **4.65** |
| 増加一定 減少一定 | なし | 46.67 | 33.33 | 13.33 | 6.67 | 0 |
| | 小 | 11.11 | 72.22 | 0 | 5.56 | 11.11 |
| | 大 | 53.33 | 0 | 6.67 | 6.67 | 33.33 |
| | **計** | **35.42** | **37.5** | **6.25** | **6.25** | **14.58** |

大を考慮して各形状ごとに発話パターンの出現割合を表 3 に示す。

### 4.3 グラフ形状と発話パターンの考察

これらの結果を踏まえ、人がグラフを言葉で表現する際の振る舞いについての考察を行う。

**(1) 着目点**

一般に、人がグラフを言葉で表現する際に最もグラフの理解に貢献している要素から順に表現する、と仮定すると、発話パターンから半分以上の文章では最初に「全体的な形」に着目していると言える。すなわち、「全体的な形」がグラフを言語化する際に最も重要であると考えられる。

**(2) 形状の変化**

表 3 より、「一定・増加・減少」、「一定増加・一定減少」・「増加一定・減少一定」、増加減少・減少増加のグラフ形状と発話パターンがそれぞれ似た傾向を示していることがわかる。これらをカテゴリとしてまとめ、その結果を表 4 に示す。

「一定・増加・減少」カテゴリはパターン 1 が多く、「一定増加・一定減少・増加一定・減少一定」カテゴリはパターン 2 が多い。このことから、途中で形状の傾向が変化するグラフはパターン 1 が多く、変化しないグラフはパターン 2 が多い、と考えられる。しかし、「増加減少・減少増加」カテゴリはこのルールには当てはまらない。ここで、表 2 において“ 山 ”や“ 谷 ”というグラフとは直接関係ないが、形状を表していると思われる単語が多く得られていることから、グラフの全体的な形を一般的に相手に通じる形に見立てることができる場合は全体的な形で表現する、すなわち「パターン 1」が多くなると考えられる。

**(3)「一定」を含む形状の変化**

「一定増加・一定減少」カテゴリと「増加一定・減少一定」カテゴリを比較すると、振動なしと振動小はよく似ているが、振動大は大きく異なっていた (表 5)。「増加一定・減少一定」カテゴリの振動大はパターン 1 とパターン 5 が多いことから、表 4 の「一定・増加・減少」カテゴリに似ている、すなわち、形状の変化がないグラフとして表現されていると言える。このことから、振動が大きく形状が「一定」に変化する場合には、形状の変化が正しく認識されていない可能性があると考えられる。

また、表 3 より、一定増加と一定減少、増加一定と減少一定がよく似ていると言える。しかし、一定増加と減少増加や増加一定と増加減少などの「一定」以外の形状が共通の場合は似ているとは言えない。このことから、形状が「一定」から、または「一定」に変化する場合には「増加」や「減少」という形状の変化内容が発話パターンに与える影響は小さいと考えられる。

**(4) 振動の変化**

表 3 より、パターン 5 は振動なしでは現れず、振動が大きくなるにつれて出現率が増える傾向にある。すなわち、振動が大きくなるにつれて、グラフを網羅的に説明するのではなく、特徴的な一部分のみを説明する文章が増えていると言える。このことから、振動が大きくなると言及対象が少なくてもグラフを説明できるようになると考えられる。

### 4.4 先行研究の評価

これらの仮説を踏まえると、先行研究において、「全体的傾向」グラフの言語化に最も重要な要素である、という我々の提案はある程度人の振る舞いに近い考え方であったと言える。しかし、形状の変化に「一定」が含まれている場合や振動の表現については更なる検討の余地があると言える。

## 5. おわりに

本稿では、グラフを表現する語彙や着目点の選択基準、発話パターンなど、文章とグラフの対応関係を明らかにするために、被験者実験を通じてグラフを言葉で表現する際の発話パターンと使用される語彙、グラフの形状と発話パターンとの対応関係から導かれる仮説についての考察を行った。

今後は、仮説の検証や新たな仮説を検討することにより、人がグラフを言語化する際の一般的な規則を明らかにし、先行研究の改良を行うだけではなく、文章からグラフを生成するための手法についても検討していきたい。

なお、本研究は NTT コミュニケーション科学基礎研究所におけるインターンシップで行われた研究を発展させたものである。

## 参考文献

[1] 加藤 恒昭、松下 光範: 情報編纂（Information Compilation）の基盤技術, 第 20 回人工知能学会全国大会, 1D3-2 (2006).

[2] 馬野 元秀、小泉 尚之、岡村 光洋: 全体的傾向と局所的特徴に基づく時系列データの言葉による表現 –標準偏差による振動の表現–、第 2 回 MuST 成果・進捗報告会 (2007).

[3] http://chasen-legacy.sourceforge.jp/