Program Integrated Informationの統計的特徴は何に対する特徴か

加藤直孝[†] 有澤誠[††]

プログラムが人間と意思疎通するために出力する文字列を Program Integrated Information（PII）文字列とよぶ．本稿は，実験で PII 文字列の単語数には，どのような統計的な特徴があるかを調べた．そして，Brown Corpus の文の単語数と比較した．その結果，単語数が log-linear の目盛りで直線に回帰するパターンは，文中の単語に限ったものではなく，文にはならない GUI 上の単語も含んだ，より普遍的なパターンであることがわかった．PII 文字列の単語数は，2 単語が最も多く，単語数と PII の数のグラフは，50 単語以下で log-linear の目盛りで直線に回帰する．一方，Brown Corpus の文の単語数は 12 単語が最も多く，12 単語以上のみで，log-linear 目盛りで直線に回帰する．PII 文字列は，PII 文字列間の「つながり」だけではなく，PII 文字列と GUI 上のオブジェクトとの「つながり」を持つ．そのため，文よりも少ない単語数で，プログラムと人間の意思疎通を可能にできる．

Naotaka Kato[†] and Makoto Arisawa[††]

Programs use Program Integrated Information(PII) as an output string to communicate with their users. This paper examined many sets of PII strings to get the statistical characteristics of them. We counted the number of words in each PII string and compared the result with the number of words in each sentence in Brown Corpus. As a result, the log-linear regression pattern of the number of words in each sentence is not limited to the complete sentences, but also applicable to the incomplete sentences. The log-linear pattern is a universal pattern for the number of words in multimedia objects.

1. はじめに

プログラムが人間と意思疎通するために GUI（Graphical User Interface）上に出力する文字列を PII（Program Integrated Information）文字列とよぶ．本稿は PII 文字列の単語数がどのような統計的特徴を持つかを実験で調べた．PII 文字列の単語数は log-linear の目盛りで直線に回帰する．多くの PII 文字列は文ではない．そこで，PII 文字列の単語数の分布と Brown Corpus の文の単語数の分布を比較した．その結果，単語数が log-linear の目盛りで直線に回帰するパターンは，文中の単語に限ったパターンではなく，PII 文字列の単語も含んだ，文とは独立な普遍的なパターンであることがわかった．PII の統計的特徴は，PII に関する色々なオブジェクト間の「つながり」の特徴を示している．

プログラムの GUI（Graphical User Interface）上の文字列は，プログラム中に文字列として埋め込んでプログラミングを行なうこともできる．しかし，通常はプログラムを国際化しやすいように，この文字列をプログラム本体から分離し，別のテキストリソースファイル上に作成する．この分離したテキストリソースファイルのことを PII テキストリソースファイルとよぶ．略して PII ファイルとよぶ．PII ファイルは key と分離した文字列の対で構成する．PII ファイル中には次の行が繰り返す．

key=文字列

本稿では，左辺を「PII の key」または単に「key」とよび，右辺を「PII 文字列」とよぶ．PII は「key=文字列」全体を指す．プログラム中には key を埋め込むので，出力する言語を切り替えるには，参照する PII ファイルを切り替えるだけでよい．PII の具体例は次のようになる．

keyProfileLocation=Copy profile to

PII 文字列も言語である．単語や句，文や文章である．しかし，上の例のように言語単独では意味を成さないことが多い．言語は意思疎通のための手段であるが，PII 文字列は，通常の文章のように，言語単独で意思疎通の手段にはならない．PII 文字列は，プログ

† 日本 IBM(株) ナショナル・ランゲージ・サポート
Translation Services Center, IBM Japan,Ltd.
†† 慶應義塾大学 政策・メディア研究科
Graduate School of Media and Governance, Keio University

ラムの操作の流れの中で GUI 上に出現するときその役割をはたす.

そこで，PII 集合の特徴を調べるために，OS を含む 124 個のアプリケーションの PII に対して実験を行なった．結果に対して，単語数を横軸にとり，その単語数を持つ PII の個数を縦軸にしてグラフを作成した．その結果，次の特徴を発見した．PII 文字列の単語数はフラクタルの特徴を持つ．PII の大部分（約 99%）は log-linear 分布になり，全体をカバーする分布はパラボリックフラクタル分布になる．約 1 割の key をカバーする部分は線形フラクタル分布になっている．本稿では，PII の単語数が log-linear の目盛りで直線に回帰する点に注目し，Brown Corpus の文の単語数のパターンと比較した．

本稿が対象とする PII 文字列は，言語の世界に閉じておらず，言語以外のメディアである GUI 中に出現する．PII 文字列の「つながり」には，通常の文章の「つながり」に加えて，PII 文字列と GUI 上のオブジェクト（アイコンやチェックマーク等）との「つながり」がある．この GUI 上の「つながり」は PII 文字列の単語数に大きな影響を与えている．

本稿は，第 2 節で関連研究について述べ，第 3 節で実験およびその結果について述べる．第 4 節の考察では，単語同士や単語とオブジェクトの「つながり」が意味を生み出すことを述べる．そして，Brown Corpus の文の単語数の特徴と PII 文字列の単語数の特徴を比較し，考察を行なう．

2. 関 連 研 究

通常の英語のテキストが Zipf の法則に従うことは良く知られている[6]．文章中の英単語の出現頻度は，英単語を出現頻度の多い順に並べると，順位のベキ乗の逆数に比例し，そのベキ乗の値はきわめて 1 に近くなる．Zipf の法則はベキ法則ともよばれており，Adamic は Zipf の法則とベキ法則とパレートの法則が，数学的に同じ内容であることを証明している[1]．

英語の文の単語数の研究では，Marckworth[3] らによる Brown Corpus の研究に基づき，Bengt Sigurd[5] らが研究を行なっている．

3. 実　　験

3.1 実験対象アプリケーション

次の 124 個のアプリケーション（OS を含む）に対し，各 PII の単語数を数えた．

- Eclipse SDK 3.12
- CATIA(R) Version 5 Release15

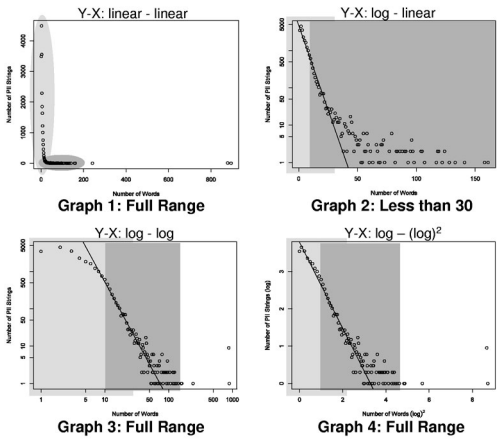

図 1　PII 文字列 - 単語数（Eclipse）
Fig. 1　PII string - Number of Words (Eclipse)

- Microsoft(R) の 122 個のアプリケーションと OS

Eclipse SDK 3.12 は，オープンソースの JAVA(R) プログラム開発環境ツールである．CATIA Version 5 Release 15 は CAD/CAM（Computer Aided Design/Computer Aided Manufacturing）システムプログラムである．実験は，アプリケーションの PII ファイル内にある PII を対象とした．

3.2 実 験 結 果

全ての実験対象アプリケーションは PII の単語数に関し共通の特徴を示した．本稿は，Eclipse を例に実験結果を示す．**図 1** に Eclipse の PII 文字列の単語数とその単語数を持った PII の数の関係を示す．X 軸（横軸）は PII 文字列の単語数で，Y 軸（縦軸）は X 軸の単語数をもった PII 文字列の個数である．該当する単語数の PII 文字列が存在しなければ，プロットしていない．Graph 1 は，linear-linear（Y 軸目盛り-X 軸目盛り）で，残りの 3 つのグラフは，Graph 1 と同じデータを異なる目盛りで表示している．各グラフの上の表示（e.g. Y-X: log-log）は，そのグラフの目盛りを示している．Graph 2 は右側から 3 プロットを除外し，200 単語以下のみを表示している．Graph 2 以外のグラフは全ての単語数の範囲を含んでいる．Graph 1 のそれぞれのグレーレベルの部分が，残りの 3 つのグラフのそれぞれのグレーレベルの部分に対応している．

PII 文字列の単語数の特徴は 3 つの領域に分けることができる．それぞれを，A 領域（薄いグレー），B 領域（濃いグレー），C 領域（全域）とよぶことにすると，1 単語から 30 ないし 50 単語以下が A 領域，10 単語から 100 ないし 200 単語以下が B 領域，全ての単語数が C 領域である．**表 1** に，分布に対しそれぞ

れの回帰式がカバーするPIIの割合，および回帰式を求めるための座標軸の目盛り（スケール）をまとめた．他のアプリケーションも同様の結果になった．

本稿は，Graph 2の結果に注目し，考察を行なう．124個中の3つのアプリケーションに関するlog-linear目盛りの回帰式を**表2**に示す．Eclipseの回帰式は，30単語以下のデータから求めている．WinXPはWindows XP SP2を示す．本稿では，log-linear座標で直線に回帰するパターンをlog-linearのパターンとよぶ．

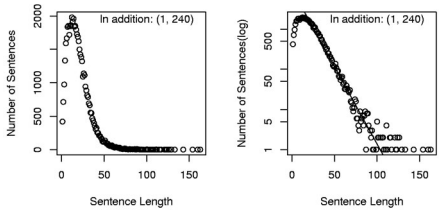

図2 文の単語数分布（Brown Corpus），右図はY軸を対数にした

Fig. 2 Sentence-Length Distribution of the Whole Brown Corpus

4. 考　　察

本節では，最初に「つながり」と意味の関係を述べ，文の「つながり」とPII文字列の「つながり」について説明する．そして，Brown Corpusの文の単語数の特徴とPII文字列の単語数の特徴を比較することで，それらの「つながり」の特徴を述べる．

4.1 「つながり」と意味の関係

本節では,「つながり」が意味と深い関係にあることを指摘する．Nils Nilssonは4)の論理命題の章の中（Section 13.5.1）で論理命題のセマンティックス（意味）について，次のように述べている．

> セマンティックス（semantics）は論理言語の要素集合（elements）を特定の分野の要素集合とつなげる（associate）ことと関係している．そこでのつながりは，我々が意味（meaning）とよぶものである．

「つながり」によって意味は生まれる．言語と何かがつながったとき意味は生まれる．

文の中の単語は何らかの分野の要素集合とつながり，意味が生まれている．句や文や文章も同様である．そして，色々なレベルの言語オブジェクトの間には,「つながり」が存在する．たとえば，単語と単語，単語と句，単語とパラグラフ，文とパラグラフ，パラグラフと文章，単語と文章等々である．文章はそれら全ての「つながり」により意味あるものになっている．

PII文字列の場合は，言語の「つながり」に加えて，GUI上のオブジェクトとの「つながり」が存在する．PII文字列の中には色々なレベルの言語オブジェクトがあり，それら言語オブジェクトとGUI上のオブジェクトとの間には，色々な「つながり」がある．たとえば，単語とアイコン，句とウィンドウ，PII文字列全体と入力域，といった「つながり」である．

表1 PIIの単語数の分布に対するカバレッジと目盛り

Table 1 Scales for the patterns of the number of words in a PII

領域	カバレッジ	Y目盛	X目盛
領域A	約99%（30-50単語以下）	(log)	(linear)
領域B	約10%（10-120単語）	(log)	(log)
領域C	約100%（Tailは無視可能）	(log)	$(\log)^2$

表2 3つのアプリケーションのGraph 2における回帰式

Table 2 Regression line formula for 3 applications

領域	回帰式	決定係数 (R^2)
Eclipse	$\log Y = 3.67 - 0.0900X$	0.979
CATIA	$\log Y = 4.52 - 0.0885X$	0.979
WinXP	$\log Y = 4.19 - 0.0440X$	0.985

4.2 文の単語数とPII文字列の単語数の比較

本節では，文の単語数が持つ特徴とPIIの単語数が持つ特徴を比較することで，それぞれの特徴に新たな解釈を行なう．そして，その解釈は意味を形成する「つながり」に関する重要な事実を含んでいる．

文の単語数は，PII文字列の単語数と共通な特徴と異なる特徴を同時に持つ．Mary Lois MarckworthとLaura M. BellのBrown Corpusの文の単語数（Sentence Length）に関する研究結果[3]を**図2**の左図に示す．横軸が単語数で縦軸がその単語数を持つ文の数である．このグラフは文献2)のTable D11（Page 380-381）の全ジャンルに対する文の単語数のデータから作成した．同文献のGraph D_I（Page 397）と同じ内容のグラフである．文献2)のデータとグラフから，Brown Corpusは，5単語から20単語にそれらの単語数を持つ文が多く集中しており，12単語の文の数が最大になっている．4単語以下では，文の数は激減する．図2の右図は，本稿が縦軸の文の数（Number of Sentences）を対数目盛りにした図である． 図2の右図のように縦軸のピーク値（12単語）より60単語までで回帰線を引くと，回帰式は

$$\log Y = 3.99 - 0.0385X$$

となり，決定係数（R^2）は0.983である．Brown Cor-

pus も 12 単語以上では log-linear のパターンに従っている．12 単語未満では従わない．ここで，PII 文字列の単語数のパターン（図 1 Graph 2）を思い出すと，PII 文字列は 12 単語未満でも log-linear のパターンに従っている．これは，log-linear のパターンが文に限ったパターンではなく，より普遍的なパターンであることを示している．次にそれを説明する．

意味が明確になるということは，上位のオブジェクトに対する下位オブジェクトの位置づけが，明確になるということである．PII 文字列の場合は，1 単語や 2 単語であっても，2 次元の GUI 上での配置により，上位のオブジェクトや同位のオブジェクトに対しての位置づけを明確にしやすい．全体の構造も簡単に理解できる．通常の文では，単語の上位のオブジェクトは文である．文は 1 次元の配置であり，文の上位のオブジェクトは通常はパラグラフである．文の単語数が 1 語や 2 語だと，パラグラフ（上位のオブジェクト）および他の文（同位のオブジェクト）に対する位置づけが困難である．そこで，通常の文の場合は，上位のオブジェクトに対する位置づけを明確にするには，PII 文字列よりも多くの単語数を費やす必要がある．その結果として，縦軸のピーク値を示す単語数は大きくなる．単語と GUI 上のオブジェクトとの「つながり」が PII 文字列の単語数と文の単語数の特徴の差になって現れている．

文の単語数と PII 文字列の単語数は，12 単語以上で共通に log-linear のパターンが存在する（図 2 の右図，図 1 の Graph 2 参照）．PII 文字列は，文にならない 1 語や 2 語といった少ない単語数も含め，12 単語未満でも，log-linear のパターンが成立している．ところが，文の場合は 12 単語未満では，log-linear のパターンに従わない．文では多くの単語を必要とし，PII 文字列では多くの単語を必要としない．PII 文字列の単語数が少なくてもよい理由は，PII 文字列は，少ない単語数のとき，GUI 上に他のオブジェクトとの「つながり」を持つからである．PII 文字列が，12 単語未満でも，log-linear のパターンが成立しているということは，言語のみの「つながり」のパターンと言語以外を含んだ「つながり」のパターンが共通であることを示している．一方，Brown Corpus の文は，文のみにより意思疎通を行なうという制約があるので，12 単語未満では log-linear のパターンに従わない．また，文では，多くの場合 5 単語以上が必要であるということは，「つながり」が一定の意味を形成するには，5 オブジェクトは必要であると解釈できる．それならば，PII 文字列が 1 単語のときは，4 単語（5 単語から 1 単語を引いた単語数）以上に相当するオブジェクトが GUI 上に存在するはずである．

最後に，図 2 の左図からわかるように，ほとんどの英語の文は 50 単語以下である．このことは，PII 文字列の単語数の分布が 50 単語以下で，log-linear の特徴を持つことと符合する．

5. おわりに

文の単語数と PII 文字列の単語数には，共通に log-linear のパターンを持つ．この事実は，言語単独の「つながり」のパターンと言語以外のメディアを含んだ「つながり」のパターンが共通であることを示している．

PII の統計的特徴は PII に関する色々な「つながり」に対する特徴を示している．言語間あるいは言語と他のオブジェクト間の「つながり」が何であるかを突き詰めることは，文脈や背景は何かという問いに等しい．大きな問題である．しかしながら，PII 文字列の単語数の分布と文の単語数の分布のように，ほとんどが同じ「つながり」を持つが，一部のみ異なる種類のつながりを持つサンプルを見つけることができれば，「つながり」の特徴を解明して行く糸口をつかむことができる．今後はこの視点から，言語が持つ「つながり」の研究を行なう予定である．

参考文献

1) Adamic, L.A.: *Zipf, Power-laws, and Pareto – a ranking tutorial*, http://www.hpl.hp.com/research/idl/papers/ranking/ranking.html (2000).
2) Kucera, H. and Francis, W. N.: *Computational Analysis of Present-Day American English*, Bronw University Press (1967).
3) Marckworth, M.L. and Bell, L.M.: Sentence Length Distribution in the Corpus, *Computational Anaysis of Present-Day American English*, Vol.58, No.1, pp.368–405 (1967).
4) Nilsson, N. J.: *Artificial Intelligence : a new synthesis*, Morgan Kaufmann Publishers (1998).
5) Sigurd, B., Eeg-Olofsson, M. and vande Weijer, J.: WORD LENGTH, SENTENCE LENGTH AND FREQUENCY - ZIPF REVISITED, *Studia Linguistica*, Vol.58, No.1, pp.37–52 (2004).
6) Zipf, G.K.: *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*, Hafner Publishing Company Inc., New York, (Reprinted version, Originally published 1949) (1965).