様式Ｆ－１９

# 科学研究費助成事業（学術研究助成基金助成金）研究成果報告書

平成 25 年 6 月 5 日現在

機関番号：１２６１３
研究種目：挑戦的萌芽研究
研究期間：2011～2012
課題番号：２３６５００４５
研究課題名（和文）　電子文書におけるスタイルの抽出・計量及び検索に関する基礎的研究
研究課題名（英文）　Foundational study in extracting, measuring and searching presentation styles and their patterns in digital documents

研究代表者
武村　知子（TAKEMURA TOMOKO）
一橋大学・大学院言語社会研究科・教授
研究者番号：60323896

研究成果の概要（和文）：本研究では、電子文書における表示のスタイルとそれを支持するマークアップ言語の相関性に着目し、情報伝達特性に関する分野横断的な研究のための基礎資料となりうる有用なデータベースシステム「スタイルコーパス」の設計・構築を試み、任意のウェブサイトからスタイルに関するデータを抽出して構造特性を可視化するデータ収集解析プログラム及び検索プログラムを試作した。データサイズが小さいため今後検討すべき多くの課題が残ったが、相関解析ツールとしてのスタイルコーパスの有用性は一定程度実証された。

研究成果の概要（英文）：A "Style Corpus" database system,which serves as infrastructure for interdisciplinary researches on various aspects of document-based communication, was constructed on the basis of a correlation analysis between visual presentation styles of digital documents and underlying markup codes. Based on a set of style data extracted from manually-selected web resources, algorithms for searching, analyzing and visualizing its underlying patterns were developed and implemented. Although, due to the limited size of the data, some areas still remain unexplored, the Style Corpus is proved to be effective as a tool for the correlation analysis.

交付決定額

（金額単位：円）

| | 直接経費 | 間接経費 | 合　計 |
|---|---|---|---|
| 交付決定額 | 2,800,000 | 840,000 | 3,640,000 |

研究分野：　文学
科研費の分科・細目：情報学　・　図書館情報学・人文社会情報学
キーワード：コーパス、スタイル、ウェブサイト、電子文書

## １．研究開始当初の背景

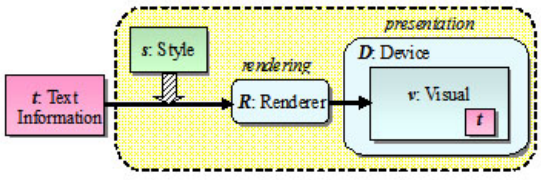
(1) テクストの「スタイル」とそれがもたらす「可読性」とは何か、書記言語における伝達特性とは何か、という問いをめぐる探究は、ルネサンス以来長らく組版技術者・タイポグラファ・造本家が担ってきた実働的領域のそれとしての長く深い伝統を持つ。この伝統に密接に関与しながら人文学全般は進展してきたのであるが、近年、テクストの表示・流通に関する実働・実践領域と人文学とは刻々と乖離しつつある。デザインないしスタイルは決して単なるパッケージングではなく、それ自体がテクストの伝達特性を形成する基体であるという認識は一定程度膾炙してきたと思われるにも関わらず、昨今の電子文書（電子書籍を含む）に関する人文学的議論はおおむね、コンテンツの入力・入手・読解のレベル、すなわちインターフェイス・レベルのそれに留まり、コンテンツの伝達特性の存立を支える深層構造の特性が表層の変化へどのように影響を与えるかという最も重要な議論はなおざりにされてきた。検索・リン

ク等の動的機能をも含めた電子テクストの「スタイル」を存立可能にしているのはHTML,CSS,スクリプト等のプログラミング言語類であるという自明なはずの事実に基づいた実際的かつ思想的な議論に人文学的アプローチをもって参与することが困難なのは、分野横断的な議論の基盤としての共有基本資料が目下欠けているからではないかと思われた。

(2) Google をはじめとする検索エンジンは情報の本体としてのコンテンツを網羅せんとするに至っているが、デザイン／スタイル要素を対象とする検索エンジンは存在しない。「誰もがウェブで発信できる」と謳われる時代でありながら、求めるスタイルとプログラムパターンの具体的な相関性、多様性にはわずかな専門家しかアクセスできず、大枠のところ以上に詳細な情報はウェブ上・紙媒体上に断片的に散在しているにすぎないため、個々人は手さぐりで参考資料を探す他はない。勢いテンプレートへの依存が強まり、ウェブテクストが本来持っているはずの多様な可能性が活かされず、返ってスタイルの画一化へ向かうのは、惜しまれるべきことに思われた。

２．研究の目的

(1) ウェブ・紙を問わず、認知言語学等の分野におけるテクスト認知の研究はもっぱら、デバイス上に表示されたテクスト情報をユーザーが認知するインターフェイスの局面に関して行われる。本研究が対象とするのは、プログラム言語上のマークアップによりテクストにスタイルが適用され（rendering）、そのスタイルにおいて画面上に表示される（presentation）までの処理過程において生じるものごとの部分である（下図）。




本研究の第一の目的は、この、人文学領野では一般にブラックボックスとして看過されている部分に着目し、電子デバイスにおける可読性すなわち言語の伝達特性に関する、分野を問わない研究に資するべき共通基礎資料を作成する方途を発見することである。ウェブ上の言語資源に基づく自然言語コーパスの作成と関連研究は盛んだが、これらの研究はあくまでもテクスト情報が対象であり、HTML タグや CSS 等の人工マークアップ言語は、テクスト情報の抽出に役立つ限られたものを除き、基本的に収集の対象外とされる。本研究では、コンテンツが伝達特性を発揮する過程でいわゆる情報組織化の役割を担うこれらプログラミング言語の発現パターンを素材として、電子文書におけるスタイル・コーパス「KATAGAMI」の作成のためのメソッドの開発に着手する。さらに、コンテンツではなくスタイルの特性をキーとして検索可能なシステムを構築する手段を検討し、実装・実験を行う。

(2) 上記作業と並行して、テクストとその基体の関係に関して数百年来蓄積されてきた人文学的認識をマークアップ言語に対して応用することを試みる。

３．研究の方法

(1) 試験的スタイルコーパス及び検索システムの構築

①素材選定作業

ウェブ上で公開されている一般のページから、なるべく異なるスタイル上の特徴を持つドキュメントを 200 件程度採集する。

②スタイルデータの抽出

採集されたサンプル・ドキュメントの HTML 及び CSS からスタイル要素を抽出し、構造表現に変換する（コンテンツはデータに含めない）。データ表現としては、CSS ボックスモデルにおけるボックスを基本単位（ノード）と見なし、その CSS 属性をブラウザエンジンからダイナミックに取得する。ノード間の関係の表現には、HTML DOM ツリーから得られる親子関係を用いる。このとき、HTML タグもノードのスタイルの一つとして取り込む。

　この作業は専用ソフトウェアを設計・実装して行う。

③コーパスの構築

得られたスタイル構造表現をデータベース化し、PC 上で可視化する方法について検討する。このとき、スタイルの研究者がドキュメントの内容と切り離してスタイルそのものを認識しやすい「パターンカタログ」となることを目的として設計する。

④検索システムの構築実験

上で構築したデータベースに対する検索方法を、有効な検索キー及び使いやすいインターフェイスの観点から選定・実装・試験する。コンテンツコーパスが一般には文字列のパターンマッチングの手法で検索を行うのに対し、スタイルコーパスはスタイルが持つ様々な属性の組み合わせによりパターンマッチングを行うことになる。

(2) 人文学的検討

紙媒体基盤における各種テクスト論を整理検討すると共に、HTML5 の策定過程などの

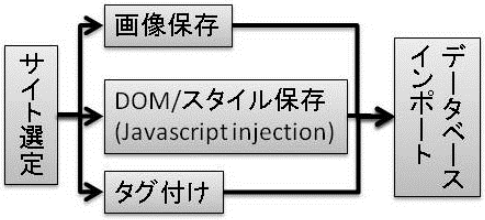
把握を介して、マークアップ言語におけるセマンティック及び非セマンティック要素に対する有効な人文学的考察メソッドを模索する。

４．研究成果

(1) データ収集プログラムの開発

HTML/CSS をそのまま収集するのではなく、実際にブラウザエンジンによって解釈・表示された状態から CSS タグを収集するため、新たに Javascript で HTML DOM 解析プログラムを作成した。

データ収集のフローを下図に示す。




データベースへのインポートの段階は完全に自動処理であるが、DOM/スタイル保存の部分で一部手作業が入るため自動収集の足かせとなっている。ただ、人手が入ることにより、ページの目視による確認が可能となることや、スクリプトしか含まないような、収集の対象とすべきでないページを同時にチェックできるため、一概に非効率とは言えない面がある。

なお、今回はあくまでも試作であり、コーパスそのものを公開する予定がないため、著作権処理は行わなかった。

(2) スタイルのパターンマイニングプログラムの開発

収集したデータはツリー構造であるため、既に知られているツリーマイニングのアルゴリズムを基本とし、部分ノードツリーの共起パターンをマイニングできるように拡張したアルゴリズムの実装を C#言語により試行した。パターンの区分には、スタイル値の絶対値を使う方法と相対的な変化に着目する方法の両方の適用を試みたが、スタイルの種類によって有効性が異なること、例えばフォントサイズでは後者が有効であることなどがわかった。

「文書スタイルに対するパターンマイニングは可能か」という設問への明確な形式的解答をこうして得た。ただし、今後、意味のある解釈が可能であるようなパターンを得つつ、大規模なテクストをコーパスとして収集し主に統計学的な観点から有用な言語学的知見を得ることを狙うコーパス言語学に倣って「コーパススタイル学」を確立するには、より大きなデータセットが必須である。例えば複数群からのスタイルデータの自動分類などの作業に今後着手すべきであるが、データの持つスパースネスと多様性を考慮するならばそのためのサンプル数は、常識的に考えても少なくとも今回の数百倍を要するだろう。前述のように、人力で採取することには利点もあるとはいえ限界が大きいので、ぜひとも自動収集プログラムの開発が先行しなければならない。

(3) 検索システム

数量的なアプローチとは別に、スタイルのパターンを検索するに当ってどのようなユーザーインタフェイスが可能かという課題に対しては、グラフィカルなインターフェイスを備えたスタイルコーパス検索システムを Windows アプリケーションとして試験的に構築した。このシステムにおいて検証したのは、(a)使いやすい条件指定の方法、(b)結果の表示方法、の２点である。

(a)に関しては、既存手法としては Tgrep のように特殊な論理式・文法に従った文字列を用いて検索条件パターンを指定する方法があるが、こうした方法では、スタイル属性の指定部分が長くなった時にパターンの構造を直感的に把握できなくなってしまうこと、また、コンピュータ言語に習熟した者を念頭に置いた手法であるため多くの非熟練者にとって利用困難と考えられることから、インタラクティブかつ直感的なインターフェイスの開発を行うことにした。具体的には、画面上で編集可能な複数のノード（箱）とそれを結ぶ関連線によって、ノードとその属性、およびノード間の関連を表現することにした。

(b)については、DOM ツリー構造のどの部分が検索結果として該当したかを色分けして示すとともに、ワイヤフレーム表示により文書のプレビューが可能となるようにした。検索システムの動作画面スナップショットを下図に示す。

マークアップ言語に基づいたスタイルデータマイニング及び構造特性の可視化のメ

ソッド自体は、ウェブページをワイヤーフレーム化して呈示するアプリケーション等をはじめとして、この間に諸所において開発・公開がなされてきている。だが、それらにおいても、ブラウザが現に表示しているウェブページの構造特性を可視化することができるに留まり、視覚的に把握できるスタイルから逆にウェブサイトを（そのソースコードを）検索するシステムがなお知られていないのは、当然ながら、そもそもスタイルはソースコード／CSS に直接書き込まれているものではなく、ブラウザ表示において初めて発現するものであるゆえに、原理上ネット検索が困難だからであって、そのためウェブサイトを作成したい一般ユーザーにデザイン見本を提供する紙媒体カタログや「まとめ」サイトなどが多く出ているのであるが、それらの殆どは、特定の個人がその主観的感性ないしは経験に基づく判断によって任意のサイトを分類抽出したものであり、そうしたカタログを見て「例えばこんなような感じ」のサイトのコードの書き方を知りたいと思えば結局は直接そのサイトに入ってコードを見るしかない。その上これら個別のカタログは流行に応じて日々刷新されてゆき、研究資料とするに充分な蓄積がなされることがない（事実、今回ウェブサイト選定の参考に用いたその種のカタログに記載されていたサイトの多くは、採集時点で早くも消えていた）。大規模データベース構築が前提として必要であるゆえんである。大規模データベースが構築できれば、そこに立脚した検索システムが、ネット検索システムの不在を一定程度補いうると期待されるし、よりインテリジェントな検索システムを発展させる方途を見出すことも可能になるかもしれない。

そこで今後解決すべき問題点としては、まず上述のように⓪数量的解析を可能にするためのデータベースの規模的な拡充、およびその際に問題となる収集とプリプロセスの自動化、が挙げられる。その上で更に⓫検索キーを与える際に、厳密なスタイル値のみならず、あいまい表現・直感的表現・図形情報・動的表現をキーとして用いられるようにし、検索性能を実用レベルまで向上させること、⓬「読みやすさ」など主観的なタグ付けのサポートの３点が主なものとして挙げられる。

⓪,⓫に関して、本研究では、データにおける値の分散が大きいスタイルを対象にしたため、結果的に行幅やフォントサイズなど数値化が容易なスタイルが解析の条件として選択されることになったが、今後、タイプフェイスなど本来数値化できないスタイル値の数量化の規範を確立する必要がある。また、スタイルのパターンにとどまらず、スタイル値の確率分布や相関を解析できるようなデータ処理方法の適用によって、より実用的な検索が可能になるはずである。

なお、ユーザーインターフェイスの観点からみて、スタイル条件の指定が煩雑であることが問題として残されている――つまり一般ユーザーにとって最終的に必要かつ有意義なのは、コードの記述に使われる用語や正確な数値による検索ではなく、あくまでも intuitive に遂行しうる類似検索ないし視覚イメージ検索であるからである。

⓬に関しては、スタイルに実際に評価タグ（「読みやすい」「カラフル」など）をつけて検索システムに投入する具体的な方法の開発自体、もともと本研究の範囲外である。しかしながら文書のプレゼンテーション形式から読み手が受ける様々な印象を個々のスタイル要素に結びつけて説明するモデルを作成することは、本研究がやがてのひとつの到達点として想定する長期的課題である。データの自動収集プログラムの開発と並行して、収集されたデータに読み手の評価タグを追加して蓄積するシステム、つまり任意のタグが編集可能なスタイルコーパスシステムの構築は、すぐにも着手したい課題である。

(4) 上記コーパス及び検索システムの開発目的に鑑み、特に読み手から見た主観的な可読性の処理に関して問題点が多く残ったのは、人文学的考察が予想以上に難航したことが大きな原因のひとつである（これが難航せずにすむための基礎資料の作成こそが本研究の目的なのであるが）。それでも、幾つかの知見が得られた。

①規範と自由に関する古い問題を問い直すべき局面の発見

この２年間で HTML5 統一規格の広範な普及が進み、今や人文学なかんずく文学にとってアクチュアルな考察の対象となすべきは、基本的に紙媒体におけるデザイン規範の継承を専らとするいわゆる「電子書籍」ではなくウェブサイト全般であることは明らかである。一方で、HTML5 仕様書策定が進み規格化が進むにつれ、大量の便利なテンプレートが流通しつつあり、マークアップ言語と表示スタイルの相関性はいやましにブラックボックス化していく徴候も見られる。問題は、そもそも表示スタイルというものは、ソースコードが示すテクスト構造とも、CSS に列挙されてあるスタイル属性とも、実は根本的には何ら連動しないという点にある。HTML5 は(a)テクストの構造特性にスタイル要素が介入することを防ぎつつ、同時に(b)構造特性が適切な統一的基準に基づいてスタイルに反映することを目指して策定されているが、この目的はあくまでも、いわば英語論文におけるパラグラフ・ライティング的な一定の文章様式を暗黙のうちに念頭に置いた上で達成が見込まれている。だが HTML 言語によ

る記述と CSS の記述の組み合わせに関する明瞭な規定はなく、その部分に残されている自由裁量の余地があまりに広いため、穿った見方をすれば、その自由を限定して上記の策定目的を達成するためには、その部分をブラックボックス化してテンプレートを流通させるしかなく、一般ユーザーにとってもそれが最も便利で安全なのだと考えられよう（むろんその便利さ自体が、HTML 言語規格の目指すもののひとつなのであるが）。いわば人々の安心と引き換えに不可視化されてゆくこの部分にこそ、今後の電子テクストの方向性――すなわち人間の書記言語テクスト全般の方向性――を決定するモメントがあることは、今回の研究でよりいっそう明らかになった。検索システム構築の最も困難な部分と、この不可視の部分とが重なるからである。

ウェブが「誰もが自由に発信できる」媒体であるとうたわれるとき、その「自由」なるものは一般に「コンテンツの自由」であると解釈される。しかし、本研究の前提である「デザインないしスタイルは単なるパッケージングではなく、それ自体がテクストの伝達特性を形成する基体である」という認識に照らしつつこの「自由」を考えるならば、ウェブテクストにおけるそれはいわゆるコンテンツのそれに留まり得ないことは明らかである。それはむしろスタイルの発動においてあるべきものである。そしてその自由を妨げるものは HTML 言語規格ではなく、また、その統一規格に反抗することが自由の発露であるのでもない。上述のように HTML 仕様はスタイル上の自由裁量の余地を大幅に残しているし、そもそもこの統一規格の存在そのものが、これまでできなかった多くのことを可能にし、長らくデザイナーを悩ませてきたクロスブラウザ問題などをも漸近的に解消しつづけているのである。スタイルの自由な展開を妨げるものがあるとすればそれは、この自由裁量の余地の活用を単純に「アーティスティックな」それとしてあたかも例外的なものであるかのごとく許容するような態度である。言いかえれば、流通している種々のテンプレートや市販カタログ等によるスタイル上の「推奨」を「規範」と取り違えることで、ブラックボックスの更なる不可視化に加担する態度である。HTML はデザインの現場に規範と自由とを共々にもたらす。マークアップと「見かけ」の相関性を解析することは、この規範と自由との相関性を解析することに他ならず、それはまさしく古来の人文的な問いである。この認識はより広く共有される必要があるだろう、なぜならこの問いは、下述のように、言語テクストの流通形態全般、およびそこにおいて発効する権威や権利の問題にも通じてゆくからである。

②テクスト流通制度全般の見直しに関する考察と展望

紙媒体と電子媒体（ウェブ）の最も根本的な違いのひとつがマークアップ言語の介在であることは言うを俟たないが、それ以上に重要なのは、この介在物がそれ自体「目に見える」「読める」という、しばしば見逃されがちな事実である。マークアップ言語によって記述されたソースコードは、ブラウザによって表示されるウェブサイトのページ上にあるテクストを支持する媒介物であると同時にそれ自体テクストでもあるという、紙媒体において類比可能なものごとを見出すことの極めて困難なこの事態は、最終的に「読まれる」テクストというものの相対的な位置づけに大きな動揺をもたらす。ソースコードはそれ自体「スマートに」「エレガントに」書かれねばならないという、技術者間にほぼ共有されているいわば作品意識と、その意識に基づいて書かれる作品に他ならないコードとの介在が、それらが支持する表示テクストの本質にいかなる影響――ないし変容――をもたらすかという問いは、実は言語文化論的に非常に重要な問いである。すなわち、テクスト（言語作品）における「本文」とは何かという問題であって、URL にアクセスしたときにまず画面上に表示されるウェブページに載っているテクストは、実はもはや本文ではなく仮象にすぎず、むしろソースコード自体こそが本文であると考えることが可能である――すなわち、ソースコードはあくまでも「読まれうる」ものにすぎず全てのユーザーがそれを読むわけではないにせよ、ともかくも読まれることが常に可能であるのと同じ程度に、そう考えることが可能である。そしてそのように考えたとき、「本文」のみならず例えば「著者」「作者」「作品」といった文化的諸概念が大きく動揺し、言語テクスト全般の流通とそれにまつわる経済システムに変革が生じうる。というのは、書籍出版全盛の時代に築き上げられた従来の言語テクスト流通制度は、「テクストを書く者＝著者」を頂点として、その「テクスト」に最終的構造とスタイルと物質的形態を付与する種々の職能者が順次連なる形で形成されたヒエラルキーに密に寄り添うものだったからである。そのヒエラルキー及び旧来の分業体制は一方では速やかに崩れつつあり、一方ではなお強固に維持され、まさしく過渡期的な様相を呈している。

例えば著作権に関しては、映像・音楽コンテンツをはじめとする諸作品の著作権に関する議論は昨今かまびすしいが、書籍におけるデザインの著作権およびプログラムのそれに関する議論や意識は未だに、誰もが満足するところへ至っているとは言い難い。ウェブサイトにおけるデザイン著作権とプログ

ラミングのそれは極めて隣接し、ほぼ融合的であると言えるが、これに関する議論が今後どのように導かれて行きうるかは、上記の「著者」「作者」「作品」等の概念の変動と歩を一にするはずである。例えばユーザーは何に対して誰に対価を払うべきなのかといった問題を含めて、制度全般が持っている今なお極めて可塑性の高い潜在的可能性は、まさしく、一種のグローバル言語に他ならないマークアップ言語が呈示する規範と自由の相関性の内にひそんでいるということ、それゆえ、この相関性を解析することははなはだ火急の任務であるということが、改めて深く認識された。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔学会発表〕（計１件）
①橋本喜代太、武村知子、森田敏生　「文書スタイル再考――見栄えと意図」電気学会／情報システム研究会、2013 年 5 月 14 日、機械振興会館地下 3 階 2 号室

６．研究組織
(1)研究代表者
武村 知子（TAKEMURA TOMOKO）
一橋大学・大学院言語社会研究科・教授
研究者番号：60323896

(2)研究分担者
橋本喜代太（HASHIMOTO KIYOTA）
大阪府立大学・人間社会学部・准教授
研究者番号：50278818