

HTMLの拡張による 電子文書の添削の試み

春 木 良 且
安 部 清 哉

1. 電子文書とその問題点

Internetを中心とした世界規模の情報通信手段の現実化によって、昨今の情報交換は、コード化された電子文書の形式をとるものが増加してきた。紙を媒体とした文書に比べると、電子文書には以下のような利点がある。

- ・加工、再編集が可能であるため再利用性が高い
- ・媒体の物理的な滅却、摩耗による情報の劣化が無い
- ・保管や伝達に要する労力、費用が低い
- ・入力者の個性が文書には反映しないため、一定の品質が保証される

逆にその欠点としては、以下の2点が重要なものとして指摘できる。

1. 文書の可読性が保証されていない。

電子文書は、その作成と表示の両局面において、必ず電子機器を必要とする。さらに、媒体自体あるいは文書処理ソフトウェアに様々な仕様、製品があり、それらは必ずしも互換性が保証されているわけではない。そのため、特定の文書の受信者が、必ずしも可読であることが保証されているわけではなく、電子文書を伝達する対象は限定されることになる。

2. 文書の証憑性が保証されない。

電子文書において再加工がなされた部分は、原データと区別をすることが不可能である。さらに削除がなされることによって、原データ自体が存在しなくなってしまうことすら考えられる。そのため、文書が特定の作者により作成されたものであるか、どのような変更がなされてきたのかを、文書自体から知る

ことはできない。

本論では、電子文書ベースによる語学教育における文書の添削作業を例に、こうした問題に対するアプローチの一つとして、ページ記述言語であるHTML (Hyper Text Markup Language)の拡張による試みを提案する。

2. HTMLの機能と限界

2.1 文書記述の標準としてのHTML

Internet上でWWW(World Wide Web)を作成するための記述言語として、HTML(Hyper Text Markup Language)が注目されている。WWWを巡る昨今の状況から鑑みると、ページ記述言語としてはHTMLがデファクトスタンダードの地位を獲得していると言って良いであろう。

現在までHTMLは、NetscapeなどHTML文書表示ソフトウェアであるブラウザをベースに仕様化されてきており、またブラウザのベンダ各社が独自の使用拡張を行ってきたこともあり、WWWのビューアに対するソース言語としての側面が非常に強い [1]。W3Cによる最新の仕様案であるHTML3.2においても、その傾向には基本的な変更はない [2]。

HTMLをWWWに限らない電子文書の記述手段として考えると、Internetベースで開発がなされたという経緯から明らかなように、ホームページの作成のみならず、機種やドキュメント処理ソフトウェアの仕様の違いを越えて、電子文書を統一的に扱うための規範としての側面に注目することができる。すなわちHTMLは、文書の構造をコンテンツとは分離して記述するものであり、マルチフォーマットのデータに対しても、その表示処理自体はそのデータ内にカプセル化してオブジェクト指向的な扱いをしている。そのためどのような文書であっても、標準的な入力デバイスであるキーボードに割り当てられている文字記号のみを使ったテキストファイルによって、その構造を完全に表現することが可能である。こうした性質により、HTMLは異機種間での情報交換に対応した文書記述手段として、電子文書の持つ欠点の一つである文書の可読性を保証するものである。

HTMLの拡張による電子文書の添削の試み

電子文書を扱う場合においては、異機種、異ソフトウェア間でのデータ互換の問題に悩まされるのが常であった。それどころか、昨今のような情報技術の進歩が著しい状況においては、同一プラットフォーム上であっても、バージョンによってはデータの互換が保証されない場合が多々存在する。こうした問題に対する現状での最も有効な解決手段は、このHTMLであろう。それはWWWの急激な一般化を見ても明らかである。

2.2 紙媒体とHTML

紙媒体の文書は、非常に多様な情報を包含している。それらは通常、コンテンツ（文書内容）と文書構造に分類されている。電子化文書においてもこの区分は踏襲されており、例えばワードプロセッサには、コンテンツのみを扱うファイル形式（テキストファイル）と文書構造をも含めて扱うファイル形式（バイナリファイル）の両者が用意されている。HTMLは、文書構造のみを扱う記述形式であるが、ファイル形式としてはコンテンツと同様のテキストファイルであるところにその特徴がある。つまり電子化された文書は、文書のコンテンツ及び構造情報を電子化したものである。

しかし我々が紙媒体を用いて扱うことができる情報には、文書のコンテンツとその構造以外に、文書の履歴も含まれている。手書き文書の場合には、書体やインクの色調、印刷形式の文書ではフォントや印刷位置など、また紙の色や質など媒体自体や、時には表現形式などコンテンツ自体などから、その文書のその時点までの改変の履歴を、文書から情報として読みとることができる。現実の社会関係においては、文書の履歴は非常に重要な機能を持っており、有印文書では改変項目に対しても印章を必要とするなど、こうした情報が実際に重視されている例は多い。

改変されるのは文書のコンテンツであるが、文書履歴は文書のコンテンツ自身と考えることはできない。しかし文書構造にも当てはめることはできない。なぜならば、構造の変更を伴わない文書の改変も考えられるためである。つまり紙媒体の文書は、固定的なものではなく可変的な存在であるが、印章以外に

HTMLの拡張による電子文書の添削の試み

それらを明示的に表現する手段を持たないため、我々は通常それらを余り意識せずに文書を扱っているということになる。

前述したように、電子文書は文書をコンテンツと構造の2面から情報を抜き出し、電子化したものである。そのため、文書の加工や再編集作業などの履歴情報を含めて、直接電子文書上に記述することは、本来不可能である。さらに電子文書の利点である再利用性の高さが、文書から完全に履歴を推定させるための情報を捨てさせてしまう結果となる。こうした理由から、電子文書には証憑性が認められることがない。それはHTMLにより作成された文書でも全く同じである。タグ「<!..>」及び<ADDRESS>によって、コメントや文書の作成者を明示することができるが、コンテンツの履歴を記述する手段は無い。さらにタグ自体を消去してしまえば、それらは無意味となってしまう。結局HTMLによっても、電子文書の包含する問題の2番目として指摘した、文書の証憑性を実現することはできない。

つまり、紙媒体の文書が暗黙の形式で保持している履歴情報を明示的に保持する手段がなければ、この問題点は解決できない。それが顕在化する場合として、ここでは文書添削という語学教育上非常に重要な文書処理の応用を考える。語学教育は、異なった国の様々な学習者を想定するため、電子文書の持つ問題点の1番目として指摘した、広範な受信者による可読性はまず前提である。前述のように、この問題に対する最も有効な解決手段は、現在のところプラットフォームを限定しないHTMLである。そのためHTMLベースでの作業は、語学教育においても、有効であると結論付けることができる。

添削作業は、完全な文書を作成することが目的ではなく、そこまでの過程を文書として記録していくことがその本質である。つまり、電子文書が本来包含していない履歴情報が最も重要な文書の内容となる応用分野である。語学教育が、昨今の情報技術の進歩と世界規模の情報通信網の整備の恩恵を受けるには、電子文書の証憑性を保証する手段を現実化することが急務であるといっても過言では無かろう。以降には、文書添削における要求事項を明らかにし、HTMLによるその実現を考察する。

3. 文書添削における要求

3.1 語学学習と添削

近年では、世界規模での情報通信手段が充実してきており、それにともない様々な言語に触れる機会も増加してきている。しかしこと日本語教育について考えた場合、以下のような問題を指摘することができる。

- ・海外における日本語教師の不足
- ・日本語教師志望者の実習機会が少ない
- ・日本語教師有資格者が公的場面でその能力を発揮する機会が少ない

これらにはいくつか理由が考えられるが、日本語の言語特性がかなり大きなウェイトを占めていると思われる。日本語は、どちらかと言えば記述が重視される言語体系であるため、添削という学習形態が、特に国語教育、日本語教育の上では非常に重要である。しかし、紙を媒体とした文章形式のやりとりでは、冗長かつ不便である。黒板を使って指導するように、リアルタイムかつグラフィカルに原文書を直接添削するような機会が必要とされているが、現状ではまだ実現されてはいない。

3.2 校正記号と添削

紙媒体に作成された文書を修正、変更するための記述形式としては、日本工業規格に印刷校正記号として定められている。これは、印刷出版時における文書の校正を想定したもので、非常に多岐に渡る処理を包含している。これらを文書に対する機能の観点から分類すると、およそ以下に示すようになる。

1. 文書のコンテンツを対象とするもの。
2. 文書の構造のみを対象とするもの。
3. 文書の表記を対象とするもの。

おのこのの区別は明確なものではないが、1は主記号1.1(文字の変更、削除)、1.3(字間への文字挿入)、1.12(文字、行の入れ替え)、1.14、1.15(文字の移動)などのように文書内容を変更の対象とし、その結果として、文書の構造も

HTMLの拡張による電子文書の添削の試み

変更されてしまうものである。2は、文書内容ではなく構造だけを変更するもので、1.7(字間, 行間アケ), 1.9, 1.10(行移し), 1.11(行起こし), 1.13(行継続)などがある。3には、字体, 字の並び, 大文字小文字化, あるいは不良の文字を校正するなど, コンテンツや構造よりも, 主に印刷時での見栄えを意識したものが含まれる。これらのうち, 言語教育における添削で重視されるのは, 主にコンテンツに関係する1であり, 2, 3は, 使われない。そこで, 2, 3は本論の対象外とする。

校正記号は, 主記号と併用記号から構成されている。主記号は図形表示を中心に構成されており, 処理を行う位置と処理の内容を指定する。併用記号は, 主記号に付加することによって, その処理の具体的な属性を指定する。

以上から, 文書の校正に必要な情報は,

- ・位置情報
- ・処理内容情報

の2つであることがわかる。

コンテンツを対象とした処理に限定すると, 処理内容は,

- ・追加
- ・削除
- ・変更
- ・移動

に集約される。位置指定は, 処理内容によっては, 位置情報が点であるものと幅情報を必要とするものが考えられる。

4. HTMLの拡張仕様

4.1 原文書の扱いと添削の基礎情報

前項で述べたように, 添削作業では原文書が存在する。原文書に対して添削がなされて, コンテンツや文書構造が変更されて行くわけであるが, 添削がなされた後でも, 原文書自体を識別できねばならない。

添削処理は, 原文書に対して複数回行われることが想定される。そのため要

HTMLの拡張による電子文書の添削の試み

求としては、添削処理が行われた回数を直接記述できる機能が必要となる。これは添削の内容自体ではなく、添削の基礎となる情報である。文書全体に関する情報なので、<HEAD>部で</TITLE>以降に指定する。これらを示すタグを以下に示す。

<CORRECT>

<G0 原文書のタイムスタンプまたは作成者識別名>付加情報

<G1 期間指定または添削者識別名>

<G2 ...>

<G2.5 ...>

</CORRECT>

タグ<CORRECT>から</CORRECT>まで、添削の基礎情報として、原文書からの世代別の添削履歴を記述する。タグ<G0>は原文書を示し、オプションとして、原文書作成時のタイムスタンプ、作成者の識別名などを指定できる。さらにタグの後には、テキストとして付加情報を記述できる。このテキストには終了タグが無く、次のタグまでが付加情報として解釈される。

<Gn>は、n世代目を示すタグである。世代が整数値の場合「校了」であり、「要再校」などの場合は、n.xの形式で示される。ここでも、オプションとしてタイムスタンプか添削者の識別名を指定することが可能である。タイムスタンプの場合、添削時点を直接指定する絶対指定と、原文書の作成時に対して経過時間を指定する相対指定がある。ここで必須な情報は、いつ、あるいは誰によって添削が行われたかを特定する情報ではなく、その文書が何世代目かという点である。そのため、オプションの選択が世代によって異なっていたり、あるいはオプション指定が全く省略されていても問題はない。後述するように、添削情報の具体的表現はビューアに委ねるため、これらはオプション情報の文書中での表示形態を規定するものではない。

この添削の基礎情報は、以降のタグで指定する具体的な添削指定を統一的に

HTMLの拡張による電子文書の添削の試み

管理するものである。そのため、添削指定と基礎情報に矛盾がある場合、原則として基礎情報が優先することになる。

4.2 具体的な添削の指定

前述のように、校正情報は、位置情報と処理内容情報から構成されている。そのため個々の添削処理では、以下を指定する必要がある。

1. 文書中における添削位置の指定

2. 処理内容の指定

このうち1は、標準HTMLと同様にタグ位置によって指定するのを原則とし、2はタグ名によって示すことを原則とする。

前述のように、添削処理には追加、削除、変更、移動が考えられる。このうち変更と移動は、削除と追加の組み合わせによって実現が可能であるが、変更は頻度が高く重要な処理であるため、独立した操作として考える。移動は、多くの情報を指定しなければならないが、添削にはそれほど頻出するものではないため割愛した。

添削作業においては、対象自体を文書中から実際に削除することは許されない。そのため、削除は削除情報の追加という形式で、行われねばならない。つまり削除は、位置指定のみで実現できることになる。但し、この位置指定は点指定ではなく、幅指定である。追加に関しては、位置指定と追加テキストの指定が必要である。変更は、位置指定と追加テキスト、さらに削除対象の指定が必要である。

前述のように、添削作業では削除もテキストの追加という形式で行われるため、変更や追加テキストが大きい場合、文書が複雑化してしまうことが考えられる。そこで、位置指定と追加、変更するテキストを分離して指定することを可能とした。具体的には、位置指定とテキストの指定を同時に行う直接指定と、両者を別個に行う間接指定の2つである。

・直接位置指定

HTMLの拡張による電子文書の添削の試み

添削内容を示すには、必ず位置指定が必要である。添削のタグは文書の< BODY>中で他のHTMLタグと混在するため、タグ文字の先頭に@を付けて「< @..>」の形式を取ることによって明示化する。前述のように、添削で行われる処理は、削除、追加、変更である。これらは、おのおのタグ「< @D..>」「< @A..>」、「< @R..>」によって示される。さらに各処理には、基礎情報として添削の履歴が存在する。これは、タグオプションとして扱われるべきであるが、これ以外のオプションは想定できないため、直接タグに対して「< @X=Gn>」の形式で情報を付加して記述する。このタグは、n世代目の添削作業において処理Xがなされたという解釈をすることができることになる。なお前述のように、添削の基礎情報が内容よりも優先するため、例えば「=Gn」として与えられたn世代が基礎情報として定義されていない場合、この添削は無効となる。

以下に、直接指定の形式例を、削除、追加、変更の順に示す。

< @D=Gn>削除対象</@D=Gn>

< @A=Gn>追加文</@A=Gn>

< @R=Gn>変更対象</@R=Gn>変更文</@R=Gn>

このように直接指定では、タグに追加する文や変更する文が続く形式をとる。変更の場合、処理内容は実質的には「削除+追加」であるため、原文書から変更対象を指定し、さらに追加する文を指定しなければならない。そのため、終了タグが2つ無ければならない。

・間接位置指定

間接指定の場合、位置指定と内容の指定は分離する。位置指定タグは、< BODY>内に記述される。ただし削除には、間接指定は有り得ない。つまり間接の位置指定がある場合には、必ず内容指定が存在することになる。そのため記述の手間を省くために、タグは「< @Gn=s>」の形式をとる。ここでsは、同一世代における位置指定タグの識別子である。これには、1文字の数字あるいは

HTMLの拡張による電子文書の添削の試み

文字を与えることができるが、連続している必要はなく、さらに同一のものがある場合、両者は同じ添削内容を持つものとされる。追加の場合、開始タグのみが必要であり、終了タグは不要である。また変更の場合、終了タグは一つだけである。終了タグは、他の終了タグと同様に「</@Gn=s>」となる。終了タグが無い場合は、追加処理とみなされることになる。

間接指定では、添削内容の指定が必要である。これは、勿論コンテンツの一部となるため、本来<BODY>部に記述されねばならないが、他の文書内容と区別するために、</BODY>と</HTML>の間で、タグ<EDIT>から</EDIT>内に指定される。各内容はタグ<Gn=s>以降にテキストとして指定される。またタグ<Gn></Gn>間で、複数項目をタグ<s>以降に指定する列挙表示も可能である。これによって世代毎に添削内容をまとめて記述することが可能となる。

位置の指定と内容の指定に矛盾がある場合、原則として指定自体が無効となる。位置指定のみで内容指定が無い場合、あるいはその逆の場合、指定は無視されることになる。以上をまとめて示す。

```
<EDIT>
<Gn=s>添削内容
<Gn>列挙表示
  <s>添削内容
  <s>...
</Gn>
</EDIT>
```

5. 考察

以上の拡張タグで、添削に要する主要な情報は、電子文書中に記述することが概ね可能である。ただしこれらは、具体的な表現形式を規定するものではな

HTMLの拡張による電子文書の添削の試み

い。これらタグ情報が文書上でどのように表現されるかは、HTML同様にビューア依存である。具体的には、

- ・削除部分の表示方法
- ・変更部分と追加部分の表示位置
- ・原文書と世代別の添削情報の表現形式
- ・文書中での位置指定方法

などは添削タグ側では規定しない。さらに添削タグは原文書のタグとは独立であるため、添削の対象文書は少なくとも原文書レベルではHTML文書である必要はない。そのためここで述べたHTMLの拡張仕様は、HTML自体とは別個に実装が可能であり、WWWブラウザへの組み込みのみならず、独立した添削用ビューアとしての実現が可能である。

問題としては、

- ・意図的な原文書の削除や改算を防ぐことが不可能であること
- ・HTML形式がより複雑化するため、文書作成ツールが必須であること
- ・既存のHTML自体との統合性が低いこと（特にSGMLの文書型定義・DTDとの関係）
- ・一文字処理や「イキ」指定など、よりきめの細かいタグ指定が必要なこと
- ・ページ区切り指定が必要なこと

などが指摘できる。

その他にも、まだ解決しなければならない課題を包含しているが、本手段によって電子文書の広範な応用が考えられる。たとえばInternet上の電子メールを利用した、日本語教育なども可能になるとと思われる。

なお本試みは、W3C(World Wide Web Consortium)に対する提案として準備を進めている。また、平成8年度フェリス女学院大学大学院特別研究費（研究代表者 安部清哉）による成果であることを付記しておく。

参考文献

- [1] 出水法俊：World Wide Web Consortium,bit,vol.28,No.9,pp79(1996)

HTMLの拡張による電子文書の添削の試み

- [2] <http://www.w3.org/pub/WWW/Consortium/Prospectus>
- [3] Mac Bride: html-publishing on the world wide web, Hodder&Stoughton(1996)