人文科学とコンピュータ　30－1
（１９９６．５．１７）

JIS X 4051 の行組版方法に基づいた文書整形システムの設計と実現

遠瀬雅宏，早川栄一，並木美太郎，高橋延匡

東京農工大学工学部電子情報工学科

本稿では，JIS X 4051 に基づいた行組版機能をもつ文書整形システムについて述べる．従来の文書整形システムでは，日本語特有の組版処理を行なうときに問題が生じる．その原因は，日本語には明確な組版処理単位がないことである．そのような問題を JIS X 4051 で規定される不可分文字列を組版処理単位とすることで解決することにした．本研究では，JIS X 4051 の行組版方法に基づいた行組版機能と文書構造の相関関係を利用するレイアウト機能をもつ文書整形システムを設計・実現した．そして本システムを文書出力に用い，組版処理について評価を行なった．その結果，不可分文字列による組版処理の有効性を確認した．

Design and Implementation of a Document Formatting System Based on the Line Composition Rules of JIS X 4051

Masahiro TOOSE, Eiichi HAYAKAWA, Mitarou NAMIKI
and Nobumasa TAKAHASHI

Department of Computer Science, Faculty of Technology,
Tokyo University of Agriculture and Technology

This paper describes a document formatting system with the function of line composition based on JIS X 4051. Since a unit of typesetting is not clear in Japanese, many current document system cannot perform the typesetting peculiar to Japanese. To solve this problem, we adopted the concept of undividable character strings defined JIS X 4051. Also, we designed and Implemented the document formatting system with the functions of line composition based on JIS X 4051 and layouting using document structures. To evaluate the function of line composition, this system has been used in the printout of documents. The results confirmed the effectiveness of the typesetting using undividable character strings.

1. はじめに

既存の文書整形システムの多くは，欧米言語を対象に開発され，日本語を処理できるように拡張されている．このため，日本語組版特有の和欧文混植や縦書きを行なう場合に問題が生じることがある．その原因の一つとして，欧米言語を対象とした組版処理がTeXのボックス&グルーモデル[1]に見られるような単語と単語間のスペースを基本として行なわれていることが挙げられる．このような組版処理モデルをそのまま日本語に適用すると，日本語組版に出現する多様な文字種に対応できない場合がある．そこで，日本語の文字種を考慮した組版方法[2]も考案されている．

一方，最近では文書の電子化が進み，電子文書による文書交換が頻繁に行なわれるようになっている．このときの電子文書の同等性つまり可搬性を保証するために，文書交換用の規格の一つとして行組版方法の標準化が必要とされていた．そして，1993年に日本語文書の行組版方法を規定したJIS X 4051「日本語文書の行組版方法」[3]が制定された．

そこで我々は，JIS X 4051に記述されている行組版規則や処理モデルに注目し，その規格を基盤とする日本語を対象とした文書整形システムを設計・実現した．本稿では，その文書整形システムの設計と実現について述べる．そして，本システムの組版処理結果について考察する．

2. 従来の文書整形システムの問題点

従来の文書整形システムにおける問題点として，行内の文字の位置を決定する行組版の問題とその作成した行を配置するレイアウト処理の問題がある．

2.1 行組版の問題

従来の文書整形システムの行組版における問題点は次のとおりである．

(1) 日本語禁則処理の問題

日本語組版では，行頭や行末に配置してはいけない文字が存在する．また，英単語のように途中で改行してはいけない文字列が存在する．そのような文字を配置する処理が行頭禁則処理，行末禁則処理，分離禁則処理である．従来の文書整形システムでも禁則処理をできるものは数多くあるが，その処理のために行末が揃わずに体裁が悪いものになることも多い．

(2) 和欧文混植処理の問題

和欧文混植処理とは，和文の中に欧文が混在すときの組版処理である．通常，和文から欧文，欧文から和文に切り替るときに空白が挿入される．また，この空白は，行末を揃える処理のときに伸縮される．このような処理ができない文書整形システムもある．また，(1) の分離禁則処理をしようとして欧文がカラムに収らなくなるものも多い．

(3) 約物処理の問題

括弧や句読点を約物という．通常，約物の前後にそれぞれ決められた幅の空白が挿入される．しかし，約物が連続する場合や行頭，行末に約物が配置される場合には，その空白は取除かれる．従来の文書整形システムの多くは，文字の種類を考慮しないために，空白を除く処理が行なうことができない．

上記のような問題は，欧米言語がスペースによって明確に分ち書きされるのに対して日本語は分ち書きされないということと日本語組版規則には文字の種類に依存するものが多いことが原因と考える．したがって，日本語組版を行なう場合には，文字の種類を利用して文章を何らかの処理単位に区切る必要がある．また，1行ごとに組版処理を行なうと分離禁則処理などで行末が揃わなくなることがあるので，複数の行で行組版を行なう必要がある．

2.2 レイアウト処理の問題

文書整形を行なう場合，行組版で作成した行のレイアウト処理も重要である．たとえば，章の見出しがページ下端に配置され，その本文は次のページに配置されることがある．このようなことを防ぐために見出しを次のページに配置する処理をウィドウ処理という．従来の文書整形システムでは見出しの前で人手で改ページを行なって解決していたが，これは文書作成時の大きな手間となる．しかし，ウィドウ処理をシステム側で支援しようとした場合，文書構造の相関関係情報が必要となる．従来の文書整形システムでも文書構造を記述するものもあるが，構造ごとに書式を統一する程度の処理にとどまっている．

3. 全体設計

3.1 本システムの特徴

本システムの特徴は，次のとおりである．

(1) JIS X 4051 の行組版方法による文書整形

2.1 節の問題を解決するために，本システムでは JIS X 4051 を採用する．採用する理由は，次のとおりである．

○不可分文字列による行組版

JIS X 4051 では行の分割位置の決定方法とその行内の文字の配置規則を規定している．その行組版方法の特徴が，不可分文字列という概念である．不可分文字列は，JIS X 4051 における行組版処理の最小処理単位となるもので，それ以上分割できない文字の集合である．次のようなものを分離できないものとしている．

- 欧文句
- 連数字
- 行頭禁則文字とその前の文字
- 行末禁則文字とその後の文字

図 1 は不可分文字列の例である．この例文において，JOSHO は欧文句，'、' と '。' と ')' と 'J' と 'ー' が行頭禁則文字 '「' と '(' が行末禁則文字となるので図 1 のような不可分文字列の並びになる．

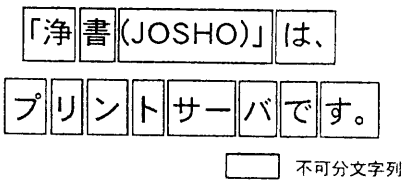

図 1 不可分文字列の例

○段落単位での行組版処理

JIS X 4051 では，行組版処理を行なう単位により，二つの規格合致性を設けている．その水準 2 が段落単位での行組版である．段落単位で行組版処理を行なうことにより，行末を揃えるための空白の幅調整（スペーシング処理）に余裕を持たせることができる．

(2) 文書構造を利用した文書整形

2.2 節の問題を解決するために，本システムでは章や節，見出しや図といった文書の論理構造を利用して次のような文書整形を行なう．このような文書整形機能を提供することにより，文書作成時の書式設定やレイアウトの手間を軽減する．

○文書構造を利用したレイアウト処理

ウィドウ処理や図表の配置処理といった文書構造の相関関係情報を必要とするレイアウト処理を行なう．

○文書構造による書式設定

章の見出し，節の見出しといった文書構造に対して使用するフォントの書体，文字の大きさなどの書式設定を行なうことによって，文書全体で同じ構造間の書式を統一する．

3.2 設計方針

本システムの設計方針は，次のとおりである．

(1) 組版規則に柔軟性を持たせる

JIS X 4051 の行組版を行なう上で文字の分類やパラメータが必要となるが，それらを外部ファイルで設定できるようにし，組版規則に柔軟性を持たせる．

(2) 科学技術文書を対象とする

本システムは科学技術文書を主な対象文書とし，その文書構造として章，節 1，節 2，節 3 の階層構造，それぞれの階層構造に属する見出し，文，図，表を定義する．

(3) 書式・組版情報を文書と分離させる

本システムでは，文字の大きさや書体などの書式情報，禁則文字の設定などの組版情報を文書ファイルとは別のファイルで設定する．このように入力ファイルを複数で構成することにより，文書ファイルの変更をしないで，さまざまな書式や組版規則の文書印刷を可能とする．

3.3 システム構成

前節の設計方針に基づき，システムの設計を行なった．そのシステム構成を図 2 に示す．

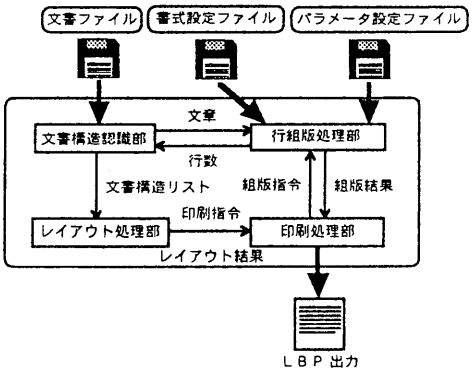

図2　システム構成

本システムは、書式設定ファイル、パラメータ設定ファイル、文書ファイルの三つの入力ファイルをもとに組版処理を行なう。各ファイルのの内容は次のとおりである。

・書式設定ファイル

ページ書式の設定、論理構造ごとの文字の大きさやフォントの設定などを行なう。

・パラメータ設定ファイル

禁則文字などの文字クラスの設定、行組版処理で必要なパラメータの設定を行なう。

・文書ファイル

印刷する文書と文書構造を表わす指令語を混在させて記述する。

またシステム内部は，文書構造認識部，レイアウト処理部，行組版処理部，印刷処理部で構成される．それぞれの機能は次のとおりである．

・文書構造認識部

文書ファイルから文書構造だけを抽出し，それぞれの構造にレイアウト処理のための情報を付加する．

・レイアウト処理部

文書構造認識部で付加された情報をもとにレイアウト処理を行なう．

・印刷処理部

レイアウト処理部の結果に従い印刷処理を行なう．その際，文章については行組版処理部へ文字列を与えてその処理結果を利用する．

・行組版処理部

JIS X 4051 の行組版方法に基づいた行組版処理を行なう．

4. 各処理部の設計

4.1 文書構造認識部

文書構造認識部では，文書構造と印刷文書が混在している文書ファイルから文書構造だけを抽出する。そして，文書の流れを表わす図 3 のような文書構造リストを作成し，その各要素に対して次の情報を付加する．

(1) 構造属性

その文書構造がどのようなもの（見出し，文，図，表）で，どの階層構造に属するものなのかを示す．

(2) 高さ

その文書構造が持つ紙面上の物理的な高さがどれくらいなのかを示す．

また，文書構造リストの各文書構造間の関係として，次の情報を付加する．

(1) 不可分性

注目している文書構造が次の文書構造と分離されてよいかどうかを示す．この情報により，見出しとその本体が違うページに配置されることを防ぐことができる．

(2) 交換可能性

処理している文書構造がそのカラムに配置不可能になった場合に，その次の文書構造との順序を入れ替えられるかどうかを示す．この情報により，図表がそのページに入りきらない場合に，その図表を次のページを送り，空いているスペースを文章で埋める処理などができる．

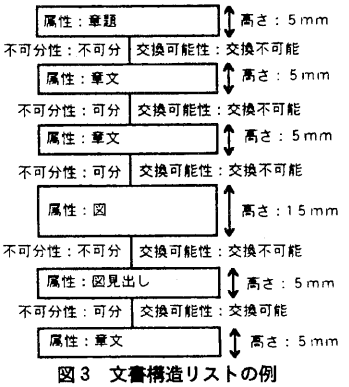

図3　文書構造リストの例

4.2 レイアウト処理部

レイアウト処理部では，文書構造認識部で作成された文書構造リストをもとにレイアウト処

理を行なう．レイアウト処理を行なう構造の単位は，文書構造リストの不可分性が不可分とされる文書構造の集合（以下，構造群とする）とする（図4参照）．

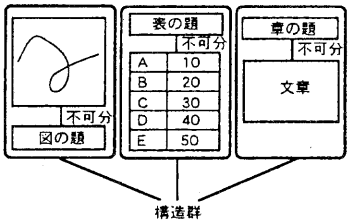

図4　構造群の例

そして，次の規則でレイアウト処理を行なう．

(1) 対象構造群がそのカラムに入るならば，そのまま配置する．
(2) 対象構造群がそのカラムに入らない場合，次の構造群との交換可能性をみて，交換可能でその構造群がそのカラムに入るならば，構造群を交換しそのカラムに配置する．
(3) (2) で交換不可能な場合や次の構造群がそのカラムに入らない場合には，次のカラムにその対象構造群を配置する．

(1) から (3) の優先順位でレイアウト処理を繰り返し行ない，ページを溢れるときには，印刷処理部へ印刷指令を出す．

4.3　行組版処理部

行組版処理部では，JIS X 4051 の行組版方法に基づいた行組版処理を行なう．本システムでは，段落単位での行組版処理を行なう．その処理の流れを次に示す．

(1) 不可分文字列リストを作成する

まず与えられた 1 段落分の文字列に対して，それぞれの文字に文字クラスを与え，前後 2 文字間のクラス関係によって不可分文字列を作成する．そして，1 段落分の連結リストを作成する．また，このときに文字のサイズ，書体，修飾などの情報を文字に付加する．

(2) 段落候補木を作成する

(1) で作成されたリストを設定した行長（以下，指定行長とする）ごとに区切る．このとき，指定行長の前後二つの区分点を行分割候補点とする．そして，それぞれの行分割候補点から再び次の行分割候補点を探す．この作業を段落の先頭から最後まで行ない、行分割候補点の二分木を作成する．つまり，N 行の段落では，最大 2^{N-1} 個の段落候補が作成されることになる．

(3) 段落構成を決定する

(2) で作成された段落候補の各行に対してペナルティ関数と呼ばれる評価関数でペナルティを算出する．そして，段落候補ごとにこのペナルティの総和を算出し，総和が最小となる段落候補を探索する．総和が同じになる段落候補については，行長調整における詰め処理回数が多いほうを優先する．

(4) 行長調整を行なう

(1) ～ (3) で決定された段落構成の各行に対して，それぞれ詰め処理，延ばし処理のいずれかを施して行長を調整する．

5. 実現

前述した設計に基づき，レイアウト処理部以外の部分を本研究室で開発したプリントサーバ「浄書」上に実現した．「浄書」のハードウェア環境は、次のとおりである。

LBP：SL－2000（日立製作所）
　　解像度：400 dpi
システムバス：VME バスシステム（Force）
CPU：MC68030 25MHz（モトローラ）
メインメモリ：8M バイト
フレームメモリ：16M バイト

また、本システムのソフトウェア環境は次のとおりである。

OS：OS／omicron
開発言語：言語 C
総行数：約 6,000 行

OS／omicron は、本研究室で開発した OS である。特徴としては、内部コード系にフル 2 バイトコードを採用していることである。本システムでも、OS の内部コード系である JIS X 0208 による文字種の分類で組版処理を行なっている。

6. 評価

本システムが正しく日本語組版処理を行なっているかどうかを評価するために，本稿を例に

禁則処理，スペーシング処理，約物処理について調査した．

6.1 禁則処理の評価

本稿について禁則処理の結果を表1に示す．

表1 禁則処理の結果

	対象総数	成功数
行頭禁則処理	485 文字	474 文字
行末禁則処理	63 文字	63 文字
分離禁則処理	129 文字列	129 文字列

総文字数： 6427 文字

行頭禁則処理が成功するとは，行頭禁則文字がその前の文字と不可分文字列になっているかで判断した．行末禁則処理の場合も同様で，行末禁則文字がその後の文字と不可分文字列になっているかで判断した．分離禁則処理については，分離されてはいけない文字列が不可分文字列になっているかどうかで判断した．行頭禁則処理が11か所失敗しているが，これは箇条書きの先頭で‘・’を配置しようしているためで，通常の組版処理では問題がない．行末禁則処理，分離禁則処理については，100%成功している．このことから，禁則処理は正しく行なわれていることが言える．

6.2 スペーシング処理の評価

図5は，本稿におけるスペーシング処理後の空白の幅による度数分布である．

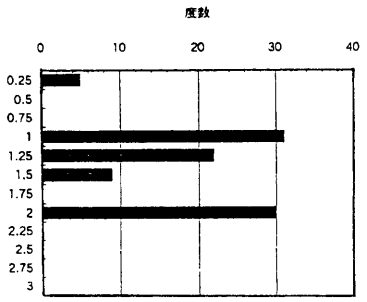

図5 空白の幅による度数分布

本稿は文字の大きさを14級(3.5mm)を印刷したので，欧文・和文間，連数字・和文間には，標準で0.875mm(四分)の空白が挿入される．図5を見ると，最大の空白でも標準のおよそ2倍(二分)程度で，これは規則で決められている範囲内である．したがって，スペーシング処理は成功していると言える．また，全体の約3分の2が標準値の近辺に分布することから，バランスのよいスペーシング処理ができていると言える．

6.3 約物処理の評価

本稿における約物の統計を表2に示す．

表2 本稿における約物の統計

本稿の総文字数	6427 文字
約物の総文字数	548 文字
約物の連続箇所	11 か所
行頭の開き括弧数	28 文字

約物が連続している箇所が11か所あったが，空白を入れず正しく処理されていた(例:3ページ目25行目)．また，行頭や行末に約物が配置された箇所も28か所あるが，同様に正しく処理されていた(例:5ページ目22行目)．このことから，約物処理は正しく行なわれていることが言える．

6.1 ～ 6.3節で述べたように日本語組版規則による組版処理を正しく行なうことができたと言える．このことから，不可分文字列による行組版処理は有効であることが言える。

7. おわりに

本稿では，JIS X 4051の行組版方法に基づいた文書整形システムについて述べた．

今後は，レイアウト処理部の実現と行組版処理部の縦書きへの拡張を行なう予定である．縦書きを行なう場合には、数字や欧文文字の配置方法がさらに複雑になるので不可分文字列のモデルなどを拡張する必要があると考えている．

参考文献

[1] D.E.Knuth:“The TeXbook”, Addison Wesley, 1984.

[2] 川崎敏治，他：“文書清書システムにおける行組版機能の一実現方式”，情処論文誌 Vol.34 No.8, 1993.

[3] 日本規格協会：“JIS X 4051「日本語文書の行組版方法」”，日本工業規格，1993.