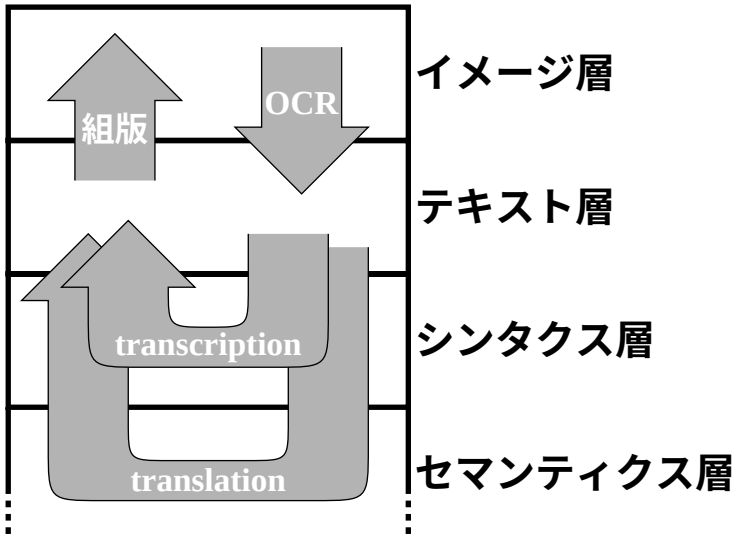
# 「漢字情報学の構築」共同研究班報告

## 1　はじめに

あとで書く。




## 2　組版に関する共同研究

テキスト層の情報を、イメージ層の情報へと変換する技術は、一般に「組版」と総称されている。本共同研究班では、組版のうち日本語組版、特に縦書の組版における種々の特徴を研究し、その背後にどのような背景知識が存在するか、またそれは実際の組版上にどのように実現されるか、について共同研究をおこなった。

ただし、組版技術というのは、工学的技術というよりは、むしろ芸術の範疇に属するらしい。言い換えると、組版は最終的には職人芸の世界であり、そこでは理論より美意識が優先するのである。しかし、美意識などというものは、主観的で個人的なものであり、それが組版に関する研究を、結果的には、かなり困難なものにしている。

### 2.1　日本語組版における禁則と行末調整

日本語組版に特徴的なルールとして、行頭・行末の禁則がある。句読点を行頭におかない、などのルールである。行末における禁則文字は、大概は括弧の起こしであり、始め括弧が直後の文字に「付いている」と理解することで、行末に始め括弧が来ないことを説明できる。一方、行頭における禁則文字は、多種多様なものがあり、しかも必ずしも意見の一致を見ていない。

行頭における禁則文字のうち、一応、意見の一致が見られるものとしては、終わり括弧、句読点、疑問符、感嘆符、中黒などがある。これらに関しては、日本語組版において、行頭に用いるべきではない。これに対し、音引き、拗音、促音に関しては、行頭の禁則文字とする意見がやや強いが、必ずしも守られていない。というのも、これらを行頭禁則とすると、字間の調整をかなりおこなわなければならず、それによって「美しくない」版面ができるくらいならば、行頭禁則にしないという割り切りも必要だ、というのである。

また、行頭禁則文字として意見が一致する「句読点」についても、その実際の処理に関しては意見の対立が存在する。特に縦組においてだが、行頭禁則によって行末に付けざるを得ない「句読点」がある場合に、その行を他の行より 1 文字分長くする (ぶら下げ) か、それとも行末を全てそろえるか、という選択肢がある。「ぶら下げ」る場合には、他の句読点のうち、ちょうど行末に収まっていて元々ぶら下げる必要がない句読点をどうするか、という問題が発生する。

さらに、文字そのものは禁則の対象となっていないが、前後関係によって、行末行頭の泣き別れを禁じる「分離禁則」と呼ばれるルール (らしきもの) も存在する。分離禁則の代表は連数字であり、たとえば「一九九五年」というテキストは、基本的に行末で切ってはならない。ただし、これに対しても、年号の後ろ二桁は分割を許すべきだという意見もあり、その場合には「一九」と「九五年」で改行してもよいということになる。この分離禁則のもう一つの例として、グループルビと呼ばれるものがあるが、それを説明する前に、まずは日本語組版におけるルビ全体を概観していこう。

## 2.2 ルビ

ルビは、いわゆる「振り仮名」の組版形式であり、日本語組版において多用される手法の一つである。通常は、漢字に対して平仮名あるいは片仮名が添えられる*。ルビは基本的には単語に対して振るものであり、たとえば「活躍」というルビの振り方は正しくない。「活躍」という形で、単語全体にルビを振るべきである。

複数の漢字にルビを振る場合、モノルビという手法と、グループルビという手法が存在する。モノルビは、各漢字ごとにルビを振るもので、たとえば「規則」のようなルビの振り方が挙げられる。グループルビは、単語全体にルビを振るもので、たとえば「規則」のようなルビの振り方が挙げられる。どちらを採用するかは、やはり美的感覚ということになるのだが、基本はモノルビで、熟字訓など特殊な場合に限ってグループルビ、というパターンが優勢なようである。

なお、グループルビは、組版における分離禁則として扱うべきである。つまり「規則」のような形でルビを振っている場合には、「規」と「則」の間で改行してはならない。ただしモノルビに関しては、この限りではない。

*本共同研究班では、その他の例もいくつか報告された。通常の例以外で最も多く目にするのは、仮名に漢字のルビが添えられているもので、これはマンガのフキダシで多用されている。また、中国の児童向け書籍の中には、横書の漢字に拼音の「ルビ」を添えている例が報告された。

## 2.3 JIS X 4051と漢文組版

ちょうど本共同研究班の発足直前、2004年3月にJIS X 4051『日本語文書の組版方法』が改正されていた。芸術の範疇に属する組版技術に対し、工業規格がどういう規定をおこなっているのか知りたかったので、本共同研究班でJIS X 4051を読んでみた。

JIS X 4051では、たとえば行頭禁則文字に、終わり括弧、句読点、疑問符、感嘆符、中黒のみならず、音引き、拗音、促音を含めている。また、行末処理は「ぶら下げなし」つまり、行末を全てそろえるやり方を規定している。さらに、グループルビにおいても、分離禁則としない方針を採用しており、その際の改行のやり方をかなり細かく規定している。ただ、JIS X 4051は組版ソフトウェアの基本仕様を目して書かれているため、オプションの付加によって、たとえば「ぶら下げあり」の組版を追加することは可能である。

JIS X 4051中で、本共同研究班の気にさわったのは、これまでに述べてきた禁則やルビではなく、漢文の組版だった。JIS X 4051の漢文組版においては、改行時の返り点を行末に置く方式になっている。しかし、実際の漢文組版においては、返り点を行末に置く方式[†]と、行頭に置く方式、さらにはそれらを混在するやり方がある。できれば、これらの方式を全てサポートしてほしいのだが、やはり工業規格としては、どれかを推奨しなければならないところなのだろう。

## 2.4 漢字フォント

組版という分野において、本共同研究班が、さらなる研究の必要性を感じた技術に、漢字フォントが挙げられる。というのも、東洋学研究者が論文を組版する際には、JIS X 0208など一般の漢字だけでは不十分であり、外字作成にいつも悩まされているのである。しかし、ビットマップフォントならまだしも、アウトラインフォントを自ら作成できる東洋学研究者など、ほとんどいない。ただ、外字と言えど、何の典拠もない文字を使用することはなく、少なくともその外字の画像くらいはスキャンなどで手に入るはずである。

そこで、外字の画像からそのアウトラインフォントを、簡単に作成できるようなツールを開発した。PBMフォーマットの2値画像からアウトライン情報を抽出する部分は、Peter Selinger作の「potrace」[‡]を借用し、アウトライン情報をOpenTypeに組み上げる部分は、班長自作の「eps2otf」[§]を用いた。これにより、2値画像さえあれば、アウトラインフォントを自由に作成できるようになった。

# 3 画像からの文字切り出しに関する共同研究

イメージ層の情報を、テキスト層の情報へと変換する技術は、一般に「OCR」と総称されている。本共同研究班では、刊本や拓本に対する「OCR」に挑戦した

[†]『漢文教授ニ關スル調査報告』, 官報, 第8630號 (明治45年3月29日), pp.703-707.

[‡]http://potrace.sourceforge.net/で公開。

[§]http://kanji.zinbun.kyoto-u.ac.jp/~yasuoka/publications/eps2otfで公開。

かったのだが、漢字を自動認識する技術は、残念ながら本共同研究班では達成できなかった。排印本に印刷された漢字と違って、木版や石刻の漢字を自動認識するのは、現状では、かなり困難である。

そこで「OCR」に至る処理過程の一つとして、拓本デジタル画像からの文字切り出しに挑戦してみた。これに関しては、かなり良好な結果が得られたので、ここに報告する。

## 3.1 文字切り出しの手法

拓本デジタル画像から、一文字一文字を切り出すには、まず各行の切り出しをおこない、さらに各行の中から文字を切り出す、という手順を取る。本共同研究班でも、この手法に則ることにした。

行の切り出しをおこなうには、各行の平均的な幅を知る必要がある。本共同研究班では、拓本画像の濃淡の垂直射影分布を取って、それに対し Sondhi の自己相関関数*を 0 から順に調べていき、最初に極大値を取ったところを行幅の平均値とみなした。ただ、行幅の平均値を取っただけでは、その幅の行が画像のどこにあるのかはわからない。そこで、行幅平均値の 2 倍幅のウィンドウを想定し、そのウィンドウ内の垂直射影分布に対して、大津のフタコブラクダ関数しきい値選定法†を用いて、行と行の境目を検出した。さらに、このウィンドウを右から左へ順に動かしていくことで、全ての行の境目を検出した。これにより、全ての行を抽出することができる。

各行から文字を切り出す部分は、行の切り出しと同様の方法を、各行に対して今度は上下におこなった。すなわち、各行画像の濃淡の水平射影分布を取って、Sondhi の自己相関関数によって文字の高さの平均値を求め、その高さ 2 倍のウィンドウで大津のフタコブラクダ関数しきい値選定法を用いることにより、文字と文字の境目を検出した。

## 3.2 結果および考察

ここまでに述べた方法を、C プログラム‡で実現し、『尹宙碑』のデジタル画像に対して、文字切り出しをおこなってみた。結果を次ページに示す。ほぼ全ての文字が、完全に切り出されているのがわかる。

本共同研究班の方法により、拓本デジタル画像から文字切り出しが可能であることが確認できた。ただし、本手法は、各行が画像に対して垂直に配置されていることを仮定している。これが一般的なデジタル画像であれば、行が斜めになっていたり、あるいはスキューがかかっていることもしばしばだ。デジタル画像の質によっては、事前に補正をかけて、各行を垂直にしておく必要があるだろう。

---

*Man Mohan Sondhi: “New Methods of Pitch Extraction”, IEEE Transactions on Audio and Electroacoustics, Vol.AU-16, No.2 (June 1968), pp.262-266.

†大津展之: 『判別および最小 2 乗規準に基づく自動しきい値選定法』, 電子通信学会論文誌, Vol.J-63D, No.4 (1980 年 4 月), pp.349-356.

‡`http://kanji.zinbun.kyoto-u.ac.jp/~yasuoka/kyodokenkyu/2005-07-05/pbm2csv.c` で公開。