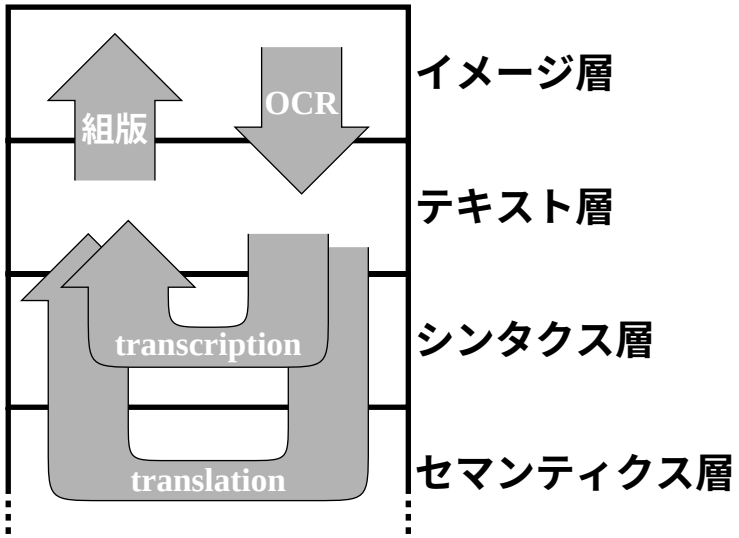
「漢字情報学の構築」共同研究班報告

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。

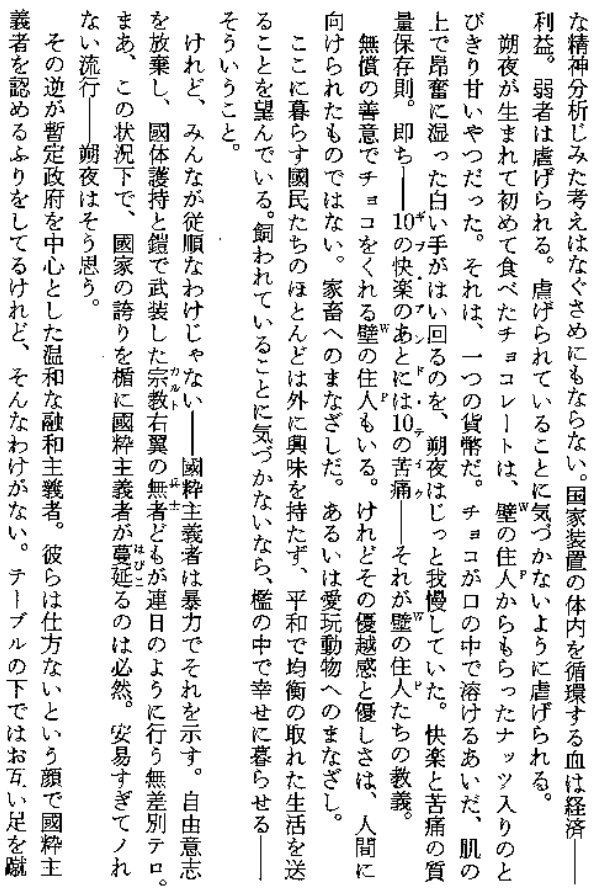
2.2 ルビ

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(やく)」というルビの振り方は正しくない。「活躍(かつやく)」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(きそく)」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(きそく)」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(きそく)」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、仮名や漢字に漢字のルビが添えられているものである (次ページに、海猫沢めろん:『零式』(ハヤカワ文庫 JA877, 2007 年 1 月) の p.35 を示す)。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。


35　第壱章「零星」

な精神分析じみた考えはなぐさめにもならない。国家装置の体内を循環する血は経済――利益。弱者は虐げられる。虐げられていることに気づかないように虐げられる。

朔夜が生まれて初めて食べたチョコレートは、壁の住人からもらったナッツ入りのとびきり甘いやつだった。それは、一つの貨幣だ。チョコが口の中で溶けるあいだ、肌の上で昂奮に湿った白い手がはい回るのを、朔夜はじっと我慢していた。快楽と苦痛の質量保存則。即ち――10の快楽のあとには10の苦痛――それが壁の住人たちの教義。

無償の善意でチョコをくれる壁の住人もいる。けれどその優越感と優しさは、人間に向けられたものではない。家畜へのまなざしだ。あるいは愛玩動物へのまなざし。

ここに暮らす國民たちのほとんどは外に興味を持たず、平和で均衡の取れた生活を送ることを望んでいる。飼われていることに気づかないなら、檻の中で幸せに暮らせる――そういうこと。

けれど、みんなが従順なわけじゃない――國粋主義者は暴力でそれを示す。自由意志を放棄し、國体護持と鎧で武装した宗教右翼の無者どもが連日のように行う無差別テロ。まあ、この状況下で、國家の誇りを楯に國粋主義者が蔓延るのは必然。安易すぎてノれない流行――朔夜はそう思う。

その逆が暫定政府を中心とした温和な融和主義者。彼らは仕方ないという顔で國粋主義者を認めるふりをしてるけれど、そんなわけがない。テーブルの下ではお互い足を蹴

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

†『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.

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

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

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

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

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。また、拓本に対しては、かなり良好な文字切り出しがおこなえたが、罫線のある刊本に対しては、本手法はあまり良い結果を出せなかった。罫線の部分が強い雑音となってしまい、行の切り出しにしくじってしまうのである。罫線のある刊本に関して

[†]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。

は、事前に罫線を消去するなどの処理を、おこなっておく必要があるだろう。

4　白文に対する自動「点」打ちの共同研究

テキスト層から一旦シンタクス層を通って、またテキスト層に戻ってくる情報処理の例として、漢文の白文に対する「点」打ちを共同研究してみよう、ということになった。というのも、本共同研究班の班長は、本共同研究と並行して『拓本文字データベース』*というプロジェクトを進めていた。『拓本文字データベース』での検索精度を上げるためには、釈文に「点」が打ってある方が良いが、拓本それ自体には、もちろん「点」など付いていない。そこで、拓本の白文に自動で「点」を打ってくれるようなプログラムがあれば、『拓本文字データベース』の役に立つし、それは今後、白文を読めない学生 (あるいは班長自身) の役にも立つと考えられる。

しかしながら結論を先に書くと、白文に対する自動「点」打ちがシンタクス層の処理だけで可能だ、というのは、そもそも読みが甘かった。現実には、セマンティクス層に立ち入る必要がどうしてもある、ということが、この共同研究の結果、明らかになった。そんな中でも、末字に注目する方法や、韻に注目する方法に関しては、シンタクス層の処理だけでも、そこそこの結果が得られた。以下、どのような方法がうまくいって、どのような方法がうまくいかなかったか、かいつまんで述べる。

4.1　末字に注目する方法

漢文には、文末にしか出てこない文字というのがある。たとえば「也」「矣」「焉」は、かなり高い確率で文末に出てくるので、これらを頼りに白文を切っていくというのは、漢文のプロでも用いている方法である。ただし、これらにもわずかながら例外があり、それにひっかからないようにするためには、やはりセマンティクス層の処理が必要となりそうだった。

4.2　頭字に注目する方法

これとは逆に、文頭に出てくる確率の高い文字もある。たとえば「嗚」「粤」「豈」などがそうだが、これらは残念ながら、白文に「点」を打つためにはあまりうまく使えなかった。これらの文字は、末字の「也」「矣」「焉」に比べると、使用頻度が 1/3 もないのだ。つまり、これらの頭字に注目して白文を切れる部分は、文全体のごくわずかな部分だったのである。

*http://coe21.zinbun.kyoto-u.ac.jp/djvuchar で公開。

4.3 2-gramに着目する方法

すでに「点」を打った文から、連続する2-gramを取り出し、それによって、白文の「点」を打つべきでない部分を求める、という方法に挑戦してみた。確かに「夫人」「春秋」「将軍」「墓誌」「邙山」など、絶対に切ることのできない2-gramは集まった[†]が、これも残念ながら、白文に「点」を打つためにはあまりうまく使えなかった。切ってはいけない部分がいくら集まっても、切っていい部分を決定することができないのだ。

4.4 韻に注目する方法

末字、頭字、2-gramなどの方法を試していく過程で、それらでは全く歯が立たない部分が、それぞれの碑文のやや後ろの方に必ず見つかった。韻文である。ただ、韻文であることがわかりさえすれば、むしろ「点」を打つのは簡単である。等間隔に打てばいいのだ。では、韻文であることを判定する方法だが、これはオーソドックスに『廣韻』[‡]を使って、同韻の字が等間隔になっているところを白文中で抽出し、その半分の間隔で「点」を打てばよい。この方法は、かなり成績が良く、手元の白文に現れる韻文をほぼ100%抽出することが可能となった。

4.5 現代中国語の文法解析を援用する方法

ここまでは、シンタクス層の処理によって「点」を自動で打つ方法を模索してきたが、どうやらそれは無理がある、ということが明らかになってきた。やはりセマンティクス層に立ち入らざるを得ない。ただ、セマンティクス層に立ち入る、と言っても、そう簡単なことではない。特に典故のあるものは、どういう典故がどういう形に変形して持ち込まれているのか、それこそ漢文に関する膨大な知識がなければわかるものではない。あるいは、漢文世界における「お約束」のようなものも、かなり難しい。たとえば「天帝使我百獣王」という一文における「天帝」など、そういう「お約束」の好例だろう。

しかし、難しいとばかりも言っていられないので、本共同研究班としては、とりあえず漢文の文法解析を、現代中国語の文法解析エンジンを援用しておこなうことを考えた。だが、これはあまりうまくいきそうになかった。漢文の語彙と、現代中国語の語彙は、かなりかけ離れており、現代中国語の文法解析エンジンでは、漢文の文法解析は困難が予想された。実際、北京大学計算語言学研究所の『漢語文本切分与詞標注』[§]に、いくつかの碑文を白文のまま入力してみたが、結果はすこぶる悪かった。

[†]対象とした釈文の大部分が墓誌だったため、かなり偏った2-gramが集まる結果となった。

[‡]`http://homewww.osaka-gaidai.ac.jp/~suzukish/inkyo/index4.htm` で公開されているWeb韻図のデータを使用。

[§]`http://icl.pku.edu.cn/icl_res/segtag98/` で公開。

4.6　返り点から漢文の構造を抽出する方法

本共同研究班は、漢文の文法解析を現代中国語の助けなしにおこなう、という目標に、立ち向かわざるを得なくなった。この目標に対し本共同研究班は、日本における漢文読解の先人達が、漢文の文法解析をどのようにおこなってきたのか、という点にヒントを求めることにした。先人達がおこなってきた手法は、大きく分けて 2 つ、音読と訓読である。これらのうち音読は、漢文の構造を目に見える形で抽出せず、むしろ、漢文を漢文のまま処理する方法である。一方、訓読は、返り点という形で、漢文の構造を抽出する。どちらかと言えば、音読より訓読の方が、現代の情報処理技術に向いていると考えられた。訓読という手法に着目するなら、コンピュータに白文を与えて、それに返り点を自動で打つことができれば、漢文の構造は抽出できたことになる。

漢文の各文字の品詞は、前後関係によって大きく変化することから、漢文の文法解析に演繹的手法を用いることは、かなりの困難が予想された。むしろ、確率的モデルを導入した帰納的手法の方が、漢文の文法解析に向いていそうである。訓点漢文を用いた帰納的手法で、白文の文法解析をおこなおうとすると、一般には以下のような手順が考えられる。

1. 大量の返り点つき例文を元に、単語辞書を作成し、単語辞書の各単語を機能別に分類して、品詞辞書とする。
2. 汎用の形態素解析エンジンを、大量の返り点つき例文にかけ、各単語間の接続確率を導出する方法によって、訓読漢文スキーマを作成する。
3. 訓読漢文スキーマと形態素解析エンジンを、対象となる白文にかけ、自動で返り点を打つ。

しかしながら、現時点の本共同研究班の手元には、わずかに 5000 文程度の返り点つき例文しかなく、品詞辞書や訓読漢文スキーマを作るには不十分だった。とりあえず、品詞辞書は日本語用のものを改造し、スキーマは上記の 5000 文からデッチあげてみたところ、完全とは言えないものの、かなり良好な結果を得ることができた。だが、この結果を自動「点」打ちにまで押し上げるには、少なくとも、大量の返り点つき例文と、十分な量の品詞辞書が必要だと考えられる。