東方學報 京都第 83 册(2008)： 360－349 頁

「漢字情報學の構築」共同研究班報告

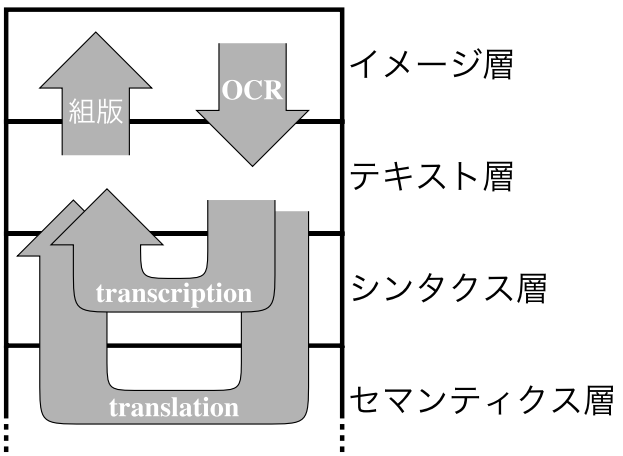
安 岡 孝 一

1 は じ め に

2004 年 4 月から 2008 年 3 月にかけて，われわれは「漢字情報學の構築」共同研究班（班長：安岡孝一）を組織し，漢字情報學なるものの構築に向けて，共同研究をおこなった。本稿では，足かけ 4 年に渡る共同研究の紆餘曲折を文書化しておくことで，何らかの意味での後世の記錄としたい。

漢字情報學を構築していくにあたり，本共同研究班では，とりあえず「テキスト情報階層モデル」なるものをデッチあげた（下圖）[1)]。漢字テキストに對して，様々な「テキスト處理」がおこなわれる中，それらの「テキスト處理」を分類・整理するために，この「テキスト情報階層モデル」を用いることにしたのである。

この「テキスト情報階層モデル」は，イメージ層，テキスト層，シンタクス層，セマンティクス層の 4 層からなる。もともと班長のアイデアでは，この 4 層をそれぞれ，畫像層，



1）安岡孝一：テキスト情報階層モデル，漢字と文化，第 2 號（2004 年 2 月），p. 6．

文字列層，文脈自由層，意味層と呼ぶつもりだったのだが，どうもアチコチがアレコレしそうだったので，カタカナ語でごまかした。このモデルに據れば，「組版」はテキスト層の情報をイメージ層の情報に變換する處理，ということになる。あるいは「OCR」は，イメージ層の情報をテキスト層の情報に變換する處理だ。また，班長が本共同研究班でやろうとしていたことの一つとして，漢文の白文に自動で「點」を打つ處理，というのがあったのだが，これは，テキスト層から一詹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處理だと考えられた。これらの大雜把な分類とモデルに基づき，本研究班では，組版，OCR，自動「點」打ち，の 3 つの情報處理に關して，共同研究をおこなうことになった。

2　組版に關する共同研究

テキスト層の情報を，イメージ層の情報へと變換する技術は，一般に「組版」と總稱されている。本共同研究班では，組版のうち日本語組版，特に縱書の組版における種々の特徵を研究し，その背後にどのような背景知識が存在するか，またそれは實際の組版上にどのように實現されるか，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，工學的技術というよりは，むしろ藝術の範疇に屬するらしい。言い換えると，組版は最終的には職人藝の世界であり，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，主觀的で個人的なものであり，それが組版に關する研究を，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

日本語組版に特徵的なルールとして，行頭・行末の禁則がある。句讀點を行頭におかない，などのルールである。行末における禁則文字は，大槪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，行末に始め括弧が來ないことを說明できる。一方，行頭における禁則文字は，多種多樣なものがあり，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，一應，意見の一致が見られるものとしては，終わり括弧，句讀點，疑問符，感嘆符，中黑などがある。これらに關しては，日本語組版において，行頭に用いるべきではない。これに對し，音引き，拗音，促音に關しては，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「句讀點」についても，その實際の處理に關

しては意見の對立が存在する。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「句讀點」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，という選擇肢がある。「ぶら下げ」る場合には，他の句讀點のうち，ちょうど行末に收まっていて元々ぶら下げる必要がない句讀點をどうするか，という問題が發生する。

さらに，文字そのものは禁則の對象となっていないが，前後關係によって，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數字であり，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，これに對しても，年號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，それを說明する前に，まずは日本語組版におけるルビ全體を概觀していこう。

2.2　ルビ

ルビは，いわゆる「振り假名」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，漢字に對して平假名あるいは片假名が添えられる[2)]。ルビは基本的には單語に對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（かつやく）」という形で，單語全體にルビを振るべきである。

複數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擧げられる。グループルビは，單語全體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擧げられる。どちらを採用するかは，やはり美的感覺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。ただしモノルビに關しては，この限りではない。

2）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，假名や漢字に漢字のルビが添えられているものである（次ページ上圖に，海猫澤めろん：『零式』（ハヤカワ文庫 JA 877，2007 年 1 月）の p. 35 を示す）。また，中國の兒童向け書籍の中には，橫書の漢字に拼音の「ルビ」を添えている例（次ページ下圖）が報告された。

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ポイントの快楽のあとには10ポイントの苦痛——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發足直前，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。藝術の範疇に屬する組版技術に對し，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 を讀んでみた。

JIS X 4051 では，たとえば行頭禁則文字に，終わり括弧，句讀點，疑問符，感嘆符，中黑のみならず，音引き，拗音，促音を含めている。また，行末處理は「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。

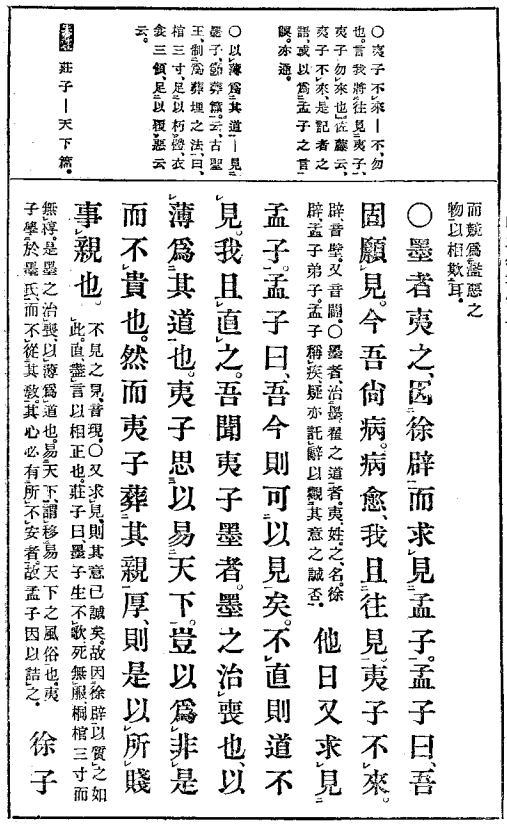
JIS X 4051 中で，本共同研究班の氣にさわったのは，これまでに述べてきた禁則やルビではなく，漢文の組版だった。JIS X 4051 の漢文組版においては，改行時の返り點を行末に置く方式になっている。しかし，實際の漢文組版においては，返り點を行末に置く方式[3]と，行頭に置く方式，さらにはそれらを混在するやり方（次頁圖）[4]がある。できれば，これらの方式を全てサポートしてほしいのだが，やはり工業規格としては，どれかを推奨しなければならないところなのだろう。

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，漢字フォントが擧げられる。というのも，東洋學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，外字作成にいつも惱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學研究者など，ほとんどいない。ただ，外字と言えど，何の典據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畫像くらいはスキャンなどで手に入るはずである。

そこで，外字の畫像からそのアウトラインフォントを，簡單に作成できるようなツールを開發した。PBM フォーマットの 2 値畫像からアウトライン情報を抽出する部分は，Peter Selinger 作の「potrace」[5]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」[6]を用いた。これにより，2 値畫像さえあれば，アウトラ

3）『漢文敎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.
4）『孟子集註』（明治書院，1940 年 11 月）.
5）http : //potrace.sourceforge.net/ で公開。
6）http : //kanji.zinbun.kyoto-u.ac.jp/~yasuoka/publications/eps2otf で公開。


滕文公章句上　一二〇

而類爲盡惡之物以相欺耳。

○墨者夷之、因徐辟而求見孟子。孟子曰、吾固願見。今吾尙病。病愈、我且往見。夷子不來。

辟、音壁、又音闢。○墨者、治墨翟之道者。夷、姓之名。徐辟、孟子弟子。孟子稱疾、疑亦託辟以觀其意之誠否。他日又求見孟子。孟子曰、吾今則可以見矣。不直則道不見。我且直之。吾聞夷子墨者。墨之治喪也、以薄爲其道也。夷子思以易天下。豈以爲非是而不貴也。然而夷子葬其親厚、則是以所賤事親也。

不見之見、音現。○又求見、則其意已誠矣。故因徐辟以質之。如此、直盡言以相正也。莊子曰、墨子生不歌、死無服、桐棺三寸而無槨、是墨之治喪、以薄爲道也。易天下之風俗也。夷子學於墨氏、而不從其教。其心必有所不安者、故孟子因以詰之。

徐子

○以薄爲其道一見墨子節葬篇。云、古聖王制爲葬埋之法、曰、棺三寸、足以朽骸、衣衾三領、足以覆惡云云。

莊子　天下篇。

○夷子不來一不、勿也。言我將往見夷子、夷子勿來也。孟子云、夷子不來、是記者之語、或以爲孟子之言、傳寫誤也。

インフォントを自由に作成できるようになった。

3　畫像からの文字切り出しに關する共同研究

イメージ層の情報を，テキスト層の情報へと變換する技術は，一般に「OCR」と總稱されている。本共同研究班では，刊本や拓本に對する「OCR」に挑戰したかったのだが，漢字を自動認識する技術は，殘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，現狀では，かなり困難である。

そこで「OCR」に至る處理過程の一つとして，拓本デジタル畫像からの文字切り出しに挑戰してみた。これに關しては，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

拓本デジタル畫像から，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， 拓本畫像の濃淡の垂直射影分布を取って， それに對し Sondhi の自己相關關數[7]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，その幅の行が畫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，そのウィンドウ內の垂直射影分布に對して，大津のフタコブラクダ關數しきい値選定法[8]を用いて，行と行の境目を檢出した。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，全ての行の境目を檢出した。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，行の切り出しと同樣の方法を，各行に對して今度は上下におこなった。すなわち，各行畫像の濃淡の水平射影分布を取って，Sondhi の自己相關關數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關數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檢出した。

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[9]で實現し，『尹宙碑』のデジタル畫像に對して，文字切り出しをおこなってみた。結果を上圖に示す。ほぼ全ての文字が，完全に切り出さ

---

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

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

9 ）http : //kanji.zinbun.kyoto−u.ac.jp/~yasuoka/kyodokenkyu/ 2005 − 07 − 05/pbm2csv.c で公開。

れているのがわかる。

本共同研究班の方法により，拓本デジタル畫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，各行が畫像に對して垂直に配置されていることを假定している。これが一般的なデジタル畫像であれば，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。デジタル畫像の質によっては，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，拓本に對しては，かなり良好な文字切り出しがおこなえたが，罫線のある刊本に對しては，本手法はあまり良い結果を出せなかった。罫線の部分が強い雜音となってしまい，行の切り出しにしくじってしまうのである。罫線のある刊本に關しては，事前に罫線を消去するなどの處理を，おこなっておく必要があるだろう。

4　白文に對する自動「點」打ちの共同研究

テキスト層から一詹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處理の例として，漢文の白文に對する「點」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，本共同研究と竝行して『拓本文字データベース』[10] 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』での檢索精度を上げるためには，釋文に「點」が打ってある方が良いが，拓本それ自體には，もちろん「點」など付いていない。そこで，拓本の白文に自動で「點」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，白文を讀めない學生（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，白文に對する自動「點」打ちがシンタクス層の處理だけで可能だ，というのは，そもそも讀みが甘かった。現實には，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關しては，シンタクス層の處理だけでも，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，

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

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處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」「粵」「豈」などがそうだが，これらは殘念ながら，白文に「點」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，文全體のごくわずかな部分だったのである。

4. 3　2-gram に着目する方法

すでに「點」を打った文から，連續する 2-gram を取り出し，それによって，白文の「點」を打つべきでない部分を求める，という方法に挑戰してみた。確かに「夫人」「春秋」「將軍」「墓誌」「邙山」など，絕對に切ることのできない 2-gram は集まった[11]が，これも殘念ながら，白文に「點」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，それらでは全く齒が立たない部分が，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「點」を打つのは簡單である。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』[12] 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「點」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。

11）對象とした釋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。

12）http : //homewww.osaka-gaidai.ac.jp/~suzukish/inkyo/index4.htm で公開されている Web韻圖のデータを使用。

4. 5　現代中國語の文法解析を援用する方法

ここまでは，シンタクス層の處理によって「點」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，そう簡單なことではない。特に典故のあるものは，どういう典故がどういう形に變形して持ち込まれているのか，それこそ漢文に關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「天帝使我百獸王」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，現代中國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，現代中國語の語彙は，かなりかけ離れており，現代中國語の文法解析エンジンでは，漢文の文法解析は困難が豫想された。實際，北京大學計算語言學研究所の『漢語文本切分與詞標注』[13] に，いくつかの碑文を白文のまま入力してみたが，結果はすこぶる惡かった。

4. 6　返り點から漢文の構造を抽出する方法

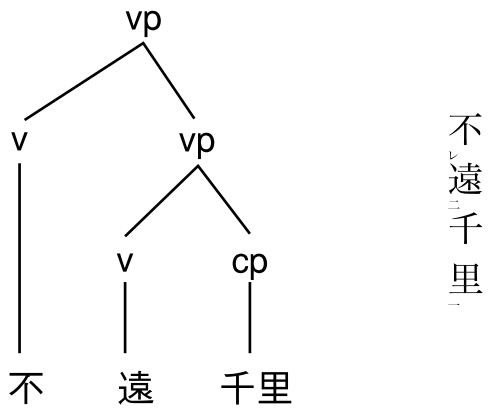
本共同研究班は，漢文の文法解析を現代中國語の助けなしにおこなう，という目標に，立ち向かわざるを得なくなった。この目標に對し本共同研究班は，日本における漢文讀解の先人達が，漢文の文法解析をどのようにおこなってきたのか，という點にヒントを求めることにした。先人達がおこなってきた手法は，大きく分けて 2 つ，音讀と訓讀である。これらのうち音讀は，漢文の構造を目に見える形で抽出せず，むしろ，漢文を漢文のまま處理する方法である。一方，訓讀は，返り點という形で，漢文の構造を抽出する。どちらかと言えば，音讀より訓讀の方が，現代の情報處理技術に向いていると考えられた。訓讀という手法に着目するなら，コンピュータに白文を與えて，それに返り點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，前後關係によって大きく變化することから，漢文の文法解析に演繹的手法を用いることは，かなりの困難が豫想された。むしろ，確率的モデルを導入した歸納的手法の方が，漢文の文法解析に向いていそうである。訓點漢文を用いた歸納的手法で，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。

13）http : //icl.pku.edu.cn/icl_res/segtag98/ で公開。

1．大量の返り點つき例文を元に，單語辭書を作成し，單語辭書の各單語を機能別に分類して，品詞辭書とする。
2．汎用の形態素解析エンジンを，大量の返り點つき例文にかけ，各單語間の接續確率を導出する方法によって，訓讀漢文スキーマを作成する。
3．訓讀漢文スキーマと形態素解析エンジンを，對象となる白文にかけ，自動で返り點を打つ。

しかしながら，現時點の本共同研究班の手元には，わずかに 5000 文程度の返り點つき例文しかなく，品詞辭書や訓讀漢文スキーマを作るには不十分だった。とりあえず，品詞辭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。



4.7　考察

だが，この結果を自動「點」打ちにまで押し上げるには，まだまだ多くの困難が立ちはだかっている。「點」を打つためには，まずは漢文の形態素解析が必要なのだが，形態素解析における曖昧性をできるかぎり少なくするためには，その深層にある構文解析が必須である。さらに，構文解析における曖昧性を少なくするには，そのさらに深層にある意味解析，ひいては，そのまたさらに深層にある文脈解析，と解析を潜っていく必要がある。

しかし，解析やその準備に對し，無限に時間をかけることはできないのだから，どこかのレベルで割り切って，一定水準の自動「點」打ちに持ち込む必要がある，ということだ。だが，その「割り切り」の見きわめと，それに必要な漢文データ量の決定にまでは，本共同研究班は至ることができなかった。

5 おわりに

本稿では，4年間に渡る「漢字情報學の構築」共同研究班の活動について，駆け足でまとめた。本共同研究班としては，漢字情報學という新たな分野に對し，一定の研究成果が得られたと確信する。もちろん，これで漢字情報學の構築が終わったわけではない。特に，自動「點」打ちプロジェクトについては，途なかばであり，2008年度より新たな共同研究班を立ち上げる豫定である。

なお，新たな共同研究班では，大量の返り點つき例文を元にして，漢文スキーマあるいは漢文コーパスを構築する，という方向の研究にとりあえずチャレンジしたい。これにより，自動「點」打ちプロジェクトを，さらに發展させることができるはずである。今後の新たな共同研究班の成果にも，ぜひ期待されたい。