「東アジア古典文献コーパスの研究」共同研究班報告

1　はじめに

2008 年 4 月から 2013 年 3 月にかけて、われわれは「東アジア古典文献コーパスの研究」共同研究班 (班長: 安岡孝一) を組織し、漢文解析に向けたコーパスの研究をおこなった。この共同研究班は、2008 年 3 月に終了した「漢字情報学の構築」班 [1] での、白文自動「点」打ちプロジェクトでの知見を受け、漢文の自動解析に特化しておこなわれた研究班である。

2　韻文の自動解析に向けて

漢文の自動解析に際し、先の「漢字情報学の構築」班の知見として、韻文と散文では全く文章構造が異なっており、それぞれに異なるアプローチを要することが、明らかになっていた。

韻文と散文の最大の違いは、もちろん、韻文が「韻」を踏んでいるという点である。すなわち韻文は、八文字、十文字、あるいは十二文字おきに脚韻を踏む、という形が一般的であり、しかもその脚韻も『廣韻』等に基づいている。このような条件から、白文の中から韻文を自動的に探し出して、文の単位で切り出すこと自体は、ほぼ 100%の精度でおこなうことが可能となった [1]。

ただ、韻文では「韻」を優先するために、語順が崩壊してしまっており、一般的な意味での形態素解析は難しいというのが、われわれの感触だった。少なくとも、散文と同一の形態素解析手法は、韻文には適用できない。したがって、韻文の自動解析においては、散文とは異なるアプローチが必要だという結論になったのである。

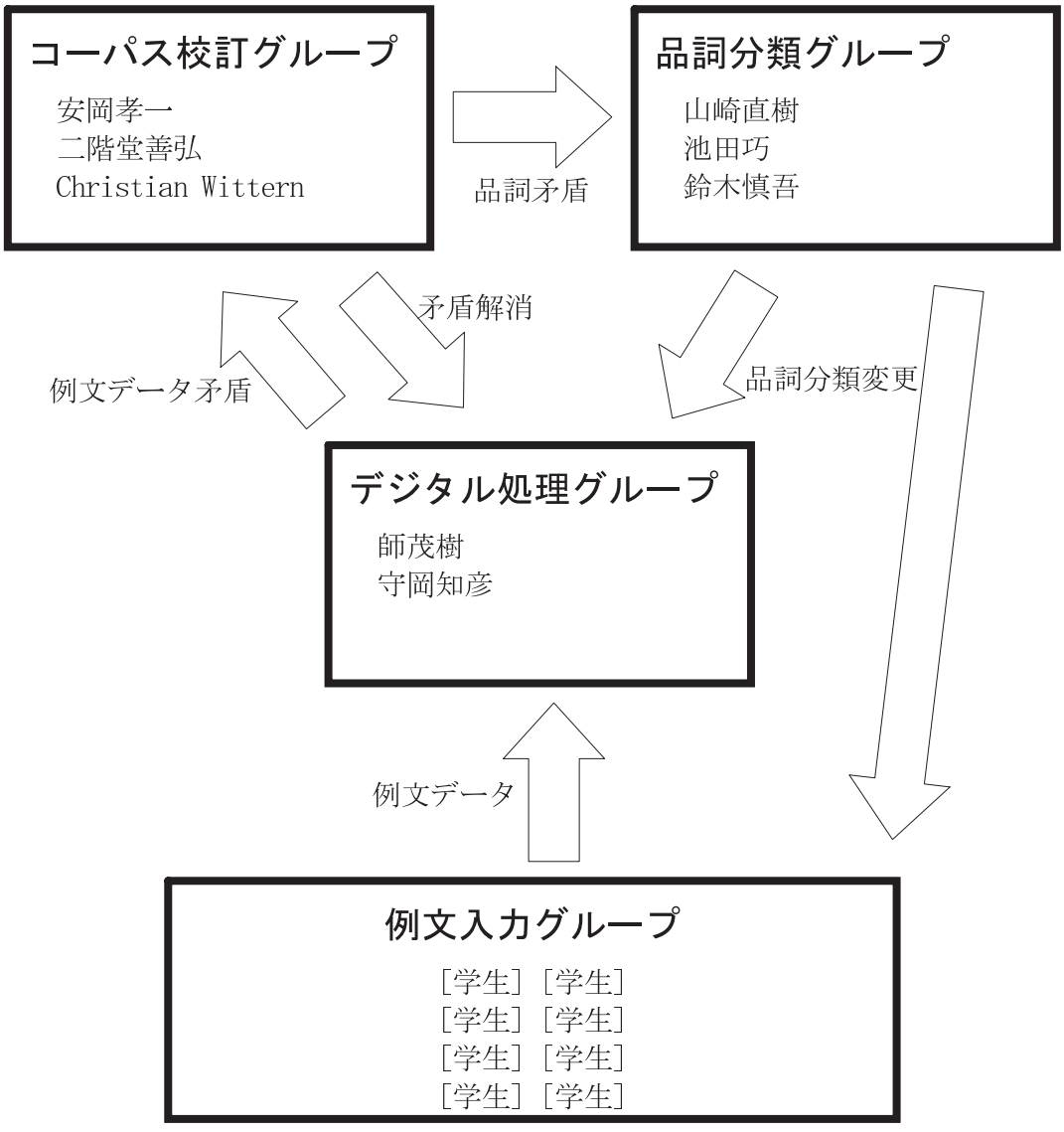
韻文の自動解析において、われわれは、対句構造に着目したアプローチを採ることにした。

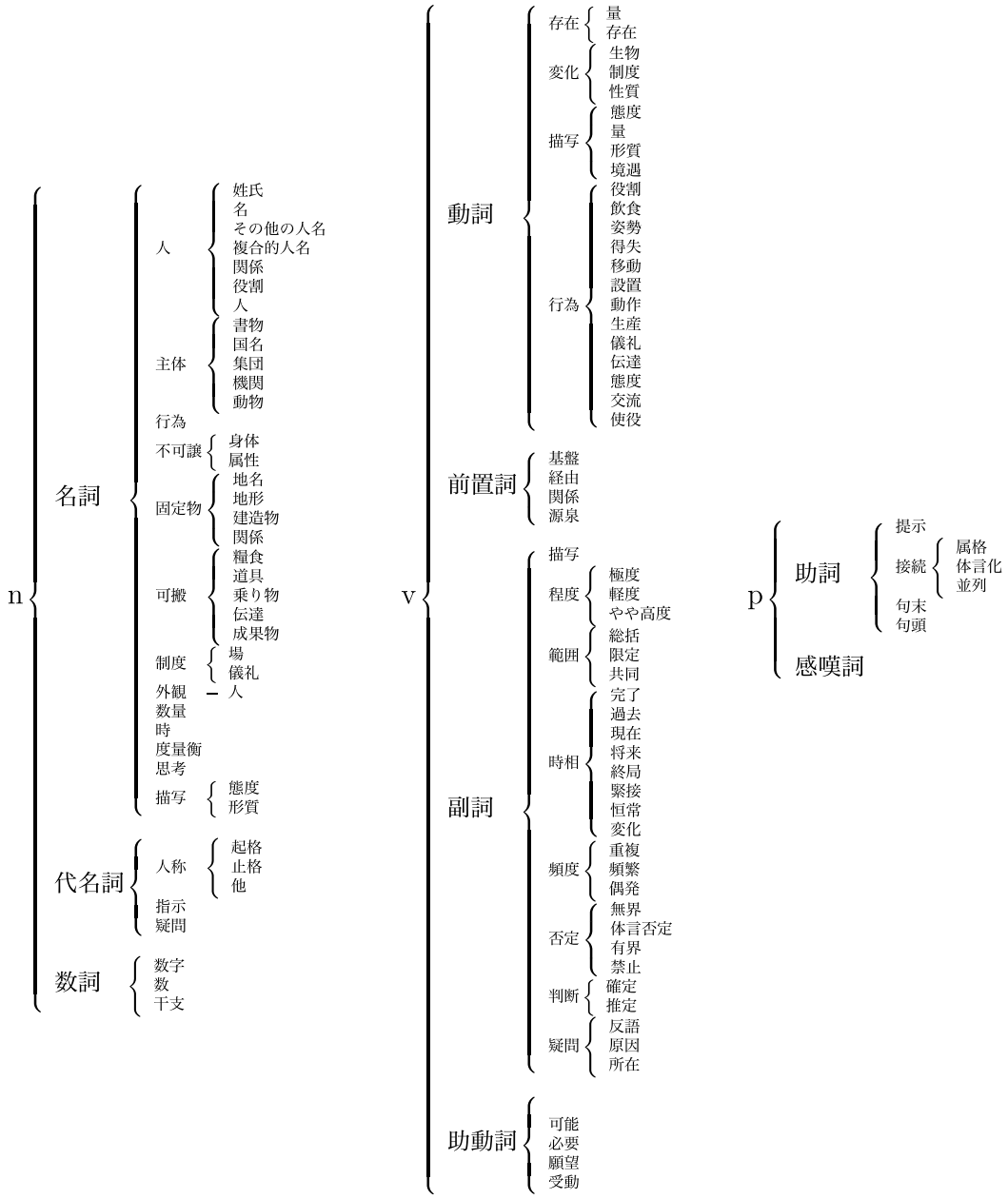
3　散文の自動解析に向けて

散文の自動解析において、われわれは、MeCab というソフトウェアを用いることにした [2]。MeCab はオープンソースの形態素解析エンジンで、言語、辞書、コーパスに依存しない汎用的な設計がなされており、辞書とコーパスを準備すればいかなる言語にも対応できる、というのが売りだった。ならば、漢文 (の散文) にも MeCab を使用できるはずだ、というのが、われわれの直感だったが、われわれ以前には誰もそれを試したことがなかった。

MeCab の辞書には 4 階層の「品詞」が必要なことから、われわれは、日本語と漢文を繋ぐ「構造」の一種である訓読に着目し、返り点を「品詞」に反映させることを考えた。すなわち、訓読における返り点を、漢文の動賓構造を表しているものとみなし、動語に「v」という「品詞」を、賓語に「n」という「品詞」を、その他の語に「p」という「品詞」を、それぞれ、MeCab 漢文辞書の「第 1 階層の品詞」(以下「大品詞」と呼ぶ) として定めることにしたのである。次に「第 2 階層の品詞」(以下「品詞」と呼ぶ) だが、これは IPA の日本語辞書から、デッチあげてみることにした [3]。「第 3 階層の品詞」(以下「意味素性」と呼ぶ) と「第 4 階層の品詞」(以下「小素性」と呼ぶ) に関しては、初期段階では付与しないことにしてみた。

この MeCab 漢文辞書 (IPA 由来版) と、それに基づいて作った MeCab 漢文コーパスを用いて、高校教科書の漢文例や、三国志呉書列伝などの白文を、MeCab で形態素解析してみた。


コーパス校訂グループ
安岡孝一
二階堂善弘
Christian Wittern
品詞矛盾
品詞分類グループ
山崎直樹
池田巧
鈴木慎吾
矛盾解消
例文データ矛盾
品詞分類変更
デジタル処理グループ
師茂樹
守岡知彦
例文データ
例文入力グループ
[学生] [学生]
[学生] [学生]
[学生] [学生]
[学生] [学生]


n
名詞
人
姓氏
名
その他の人名
複合的人名
関係
役割
主体
人
書物
国名
集団
機関
動物
行為
不可譲
身体
属性
固定物
地名
地形
建造物
関係
可搬
糧食
道具
乗り物
伝達
成果物
制度
場
儀礼
外観
人
数量
時
度量衡
思考
描写
態度
形質
代名詞
人称
起格
止格
他
指示
疑問
数詞
数字
数
干支
v
動詞
存在
量
存在
変化
生物
制度
性質
描写
態度
量
形質
境遇
行為
役割
飲食
姿勢
得失
移動
設置
動作
生産
儀礼
伝達
態度
交流
使役
前置詞
基盤
経由
関係
源泉
副詞
描写
程度
極度
軽度
やや高度
範囲
総括
限定
共同
時相
完了
過去
現在
将来
終局
緊接
恒常
変化
頻度
重複
頻繁
偶発
否定
無界
体言否定
有界
禁止
判断
確定
推定
疑問
反語
原因
所在
助動詞
可能
必要
願望
受動
p
助詞
提示
接続
属格
体言化
並列
句末
句頭
感嘆詞

4　おわりに

新規に「東アジア古典文献コーパスの応用研究」共同研究班を立ち上げる予定である。

参考文献

[1] 「漢字情報學の構築」共同研究班報告, 東方學報, 第 83 冊 (2008 年 9 月), pp.360-349.

[2] 守岡知彦: MeCab を用いた古典中国語の形態素解析の試み, 情報処理学会研究報告, Vol.2008-CH-79 (2008 年 7 月), pp.17-22.

[3] 守岡知彦: MeCab を用いた古典中国語形態素解析器の改良, 情報処理学会研究報告, Vol.2009-CH-84 (2009 年 10 月), No.3, pp.1-5.

[4] Koichi Yasuoka: Toward a Syntactic Analysis of Classical Chinese Texts, Osaka Symposium on Digital Humanities 2011 (September 2011), p.34.

[5] Naoki Yamazaki: Toward Syntactic Frame Retrieval of Classical Chinese Rhymes using Japanese 'kun' readings and Syntactic parallelism of couplets, Osaka Symposium on Digital Humanities 2011 (September 2011), p.35.

[6] Tomohiko Morioka: A Prototype of a Classical Chinese Morphological Analyzer based on MeCab, Osaka Symposium on Digital Humanities 2011 (September 2011), p.36.

[7] 守岡知彦: 古典中国語形態素コーパス編集システムの開発, 東洋学へのコンピュータ利用, 第 23 回研究セミナー (2012 年 3 月), pp.75-83.

[8] 山崎直樹, 守岡知彦, 安岡孝一: 古典中国語形態素解析のための品詞体系再構築, 人文科学とコンピュータシンポジウム「じんもんこん 2012」論文集 (2012 年 11 月), pp.39-46.