人文科学とコンピュータ　55－5
（２００２．７．２６）

漢文典故表現の抽出について

齊藤　正高

愛知大学非常勤講師

現在、ウェブ上では様々な漢文の本文データが使用に供されている。これらのデータベースでは、一つ或いは多数の本文データを同一のキーワードで検索できる。しかし、ある漢文テキストに別のテキストがどのような形で引用されているのかを自動抽出する研究は少ない。本発表では、可変長の句を持つ散文を取り上げ、明末の方以智『東西均』開章のテキストから、『論語』の引用や言い換えの表現を自動抽出した例を報告する。この作業の中で解決すべき問題、異体字問題・不要語問題・一致部分の分類の問題を、漢文の特性から述べる。

On extraction of paraphrases and similar expressions from Chinese Classics

Saito Masataka

Aichi University part time lecturer

These days there are several text data of Chinese classics on the web, and we can retrieve these data bases by keyword. On the other hand, it is rare case to study extraction of paraphrases from Chinese Classics. My study is an attempt on extract paraphrases from prose text data of Ancient Chinese. I extracted paraphrases of the *Confucian Analects*(論語) from *Tung hsi chün* (東西均) bookI written by Fang I-chih (方以智) ,a Scholar lived in China in the late Ming period. I will report some problems about unification of variants of Han characters, stop words, and term weighting.

1．研究の目的

本稿では、原文に忠実な引用やパラフレーズされた類似表現を以下、「典故表現」と呼ぶ。

本研究の目的は、２つの漢文テキストデータを用い、一方の文献の中に存在するもう一つの文献からの典故表現を自動抽出することである。これは文字コードのグローバル化によって増加傾向にある漢文テキストデータベース[1]を、相互に連係させたいという考えから出発している。

従来提供されてきた漢文テキストデータのキーワード検索[2]は、多数の文献に遍在するタームを検索し集約するというマクロな観点から見ればたいへん有益であるが、二つの文献間の引用関係というミクロな観点に限れば、データベースへの入力クエリの作成に検索者の関心が反映されるだけに、その発見に徹底を欠くことが考えられる。

このような理由から、キーワード検索とは別に、コンピュータによる典故表現の自動抽出が必要になる。この手法により、注釈のない文献を読む際の手助けになることが予測できるのである。

2．先行研究

中国語におけるテキスト検索は、現代語において Jian-Yun Nie が行った研究[3]がある。ここで提唱された、統計的アプローチと辞書的アプローチの混合手法によって、文字ベースの検索から語彙ベースの検索への道が示された。

また、古典中国語（漢文）においては、N-gram からのアプローチがある。これは、平安時代の日本語についてなされた近藤泰弘氏・近藤みゆき氏の研究[4]を受け、谷本玲大氏の漢詩文集への応用[5]、石井公成氏の提唱する NGSM[6]、師茂樹氏の般若心経の系統解析[7]、同氏の XML との連係研究[8]、山田崇仁氏の論語研究への応用[9]など、漢字文献情報処理研究会を中心に新たな試みが発表されている。

文の類似を指摘するものとしては、竹田正幸氏・福田智子氏の和歌における「本歌取り」の自動抽出の研究[10]、谷本玲大氏の「曖昧性を持たせた N-gram サーチ手法」による、『新撰万葉集』と菅原道真の詩を比較した研究[11]がある。これらは、和歌と漢詩という定型詩を扱った先駆的例である。

3．漢文にあらわれたパラフレーズの例

筆者は、中国明末安徽桐城の人、方以智の『東西均』という文献の全文データベースを作った。この書物は島田虔次氏が「難解の権化のごとき哲学エッセイ集！」[12]と述べておられた通り、随所に不明な箇所がある。近年大陸の龐樸氏が注釈[13]を発表し、その暗号めいた難解さの一端が示された。

その『東西均』開章に以下の文がある。

> 竹中之均明知無言、而何以言、因後世以不可聞者自誇其聞、嘵嘵譊譊、以傳爲市、故言其何言之行生者徵之、土型乎、鐘木乎、豈得已哉、（『東西均』開章）

下線部の「何言之行生」という部分は、以下の『論語』の文の言い換えである。

> 子曰天何言哉、四時行焉、百物生焉、天何言哉（『論語』陽貨）

この例は、『論語』の文を覚えていても、両者の関係を察知するには時間がかかった。その理由は以下に集約されると思う。

> A）三つの句の一部から構成された言い換え表現であること。（テキストの連続した部分を引いているわけではない）
>
> B）『論語』陽貨の典拠で中心となっている「天」という概念が使われていないこと。（重要な概念を中心になされた言い換えではなく、比較的目立たない言葉を集めた言い換え表現である）

このような典故表現は、原文に忠実な引用ではないために、コンピュータによって自動抽出するには、以下の方策が必要であろう。

① 検索範囲を非対称に） 分析対象は短く（例えば句単位で）切り取る必要があるが、典故をさがす文はできるだけ長く（『論語』で言えば章ごとに）切り取る必要がある。
② N-gram 検索） この例では、細切れの単語を寄せ集めて引用としている。したがって、分析対象の一句から可能なすべての N-gram を切り出し、この Ngram が典故テキストの章に存在するかどうかをすべての章にわたって調べる、「N-gram 検索」がよいと推測できる。
③ 一致した N-gram を分類する） 分析対象テキストにも典故テキストにも、多くの検索上の不要語[14]があることが予測される。そこで、何らかの不要語処理をしなくてはならない。しかし、前もって分析対象のテキストから不要語を削除することは、分かち書きの習慣のない漢文では困難である。本研究は引用を研究するものなので、分析対象ファイルと典故ファイルの中で一致した文字列を、不要語に含まれるかどうか後から判断し、全一致文字長から不要語一致長を差し引いて、内容一致長を算出している。

4．異体字の問題

具体的な検索の記述に入る前に、テキストデータの整理の問題を挙げねばならない。漢文テキストデータは作成された地域・作成した人物によってそれぞれ異なった方針で作られている[15]。本稿のように二つの漢文テキストデータの関係を調べる際には、両者の十分な校正の後で、同一と見ることができる文字が別の字としてコンピュータ処理されないように、両者の文字づかいを統一しておく必要がある。

例えば、「為」と「爲」はＪＩＳでもユニコードでも違うコードが当てられているので、両者が使われている漢文テキストデータ間では、（文献の真相を残そうという意図があるにせよ）、これらをどちらかに統一し、分析用のテキストデータに作りなおす方が、コンピュータ処理の観点からみれば便利である。

このような異体字処理を行うために、漢字同一視テーブルが公開されている[16]。しかし、日本語の文に用いることを前提として考えられているために、漢文に用いるには若干の改変を加えなくてはならない。

それは、日本語と漢文では使い分けが異なる字についての改変である。例えば、芸（うん）と藝（げい）は本来別字で漢文では書き分ける。また、「余」は人称代名詞や姓に、「餘」は「アマリ」に書き分ける。このような文字を統一してしまうと、漢文テキストデータに存在する使い分けが消滅し、テキストの分析に支障をきたす。そこで、分析用テキストデータを作成するための異体字テーブルからは、漢文独自の使い分けのある字を外しておかねばならない。

その他に、本稿ではユニコード対応としたので、「既」(SJIS:8AF9,Unicode：65E2)「旣」(Unicode:65E3)などのユニコードに於いて区別が発生する文字を異体字テーブルに加えている。

5．N-gram 検索

N-gram 検索による、典故分析は以下の手順をとった。

①分析対象ファイル断句	区切り文字「、」で句ごとに切断し、配列へ格納する。
②典故ファイル断章	改行記号で区切られている章を、配列へ格納する。
③最大 gram の算出	①と②の配列の内、大きい方の長さを最大 gram とする
④N グラムの切出し	分析句を最大 gram から 1gram まで切り出す。
⑤Ngram の一致確認	④で求めた Ngram を長いものから順に典故断章の部分と

完全に一致するかを、位置をずらしながらチェックし、一致した N-gram の内容を記録し、全一致文字数mを加算する。このとき最初に一致した N-gram の長さを最大一致長zとして記録し、最後に同一部分で再一致が起こらない様に、一致した分析ファイルの箇所を○で埋める。

⑥不要語のチェック　⑤で一致した Ngram が不要語にあるかどうかをチェックし不要語一致長sを記録する。

⑦出力　内容一致長、$m-s$が２以上のデータを出力する。

表３）N-gram サーチの例

分析句（『東西均』開章）						gram 数	典故章（『論語』述而）	一致 gram
均	罕	言	於	雅	言	6	子所雅言、詩書執禮、皆雅言也、	
均	罕	言	於	雅		5	子所雅言、詩書執禮、皆雅言也、	
	罕	言	於	雅	言	5	子所雅言、詩書執禮、皆雅言也、	
均	罕	言	於			4	子所雅言、詩書執禮、皆雅言也、	
	罕	言	於	雅		4	子所雅言、詩書執禮、皆雅言也、	
		言	於	雅	言	4	子所雅言、詩書執禮、皆雅言也、	
均	罕	言				3	子所雅言、詩書執禮、皆雅言也、	
	罕	言	於			3	子所雅言、詩書執禮、皆雅言也、	
		言	於	雅		3	子所雅言、詩書執禮、皆雅言也、	
			於	雅	言	3	子所雅言、詩書執禮、皆雅言也、	
均	罕					2	子所雅言、詩書執禮、皆雅言也、	
	罕	言				2	子所雅言、詩書執禮、皆雅言也、	
		言	於			2	子所雅言、詩書執禮、皆雅言也、	
			於	雅		2	子所雅言、詩書執禮、皆雅言也、	
				○	○	2	子所雅言、詩書執禮、皆雅言也、	雅言:不要語不一致
均						1	子所雅言、詩書執禮、皆雅言也、	
	罕					1	子所雅言、詩書執禮、皆雅言也、	
		○				1	子所雅言、詩書執禮、皆雅言也、	言:統計的不要語
			於			1	子所雅言、詩書執禮、皆雅言也、	
				○		1	子所雅言、詩書執禮、皆雅言也、	
					○	1	子所雅言、詩書執禮、皆雅言也、	

６．不要語の問題

不要語は①「機能的不要語」、②「統計的不要語」、③「経験的不要語」に分類した。

①「機能的不要語」は、テキストの内容と直接関わりのない機能語であり、これらの言葉の共有を分析しても、文法的類似点の指摘に止まり、内容的類似を指摘できない言葉である。

漢文では機能語を「助字」・「虚字」・「虚詞」などと呼ぶ。この方面で古典的な研究では、劉淇『助字辨略』や伊藤東涯『用字格』『助字考』、馬建忠『馬氏文通』などの研究がある。こうした機能語研究の中で、最近の成果として、中国社会科学院語言研究所古代漢語研究室が編纂した『古代漢語虚詞詞典』がある。これは一千八百あまりの文字や格式を扱った虚詞の辞典である。

筆者はこれをデータベース化して、その品詞を分類してみた（表１）。そして、n>1 の列に含まれる二文字以上の比較的意味の安定した虚詞を中心に不要語リストを作成した。一文字の虚詞については、上記の古典的助字研究のほかに、『新字源』など、中国古典を読むための漢和辞典に付されている助字解説を参考にして不要語リストに含めた。しかし、（否定副詞以外の）副詞や介詞のように他品詞からの転用が多い言葉は不要語に含めなかった。

②「統計的不要語」は典故テキストのなかに頻出するために、類似点として指摘され過ぎ

てしまう言葉である。

例えば、『論語』では「子」という言葉が970回余りでてくる。これは「父子」のように子供という意味でも使われているが、「子曰」のように「男子の通称」（馬融注）でもあり、先生というほどの意味でもある。代名詞ともいえないので『古代漢語虚詞詞典』には含まれていない。そこで、『論語』の、N-gram統計[17]をとって、頻度100以上の言葉について、不要語リストに含めることとした。「統計的不要語」は「機能的不要語」に含まれるものがほとんどである。しかし、「君子」や「仁」などの重要な概念も含まれている。このような「概念範疇」を不要語にしてしまうと、分析に支障がでると予測されるので、不要語にしなかった（表2）。「概念範疇」については、中国思想研究の一手法である「概念分析」の成果[18]を用い、自動的に不要語リストから外すこともできるだろうが、本稿執筆時ではこの作業を自動化していない。

③「経験的不要語」は、比較作業のなかから、経験的に不要語に含めた方がよいと考えられる言葉である。これには、一文字の数詞（一～九・十・百・千・萬）を挙げることができる。これらの数詞を不要語に含めないと、数値が書かれている部分に、すべて類似を指摘してしまう。だが、「道生一、一生二、二生三、三生萬物」（『老子』四十二章）のように、象数的用法がある文献の分析では、不要語に含めるかどうかを判断する必要がある。

①～③の不要語リストは頻度順に並べかえれば高速化できると考えたが、漢文の語彙統計については、一般的な頻度表が見あたらないので、高速化については今後の課題として残した。

表1）『古代漢語虚詞詞典』の虚詞の分類

品詞	総数	n>1	説明
副詞	1003	390	必・本など、（他品詞の虚詞用法が多い）
固定格式	271	271	與其～不如など（呼応する文型）
連詞（接続詞）	150	111	而・則など
慣用表現	149	149	由是のように虚詞を組合せてつかう表現
代詞（代名詞）	133	67	此・或・何など
介詞（前置詞）	107	14	於のように虚詞用法しかないものの他に、道のように他品詞からの転用もある。
助詞	55	2	唯・者など
語気詞	48	17	句末に付く、焉・矣など
感嘆詞	43	19	嘻・嗚呼など
助動詞	42	16	得・敢など
数詞・時間詞など	18	14	許多・幾など
総計	2019	1070	

表2）『論語』の統計的不要語（文字/頻度）

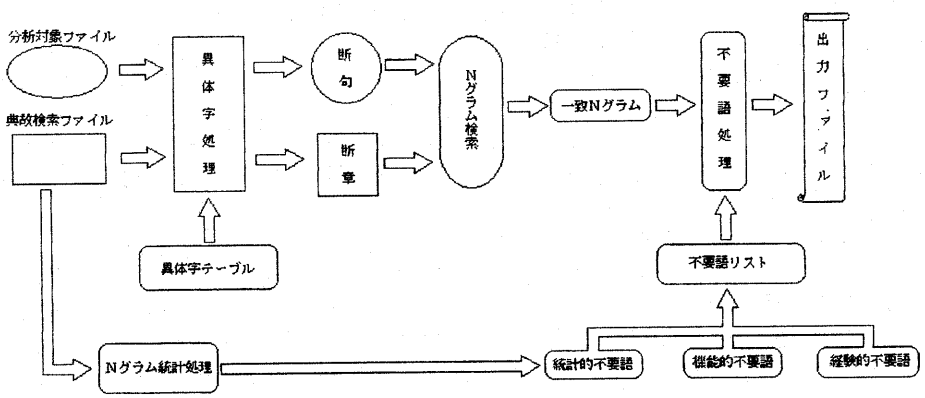
子/975	曰/756	之/612	不/582	也/532	子曰/454
而/341	其/266	人/219	者/219	以/208	有/201
矣/182	於/177	爲/170	乎/159	君/158	可/156
如/153	與/142	言/129	無/127	則/123	問/119
何/118	知/116	吾/115	仁/110	君子/109	夫/101

（下線部は、頻出語ではあるが、概念として重要なので不要語から除外した。）

7．漢文典故分析ソフトの開発

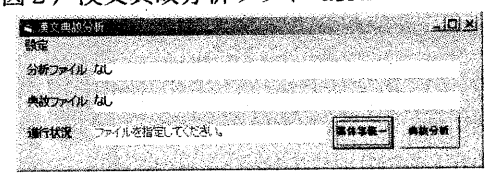
これまで指摘した観点をまとめると、以下の様になる。

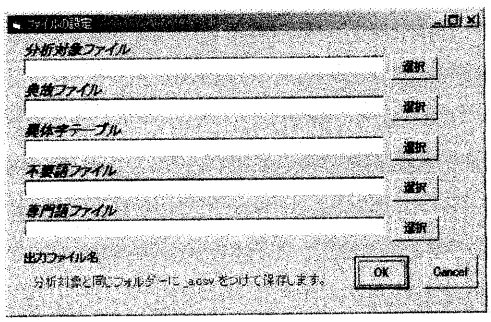
図１）　作業の流れ



上の作業を行うための簡単なソフトを、Visual Basic6.0で開発した。

図２）漢文典故分析ソフト HDFX





このソフトウェアの性能は以下の通り。

①ユニコード（2byte文字のみ）のリトルエンディアン対応

②分析句２００文字、典故章１０００文字まで対応

③必要なファイル（すべてユニコード、リトルエンディアンで保存する必要があり）

- 分析対象ファイル
- 典故検索ファイル
- 異体字統一テーブル
- 不要語ファイル
- 専門語ファイル（執筆時未対応）

④出力ファイルの書式　カンマ区切りレコード

分析句	典故章	一致グラム/一致グラム/	全一致文字長	最大一致長	不要語一致長	内容一致長

8．結果

図２のソフトを使い、『東西均』開章と『論語』の分析を行った。概要は以下の通り。

a．『東西均』開章　553句　筆者が入力
b．『論語』　550章　道気社のファイルをもとに校正、長文は断章を変更した。
c．組み合わせ　a×b = 30,4150組
d．一致行数　1681行（内容的一致が２以下の行は出力しなかった。）
e．抽出率　d÷c×100 = 0.55%
f．分析時間　３時間１０分　(Celeron500MHz)

出力した結果を、内容一致長の降順にソートすると、典故表現を確認できる。また、最大一致長の降順にソートすると、比較的長い一致が見られた引用表現を確認できる。結果の分析にはテキストデータや印刷された原文に戻って、テキストの脈絡を把握しなおす必要がある。以下に一致例を挙げる。数値は全一致数以下の４つの数値（前節④を参照）

g．引用箇所の抽出（以下●は『東西均』開章、○は『論語』括弧内は篇名）

●吾道一以貫之與一陰一陽之謂道、三一者、一一也、何謂吾、何謂道、何謂一、曾疑始否、曾同異否

○子曰、參乎、吾道一以貫之、曾子曰、唯、子出、門人問曰、何謂也、曾子曰、夫子之道、忠恕而已矣、（里仁）11,6,3,8

※破線部は『周易』繫辞傳上からの引用

●吾言無所不説者亡矣

○子曰、囘也、非助我者也、於吾言、無所不説、（先進）7,4,2,5

※「吾が言に於いて…」と記憶していると見逃すかもしれない。

h．類似表現の抽出例

●彌綸乎大一而用萬卽一之一、知之樂之、眞天不息、而容天下

○子曰、知之者、不如好之者、好之者、不如樂之者（雍也）4,2,0,3

●均罕言於雅言、使人自興自鑒自嚴自樂而自得之、以其可聞聞不可聞、吾言無所不説者亡矣

○子貢曰夫子之文章、可得而聞也、夫子之言性與天道不可得而聞也（公冶長）6,2,2,4

i．慣用表現の抽出例

●日一南而萬物死、一歲之生死也、時在其中矣

○父爲子隱、子爲父隱、直在其中矣（子路）4,4,0,4　以下同じ

○飯疏食飲水、曲肱而枕之、樂亦在其中矣（述而）

○君子謀道不謀食、耕也餒在其中矣、學也祿在其中矣、君子憂道不憂貧（衛靈公）

○言寡尤、行寡悔、祿在其中矣（為政）

○子夏曰、博學而篤志、切問而近思、仁在其中矣（子張）

9．まとめと今後の問題

『論語』は古典であり、漢文を書いた人々はこれを暗誦していたので、様々に変形し自らの文に用いている。そのような表現をコンピュータで自動抽出してみたが、この様な事実はテキストを読めば自然に分かる場合も多い。もとより機械的な手法が、優れた学者の注釈を凌ぐものではないのは、論ずるまでもない。だが、漢文を読むときに、さりげない表現の中に古典のパラフレーズが隠れていることを見逃すことも少なくないであろうし、遺漏もあり得る。機械的な手法は、徹底的に文章の関係を探りだそうという試みである。その９割は何ら関係のないデータを出力するが、なかには解釈の余地を発見することもある。また、テキ

ストの生態学として、古典の引用のされかたを歴史的に跡づける研究にも威力を発揮するかもしれない。本研究で取り上げた手法をより精密にしていくには、不要語の精度の問題・様々なタイプの漢文テキストデータの整備・統計データ整備・出力結果の分析法の確立など、解決せねばならぬ問題も多い。特に大量のデータを分析する際には、漢文を読み判断する能力が更に必要となることを指摘しておきたい。ソフトウェアの性能面では今回の比較に３時間程かかったので気楽に分析にかけてみる訳にはいかない。高速化が第一の課題である。

（了）

執筆にあたり漢字文献情報処理研究会の活動に裨益を受けました。記して感謝申し上げます。

---

1 Web 上の漢文データベースサイトは、遺漏もあるが、代表的なサイトとしては以下がある。
・台湾中央研究院漢籍電子文献　(http://www.sinica.edu.tw/ftms-bin/ftmsw3)
・寒泉　(http://210.69.170.100/s25/index.htm)
・大正新修大蔵経データベース　(http://www.l.u-tokyo.ac.jp/~sat/japan/index.html)
・電子達磨
(http://www.iijnet.or.jp/iriz/irizhtml/irizhome.htm)
・道気社　(http://www.zinbun.kyoto-u.ac.jp/~dokisha/)
・Chinese Philosophical E-text Archive　(http://sangle.web.wesleyan.edu/etext/index.html)
2 注１の台湾中央研究院や寒泉などはテキストデータをダウンロードするタイプではなく、検索型のデータベースである。
3 Jian-Yun Nie, On Chinese Text Retrieval, SIGIR'96 Zurich, 1996
4 近藤泰弘・近藤みゆき「N-gram の手法による言語テキストの分析方法」,『漢字文献情報処理研究』第２号所収, 2001
5 谷本玲大「曖昧検索性を持たせた N-gram サーチの手法」同前所収, 2001
6 石井公成「N-gram 利用の可能性」同前所収, 2001
7 師茂樹「N グラムモデルとクラスター分析を用いた漢文古典テキストの比較研究－『般若心経』の異訳を例に」(http://ya.sakura.ne.jp/~moro)
8 師茂樹「XML と NGSM によるテキスト内部の比較分析実験－『守護国界章』研究の一環として」,『漢字文献情報処理研究』第２号所収, 2001 年
9 山田崇仁「初めての N-gram」同前所収. 2001 年
10 竹田正幸・福田智子「類似歌を探せデジタル国文学の新展開」『日経サイエンス』2002/5
11 注 5 参照
12 島田虔次『朱子学と陽明学』岩波新書, 1967 年 P184
13 方以智『東西均』は東西均記によると 1652 年前後に書かれた文献である。この著作は長く安徽省博物館に抄本があるだけだったが、李学勤氏の整理を経て、1962 年中華書局が標点本を出版した。この 1962 年標点本の跋には李学勤氏によって注釈が計画されていたが完成しなかったことが書かれている。その後、この文献に初めて注釈を施した著作が龐樸『東西均注釈』中華書局 2001 年である。筆者の本文データは 1962 年本を基礎とし、龐樸氏の注釈を参考にしている。
14 不要語(ストップワード)処理については、神門典子・清水美都子・橋爪宏達・山本毅雄『全文検索技術と応用』丸善出版社, 1998 年 p39、北研二・津田和彦・獅々堀正幹『情報検索アルゴリズム』共立出版 2002 年 p30 を参照した。
15 このため各国での文字の同一視問題や、誤字の傾向などの問題である。誤字は、「李」と「季」等の単に字形に起因するもの他に、日本語入力では異音であるが、中国語入力では同音となるために誤字となる、「君」と「軍」(ともに、jun1)等がある。
16 Qgrep 用の漢字同一視テーブルが Vector (http://www.vector.co.jp) 提供されている。異体字テーブルの利用については、注 5 の谷本氏が言及している。
17 師茂樹氏の morogram を使用した。(http://ya.sakura.ne.jp/~moro)
18 張岱年『中国古代哲学概念範疇要論』中国社会科学院出版社, 1989 年など