N-gram 方式を利用した漢字文献の分析

山　田　崇　仁

１．はじめに

本稿は、「N-gram 方式を漢字文献の分析に用いる」という筆者の研究手法について、その概念・期待される効果・問題点について説明・紹介する事を目的とする。
本稿は、主に三つの部分で構成される。

１．始めに N-gram そのものを初めとする基本的な用語について説明する
２．次に、N-gram によって得られた情報を分析する手段を幾つか紹介し、各手法の特徴と期待される効果、そして N-gram 方式の限界と問題点について説明する。
３．付録として、N-gram 方式を利用して漢字文献を数値化する手順・ツールを説明する。

本稿の執筆動機として、「N-gram 方式」を用いた研究手法に対する、過度の期待や誤解を解消したいという思いが背景にある。

筆者の周辺では、「N-gram 方式」に対し、あたかも魔法の道具であるかのような過度の期待や、統計手段等を用いた分析方法に対する誤解があるように感じられる。

N-gram 自体は、自然言語学や計算言語学で用いられている極単純な言語モデルである。後述するように、単純さ故の問題は多いものの、このモデルを用いた多くの研究成果や製品が公表され続けている事からも、N-gram の有効性は既に明白である。

筆者自身も、N-gram 方式を漢字文献に用いる事への利点や問題点につき、公表済みの複数の論文に明記してある。しかし、相変わらずこの手の誤解が解けないのが現状である。

一つには、論旨が複数論文に分散している関係上、どうしても印象が散漫になる事。そして、人文の分野から提起される問題に対し、他分野の方法論（本稿では N-gram 方式が相当）を取り入れて解決を図るという、手法そのものに不信感があるためではないかと考えている。

これらの疑問点を解消するには、やはり自らの手で N-gram 方式を扱ってみるのが早道だろう。その為、以前から多くの研究者が N-gram 方式を試せるように、手法や方法論そのものの解説をしたいと考えていた。本稿は、その作業の一環に位置付けられる。

2．N-gram 方式について

2．1　N-gram とは何か？

本稿で基軸となる「N-gram」という概念は、情報理論の創始者として名高いクロード・エルウッド・シャノン（Claude Elwood Shannon 1916–2001）が想起したものである。(1)

「N-gram」の定義は、「あるテキストの総体を前から順に任意のＮ個の文字列または単語の組み合わせで分割したもの」である。また、Ｎ個の数（gram）に応じて、それぞれ「1(uni)gram, 2(bi)gram, 3(tri)gram…」と呼ばれる（本論では漢字１文字を１個として数える。以下同じ）。

2．2　単一文献に N-gram 方式を用いる

上記の定義ではわかり難いと思うので、実際の文章（『論語』学而篇第一）を事例に説明しよう。

学而篇の冒頭は「子曰、學而時習之、不亦説乎。」である。この文を二文字単位（＝ 2gram）で分割すると、以下のようになる（表１：句読点は除く）。

表１：『論語』冒頭の 2-gram で分解

子	曰									
	曰	學								
		學	而							
			而	時						
				時	習					
					習	之				
						之	不			
							不	亦		
								亦	説	
									説	乎

N-gram では「子曰」「曰學」「學而」のようなＮ個を単位とする文字列・単語の組み合わせを「共起」関係にあるとする。(2)またテキスト全体での任意の「共起」が現れる回数を「共起頻度」と呼ぶ。

「共起頻度」の例として、『論語』全篇を対象に１～ 5gram で共起頻度を集計したものから、それぞれの頻度上位10位までをあげた（表２）。

元となったテキストは、岩波文庫の金谷治校訂本を筆者がデジタルテキスト化したものである。また、デジタルテキストは、句読点や括弧等について全て削除したいわゆる「白文」状態のものを作成した。そのため、3gram の「也子曰」のように、表現として意味を持たない可能

性が高い文字列も「共起」に含まれている。

N-gram では、これら意味を持たないと判断される共起を「ノイズ（ノイズデータ）」と呼ぶ。しかし、視点によっては、これらの「ノイズ」的「共起」が有用な情報を持つと判断される可能性もあるので、ここでは句読点を跨ぐ「共起」を集計対象とした。

表２：共起頻度の例

1gram		2gram		3gram		4gram		5gram	
共起	共起頻度	共起	共起頻度	共起	共起頻度	共起	共起頻度	共起	共起頻度
子	984	子曰	458	也子曰	73	子曰君子	41	也子曰君子	8
曰	761	也子	113	曰君子	51	孔子對曰	13	不好學其蔽	6
之	614	君子	110	子曰君	41	何如子曰	10	好學其蔽也	6
不	585	而不	71	矣子曰	38	已矣子曰	9	於孔子孔子	6
也	577	孔子	68	孔子曰	31	也子曰君	8	子曰君子不	5
而	343	矣子	60	子貢曰	26	也已矣子	8	子貢曰夫子	5
其	271	曰君	55	乎子曰	21	孔子孔子	8	矣子曰君子	5
者	226	子路	47	子曰不	21	者也子曰	7	不知其仁也	4
人	220	夫子	46	子曰吾	20	不好學其	6	乎對曰未也	4
以	211	子貢	43	之子曰	19	也子曰不	6	也已矣子曰	4

表２を一通り眺めるだけでも、いくつかの特徴を見出す事ができる。

例えば、2gram の列からは、「『論語』は「子曰」を頻用し、弟子の中では「子路」と「子貢」が多く登場する」傾向が、また。4gram からは「『論語』では孔子（「子曰」）が「君子」について述べるケースが多い」傾向が見いだせるだろう。

上の例のように、あるテキストを任意の gram 単位で分解し（本稿では「N-gram 方式」と呼ぶ）、得られた共起とその頻度を数値化してその結果を集計する事で（本稿では「N-gram 統計（をとる）」と呼ぶ）、「よく使われる表現から推測されるテキストの傾向」のような、テキスト持つある側面を把握する事ができる。

２．３　NGSM で複数文献を比較

上例２．２は、単一の文献を対象とした。この場合、当該文献内での頻度傾向を確認する事は可能だが、その結果をどう意義づけるかについて判断するには、複数文献との比較が求められる。

複数文献間を対象とする N-gram 統計の比較方式は「NGSM（N-gram Based System for Multiple Documents[3]）」と呼ばれる。

実際の例を挙げてみよう。表３は、『論語』・『孟子』・『墨子』・『莊子』・『荀子』・『韓非子』を

対象に2-gramでN-gram統計をとり、そこから共起を抽出したものをNGSMで一覧表示したものである。

実際には、各テキストののべ文字数が異なるため、単純に共起頻度のみを比較する作業は厳密には有効なものではないが、それでも『論語』では頻度上位である「子貢・子路」の二人が、『孟子』以降殆ど見えないといった特徴がある事を了解されるだろう。

筆者は、複数文献間を NGSM で比較する際、通常二通りの方法を用いる。

一つめは、「複数文献間で共通する共起用例を探す」方式である。これは上述の単純比較と同義であるが、簡単に全体の傾向を把握する場合や、4, 5gram で共通する共起用例（＝引用関係が想定される）を探す際に使用する場合が多い。

二つめが「共起頻度を正規化する事で同一化し、それを統計的手法で比較する」方法である。これは、「各テキストの文字数が異なるために単純比較を行えない」という問題を解消するために用いる。その方法として、相対度数（やパーセンテージ）等の手段で各テキストを正規化し、それをグラフや統計的手法で比較し、テキストの特徴や位置づけを見いだす際に用いる。

では次に、N-gram を利用して単一・複数のテキストから共起頻度を収集し、その一覧を取得する手順について説明する。

表 3 ：諸子文献を NGSM で一覧表示

共起	論語	孟子	墨子	荘子	荀子	韓非子	全体
子曰	458	293	45	40	21	3	860
君子	110	82	51	3	174	1	421
而不	71	114	92	80	197	31	585
孔子	68	81	0	8	6	5	168
曰君	55	20	1	0	10	0	86
子路	47	6	0	0	0	0	53
子貢	43	7	0	6	0	3	59
不可	42	84	103	22	101	45	397
夫子	42	35	0	22	0	0	99
對曰	39	30	0	1	1	2	73

3 . N-gram 方式を用いた漢字文献の分析

3 . 1　デジタルテキストとしての漢字文献と N-gram

3 . 1 . 1　N-gram 方式のメリット：単なる用例検索を超えて

通常、文献の特徴を探る場合、文献に見られる用例を抽出して分析する作業を必要とする。その際、如何に「その文献の特徴を表す表現（キーワード)」を見つけるかが鍵となる。より

有効な「キーワード」を見出すには、言うまでもなく「文献の読みの深さ」に関する能力が求められる。しかし、何を「キーワード」として評価するかは、個々の読み手の関心や問題意識に規定されるという限界がある。

例えば、A 氏が問題とする「キーワード」を B 氏は問題としない。あるいはその逆。更には両者が問題としない「キーワード」が、実は重要なものである可能性すらある。

「従来問題とされてこなかったものが実は問題であった」事を明らかにするのが研究の一つの醍醐味である。しかし、「分析者が注目しない用例」を発見する事は不可能である。

このような「キーワードから捜す」手法の限界に対し、N-gram 方式の「文献に現れる全ての文字の組み合わせを収集する」という特徴は、この限界を克服する可能性を秘めている。

N-gram 方式で得られる共起用例は、「文献が内包する全ての文字の組み合わせ」である。当然、その中にはノイズデータも多い。しかし、得られた共起用例の情報を絞り込む過程で、従来見落としていた意味のある表現（キーワード）を発見し得る可能性がある。

そのため、N-gram 方式は、文献の中から特定の用字パターンを抽出するのに大変有効な方法である。また、大量の共起頻度を統計的手法を用いて分析する事で、文献の傾向を明らかにするというやり方も可能である。

3．1．2　どの頻度で共起頻度を収集したらよいか？

漢字は一文字毎に言葉の概念を包含するため、「表語文字」とも呼ばれる。

そのため、ひらがな・カタカナ・アルファベット等とは異なり、言語の最小構成要素である「形態素」の多くが「１文字」で構成されるという特徴がある。

そのため、他の言語なら余り意味のない 1gram 単位の共起頻度（＝単漢字頻度）でも有効な結果が得られる反面、4gram 以上の N-gram 単位で同様な作業を行った場合、ノイズデータが格段に増加するため、意味のある共起の発見が難しくなると言うデメリットもある。

従って、漢字文献で N-gram 方式を利用する場合、一般的には少ない gram 数を単位とした方がノイズにならない用例をより多く集められる可能性が高い。

但し、多い gram 数を単位の N-gram 統計も、決して無意味ではない。例えば、NGSM で複数文献を比較する場合に有効である。それは、NGSM で複数文献間に一致する確実に意味のある共起用例を発見した場合、その文献間に引用関係が存在する事が確実だからである。

筆者は、経験的に質・量の両面で最も有効なデータが得られる 2gram[(4)] をデータ分析の基本としている。また、補足情報として単漢字を対象とした 1gram のデータや、複数文献間の引用関係の発見を目的とした 3 〜 5gram のデータ収集を行っている。

3．1．3　N-gram 自体が内包する問題

N-gram は単純な仕組みでありながら、それなりに有効な結果をもたらすため、全文検索や

OCR と言ったデータを機械的に処理する事を前提とするシステムでよく使用される。(5)

しかし、N-gram は決して万能の道具ではない。仕組みが単純な故の問題もある。

ここでは、三つに分けてその問題を論ずる。その内、一つ目は、現在のコンピュータで文字を扱うための仕組みである「文字コード」そのものに起因し、また、後の二つは、N-gram という仕組みそのものに起因する問題である。

■文字コードに起因する問題

そもそも N-gram 方式は、「文字コード」と大変親和性が高い。なぜなら、文字コードでデジタル化されたテキストは「データの先頭から末尾に向けて文字が一次元に配列されている」状態であり、N-gram 方式は、この一次元配列に対し「先頭から N-gram 単位で文字を分割収集する」という手順で実行されるからである。従って、現状のデジタルテキストの面から言えば、N-gram は大変効率がよい方式と言える。

ところで、漢字は一文字毎に「形・音・義」等の複数の構成要素を含んでいるとされる。しかし、現在の文字コードは、極論すればその中の「形とその並び」にのみ特化したものでしかない。

そのため、漢字文献をデジタル化する事で欠落する情報について、当然、N-gram 方式ではそれ等を収集する事が難しくなる。例えば、「異体字」等の「異字同義」を同一の共起として、或いは「同字異義」を別個の共起として収集する事ができない。

■ 0 頻度（データ・スパースネス）問題

「 0 頻度問題」とは、「N-gram 統計をとった結果、ある共起の頻度が 0 となった場合、それをどう評価するか」についての問題である（北1999・松本他2004も参照されたし）。

一番単純な処理は「頻度 0 ＝その共起は存在しない」と判断する事である。しかし、実際にその共起が本当に存在しないか否かを判断するのは、実は難しい問題をはらんでいる。

例えば、複数文献を対象にした NGSM の結果である共起の頻度が 0 となった場合、NGSM の対象とならなかったテキストにその共起が存在する可能性は皆無ではない。その場合、「NGSM 結果にはその共起が存在しない」のか「それ以外の文献にもその共起が存在しないのか」を分けて考える必要がある。

この様な場合が考えられるため、N-gram 方式を用いる場合、できるだけテキストの量・質が整ったものを使用する必要がある。また、「頻度 0 」となった場合でも、他の文献に存在する可能性を考慮に入れ、適宜他のテキストデータベース等で検索する事が望ましい。

これらの補正を経た上で、なお「頻度が 0 」となった共起の存在は、情報として重要な意味を持ってくる。

例えば、皇帝の避諱に係る文字が存在するか否かが、漢字文献の版本が刻された時代を判断

するのに有効な情報となる事はよく知られている（ある文献の「特定避諱字の頻度が 0 」かつ「避諱字の代替字の頻度が 1 以上」であれば、その文献の版刻された時期の特定が可能）。それ以外にも、ある時期や地域で頻用された表現の有無を比較分析する事で、文献の成書時期や地域を確定可能である等の成果を得る事ができる。

■大量のノイズをどう扱うか

これまでに何度か述べたように、N-gram 方式では言葉として意味のある共起よりもむしろ、意味のない共起の方が大量に集計される。特に gram 数が多くなればなるほど、共有される共起が減るのと反比例して、他と共有されない頻度 1 の共起が大量に出現する。そして、その多くは、「意味を為さない共起＝ノイズデータ」に属する。

そのため、N-gram 方式で効率よく情報を抽出するには、ノイズデータをどう処理するかが問題となる。

一つは「とりあえず共起用例を全て収集し、そこから情報を絞り込む」方法である。この場合、分析者が有効と判断する共起を抽出する事になる。ノイズに対する判断基準はデータの分析者によって異なるため、全ての共起を収集する方式をとる事で、従来気づかれなかった有効な情報を発見可能という利点がある。しかし、大量の共起が得られた場合、整理と抽出に時間がとられてしまうというデメリットもある。

二つめは「ある程度の絞り込みをする」方法である。例えば、あらかじめノイズデータになりがちな句点や章を跨ぐような共起を削除して収集する仕組みを作っておけば、ノイズデータが相当減少するはずである。しかし、漢字文献では文章の区切り自体が問題となる場合もあるので、その場合は注意が必要となる。この方法の場合、情報の絞込基準（抽出したい情報の大枠）を設定しておく事が必要となる

また、応用例として、「複数文献を比較する場合、主たる比較対象に無い共起は全てノイズと見なして削除する」方法もある。この場合、相当にデータが整理されるが、上述の「 0 頻度問題」に注意する必要がある（ある共起がたまたまその文献にないのか、その時代・地域にないのか等を判別する必要が求められる[(6)]）。

筆者は主に最後の方法を使い、分析対象に応じて他の方法を使っている。

3．2　N-gram 方式を用いた漢字文献分析の実例

N-gram 方式自体は、情報収集の一手段にしか過ぎない。実際には、収集した情報をどう分析するかが重要である。N-gram 方式を用いた人文学的研究の場合、その分析方法は以下の三種に分かれる。

●パターンマッチング法

●統計的処理
●複数の手法を組み合わせる

以下、それぞれの方法について、実際の研究成果を交えながら紹介する。

3．2．1　パターンマッチング法

「パターンマッチング法」とは、対象文献に対し、ある程度の大きさの gram 数で N-gram 統計をとり、その結果得られた頻度が多い共起用例につき、文献上の使用状況を調査分析する事で対象文献の特徴を解明する方法である。

3．1．1 で述べたように、N-gram 方式では、従来の「初めに用例ありきの方式」では見いだせなかった文献の特徴を見いだせる可能性を持っている。その特徴を利用したのが、パターンマッチング法である。

この方法は、近藤みゆき氏の研究を嚆矢とするが、氏の研究は、N-gram 方式自体を人文学、特に古典学の分野に導入したという面でも、重要な研究史的意義を持っている。

近藤氏は、『古今和歌集』に対して N-gram 統計をとり、その分析結果から特定の表現が男性・或いは女性にのみ偏って出現する事を見いだした。それによって、従来着目すらされてこなかった、和歌の表現形式にジェンダー性が存在する事を明らかにしたのである（近藤みゆき2000）。

筆者は、佚文文献の収集にこの方法を用いた。

これは、「既存の文献 A（『國語』韋昭注）に別な文献 B（『世本』）の佚文が含まれている事が明らかだが（序文に明示）、それがどの部分かわからない。」という問題の解決に、N-gram 方式とパターンマッチング法を用いたものである（山田2001）。

まず、他の文献（『史記索隠』『五経正義』等）に含まれる文献 B の佚文を抽出し、それを対象に N-gram 統計をとった。次に、その結果から文献 B の記述パターンの特徴を分析し、それに基づいて文献 B の記述構造を定義した。最後に、その定義に基づき、文献 A から文献 B に相当すると推定される文字列を抽出し、それを既存の文献 B の佚分と照合する事によって、文献 B の佚文であると確定した。この作業により『國語』韋昭注所収の系譜的記述が、『世本』佚文として使用可能となった。

3．2．2　統計的処理

「統計的処理」とは、N-gram 統計で得られたデータを統計的手法で分析し、その結果からテキストの特徴を明らかにするものである。

N-gram 方式は、テキストのサイズと gram 数の増加に応じて、共起の総量が爆発的に増加する。従って、人間の目でそこから有効なデータを抽出する場合、量的・時間的に限界が生ずる。

そこで、大量のデータを処理する手段に長ける統計的手法が着目されるに至った。

この手法は、NGSM を提唱された石井公成氏の研究が先駆的である。氏は、あるテキストの異なる版本（異本）の特徴を調査し、その結果から異本を系統的に整理・分析する手段として、複数の N-gram 統計結果を一度に比較可能な NGSM を提唱した（石井2001）。

また、NGSM で得られたデータを、統計解析の手法の一つである多変量解析を用いて分析する事を試みたのが師茂樹氏である（師氏は、morogram の作者でもある）。

師氏は、NGSM で得られる大量のデータを効率よく分析する手法としてのみならず、データから何らかの情報を抽出する際に生ずる、抽出者の先入観や恣意性をできるだけ排除した分析を行うための手段として、統計的手法での分析を提唱した（師2002, 2003, 2004）。

師氏は、具体例として『般若心経』の複数の漢訳テキストに対し、NGSM のデータをクラスター分析[(7)]の手法で複数のグループに分ける事を行った（師2002）。そして、その結果を既存の分類モデルと比較する事で、手法の一般的有効性を証明した。また、師氏は別に「ばねモデル」を用いたテキスト間の距離を測る方法を提唱している（師2005）。

また、秋山陽一郎氏は、『老子』傅奕本編集時に参考としたテキストの一つである所謂「項羽妾本」の情報がどこに含まれるかについて共起頻度の傾向から解明を試みた（秋山2002）。

筆者もこの手法を用いた研究を、四種類の異なる手法を用いて行っている。

一つめは、あらかじめ抽出した共起用例を用いた分析である。

これは、「『春秋』経文を春秋期の歴史記録としてどう評価するか」という問題を解決するために利用したものである（山田2004a）。

まず、歴史的変化を捉えるために『春秋』を20年単位で分割した。次に、各単位毎に N-gram 統計をとり、NGSM に整理した。そして、NGSM の結果から諸侯や戦争・外交等の特定の用語についての共起を抽出し、時間軸で整理・分析を試みた。

その結果、『春秋』の記述が、「晋覇体制」という時代の影響を色濃く反映する一方、決して魯を中心とした視座を離れない事、すなわち、『春秋』は、魯の年代記という同時代的資料としての側面を保っている事を明らかにしたのである。

二つめは、あるテキストに見える共起頻度の傾向を分析する方法である。

筆者は、この方法を『孟子』の成書時期・地域を解明するために用いた（山田2004b）。

まず、『孟子』七篇を対象に 2gram で N-gram 統計をとり、各篇間の文字数の異なりを均一化（正規化）するために、各頻度を1000分率の数値に変換した。次に、『孟子』全体での頻度上位200位までの共起を対象に、七篇間での頻度のばらつきを、統計手法の一つである標準偏差で数値化した。

そして、各標準偏差のばらつきも標準偏差で数値化し、その結果を同様な手段で数値化した他の諸子書と比較した。諸子書には、同一の作者やテキストの異本と、成書時期の重層性が明らかなものとを用いている。

表 4 ：『春秋』に見える諸侯の時代別一覧

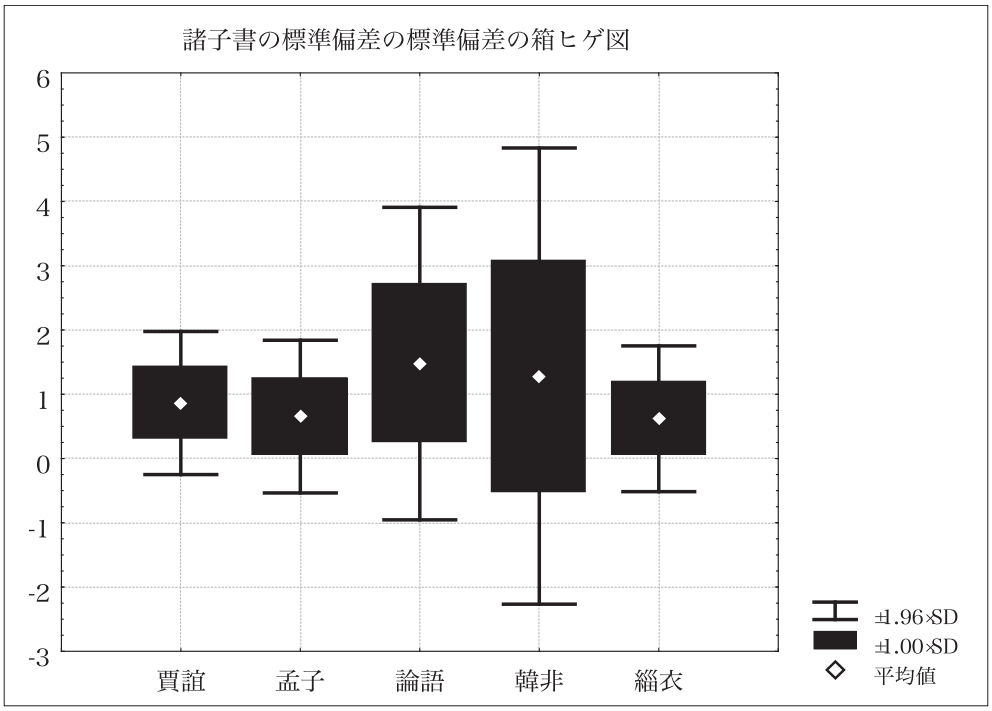
用語	隠01 桓09	桓10 荘11	荘12 荘31	荘32 僖17	僖18 文04	文05 宣06	宣07 成08	成09 襄10	襄11- 襄30	襄31 昭19	昭20 定07	定08 哀12	哀13 哀16
燕伯	0	0	0	0	0	0	0	0	0	2	0	0	0
吳子	0	0	0	0	0	0	0	0	4	1	1	1	1
秦伯	0	0	0	1	0	3	0	1	0	2	0	2	0
楚子	0	0	0	0	1	6	11	5	7	8	1	3	0
蔡侯	2	4	0	1	4	1	1	0	2	5	5	1	0
陳侯	4	6	4	7	5	3	1	5	4	6	1	2	0
許男	0	0	1	9	3	2	1	0	2	3	2	1	1
鄭伯	13	9	5	9	6	3	7	13	9	3	3	2	1
晉侯	0	0	0	4	13	2	14	30	13	4	4	0	2
曹伯	1	1	1	10	4	5	8	15	11	3	2	2	0
宋公	6	11	5	13	11	3	9	14	11	4	2	7	0
衛侯	3	9	4	8	8	3	10	20	14	2	3	9	1
齊侯	14	12	12	18	5	9	10	7	12	3	7	12	0
杞伯	2	1	2	0	2	1	3	7	10	1	4	1	0
薛伯	1	0	1	0	0	0	0	5	8	1	2	2	0
莒子	1	1	0	0	3	0	1	9	10	2	3	0	1
邢侯	0	0	0	1	0	0	0	0	0	0	0	0	0
邾子	0	0	2	1	0	2	4	15	21	7	5	7	0
鄫子	0	0	0	2	2	0	1	0	0	0	0	0	0

これは、複数のテキスト（篇・異本）を比較した場合、作者が同一（同一テキストの異本）であれば、よく使われる表現（共起頻度が多い）の偏り具合が同程度（標準偏差の値が同程度の範囲で収まる）である事が想定される（異なる場合はその逆）からである。

実際に比較した結果（図参照）、『孟子』のよく使われる表現の偏り具合が、自著である「賈誼」（過秦論等）や異本関係にある「緇衣」（『礼記』・郭店楚墓竹簡・上海博楚簡）と同程度である事が確認された。従って、『孟子』は、孟子自身の言説を基本に編纂されたものである事が明らかになったのである。

本研究の成果により、『孟子』に見られる表現が前 4 世紀後半期のものあると認められる事となった。従って、他の諸子書の成書時期を考える上で、『孟子』以前・以後という具体的な定点が提示可能となったのである。

三つ目は、クラスター分析を用いた手法である。



図：各書の共起頻度の出現率のばらつき

この方法は、『孫子』の成書時期を探るために使用した（山田2004e）。

まず、『孫子』と他の諸子百家の文献を対象に 2gram で N-gram 統計をとり、NGSM でひとまとめにする。次に、各文献の文字総数が異なる事から、数値を正規化する。そして、正規化したデータを統計用ソフトでクラスター分析した。

その結果、『孫子』は前 3 世紀に成書時期が見積もられる諸文献と同じ群（クラスター）に属するという結論を得た。銀雀山漢簡の出土以来、孫武による『孫子』編纂説が復活してきたが、本手法での分析は、むしろ旧来の学説に近い結果を示す事となった。

四つ目は、語彙の総量を数値化するユールの K- 特性値を用いた分析である（山田2004c）。

具体的な分析の過程は省略するが、語彙の総量や独自あるいは共用される語彙を数値化した値を分析した結果、儒家・道家は多様な言葉を保持し、墨家は貧困、法家（韓非子）は洗練、雑家（『呂氏春秋』）は諸学を折衷した言葉を用いている事が明らかとなった。

また、時系列で見れば、言葉の多様性は前 3 世紀初唐を一つのピークとなり、前 3 世紀後半には言葉を整理して絞り込む方向性を見せる事が確認された。この推移は、天下統一へと移る時代を反映して、言葉も前 3 世紀後半以降新たな言葉（概念）の乱立が収まり、一定の整理の方向へと向かっていたことを反映していると評価することができるだろう。

3．2．3　複数の手段で分析

この方法は N-gram 統計の結果を、複数の手段を用いて分析を行うものである。

筆者は、先秦の諸子百家や経書等（以下、先秦文献と称す）の成書時期や地域を解明するという課題解決のために、この方法を用いた（山田2005, 2006）。

具体的な方法は以下の通り。

1．先秦文献は、長期間にわたる重層的な成書過程が想定される場合が多い。そのため、分析対象の文献（対象文献）が、重層的成書過程を持つか否かの分析から始める。
2．対象文献を篇・章等の複数の群に分けて N-gram 統計を取り（2gram を使う場合が多い）、その結果を NGSM で一つにまとめる。
3．次に、上述のクラスター分析や標準偏差等の方法を用いて、重層的編纂過程を持つか否か、もし持つ場合はどの程度の群に分かれるかの仮分類を行う。
4．次に、対象文献以外の先秦文献（比較文献）を対象に、同じく N-gram 統計をとる。
5．対象文献と比較文献の両者を NGSM で一つにまとめる。
6．NGSM の結果から、対象文献と比較文献との間に共通する共起用例を見出す。
7．引用関係の有無や、ある時期以降見られる表現の出現状況等、複数の要素を比較する過程を積み重ねて、対象文献の成書時期を明らかにする。

文献間の引用関係や共用する共起用例を分析する事で、複数文献間の先後関係を確定する事ができる。そして、『孟子』や『呂氏春秋』のような、確実な成書時期に関する情報を持った文献と比較する事で、対象文献の成書時期を位置付ける事も可能となる。(8)

表 5 は、筆者のこれまでの研究と他者の研究成果とを加味した、筆者が現在認識する主な先秦文献の成書時期の一覧である。

4．おわりに

以上、N-gram を用いた人文学研究について、その仕組みと実例とを交えて説明した。

3．1．3 でも述べたように、N-gram は仕組みが単純な分（実は、人間が手作業でやる事も可能である）、多くの問題を持っている。

また、分析方法についても、問題がある。筆者は、既存の文献学的な研究成果と比較した場合、統計分析の結果を単純に受け入れる事は危険であると認識している。その為、3．2．3 で述べたように、N-gram 統計や統計的手段で得られた結果を、他の手段で確認しながら作業を進めている。

しかし、N-gram 方式は、単純故に気軽に試せるのも事実である。

表 5 ：先秦文献の成書時期

時　代	諸　　子　　書	
前五世紀晩期以前	『詩經』・『尚書』五誥	
前五世紀晩期～ 前四世紀初頭	『論語』	
前四世紀前期 前四世紀中期	坊記・緇衣 檀弓・表記・「中庸本書」 玉藻・曲礼 少儀・内則（『儀禮』士相見禮に引用有） 『左傳』 『中庸』A（『左傳』以前の記述もあり） 『中庸』B（『左傳』以後『孟子』以前）	
前四世紀後期	『孟子』	『包山楚簡』『郭店楚簡』・『上海博楚簡』（下限：前三世紀初頭。孟子とほぼ同時代）
前三世紀前期	『公羊伝』	『莊子』内篇（逍遙遊・養生主が『呂氏春秋』に引用） 『管子』経言（牧民・形勢は『荀子』に先行） 『莊子』外雑（外篇：天地・達生・山木・田子方、雑篇：庚桑楚・外物・徐无鬼・則陽・譲王は『呂氏春秋』に引用有）
	『竹書紀年』・『穆天子傳』 等のいわゆる『汲冢書』	
前三世紀中期	『穀梁傳』	
	『中庸』C・『呂氏春秋』（前 239 年序）	
前三世紀後期 ～それ以降	『國語』『墨子』・『荀子』・『韓非子』・『莊子』（上記の諸篇を除く）。『商君書』（前 3 世紀以降）	

例えば、先秦文献20種ばかりを対象に、1-5gram の共起頻度を集計し、それを NGSM 化する作業を行おうとすると、手作業ではどのくらいの時間を要するか簡単には見積もれないだろう。しかし、付録（後述）で紹介した各種プログラムとその手順をまとめ書きした bat ファイルを用意して実行すれば、上記の作業は一晩寝ている間に終わるのだ。

そのため、NGSM を用いて、対象文献がどの文献と言葉が重なるかをとりあえず調べてみるという作業が、かなり気軽に実行できるのである。

後は、その中から今までに気がつかれなかった用例の傾向を見いだし、それを分析するだけで、新たな研究が見いだされるかもしれない。

また、３．１でも述べたように、漢字文献を対象に N-gram 統計をとる方法は、漢字が持つ「形」の要素にのみ特化したものでしかない。そのため、N-gram 方式では決して収集できない情報もまた多いのである。N-gram 方式は決して万能ではない。しかし、「形」から得られる情報を集積するのには有効な手段である。

筆者も、先秦文献を対象にした単純な N-gram 方式によるデータ収集のみではなく、あらかじめ文献の構造を踏まえた情報を付加しておき、それを対象に N-gram 統計をとる等の手段で、より絞り込んだ情報からその特徴を見いだすという研究を試みている(9)。

付　記：本稿脱稿後、井上2006を著者井上氏より頂戴した。氏の論攷は、筆者の研究を対象として、その方法論を評価しつつも、情報分析の過程におけるデータ収集方法の妥当性や、情報の取捨選択についての基準に対する疑義を提示したものである。

氏の指摘は納得する部分もあり、また筆者からの反駁を要する部分もある。しかし、筆者の研究に対し、正面から評価と批判をしていただいた事については大変感謝する次第である。また、本来なら本稿で反駁等を行うべきではあるが、これについては別途論ずる予定とした。それは、井上氏の批判が歴史言語学的な情報分析に対してその主点が置かれていると筆者が判断した事による（本稿は、N-gram 方式とその応用についての紹介が主題であり、論の主題が曖昧になってしまう事を避けたく判断した。また、実際的には、井上氏への反駁を行うには、原稿の余量と時間が残っていない点も理由である）。

5．付録：Windows 環境での N-gram 統計の取り方

5．1　morogram について

ここでは、N-gram 統計のとり方について、筆者が用いている師茂樹氏作成のプログラム「morogram」を事例に説明する（実際には、極悪氏[(10)]が「morogram」を Windows 用アプリケーションに改変した「morogram win32/free-standing バージョン」を使う）。（以下、動作環境は何れも Windows XP SP2）

元々の morogram の特徴は以下の通り。

- Nagao and Mori1994 のアルゴリズムで高速に N-gram 分解と共起頻度の収集が実行可能。
- 0 ～16面の Unicode に対応（入出力は UTF-8 のみ）
- 実体参照形式 &Mnnnnnn;(1 ≦ nnnnn ≦ 131,072) を一文字として扱うことが可能。
- 4,294,967,296文字まで対応
- 1 ～4,294,967,296グラムに対応
- 頻度 1 ～4,294,967,296に対応

本来の「morogram」は、UTF-8 で書かれたテキストファイルのみが動作対象となる。従って、N-gram 統計をとりたいデジタルテキストは、あらかじめ UTF-8 形式で保存しておく必要がある（極悪氏版を使用する場合はこの限りではない。後述）。

5．2　morogram の使い方

5．2．1　事前準備

1．あらかじめ、N-gram 統計をとるデジタルテキストを準備する。ここでは、論語 .txt という名のデジタルテキストを用意した。

2．以下の Web ページを開き、ページの指示に従って、極悪氏版 morogram の最新バージョ

ンをダウンロードする。

http://hpcgi1.nifty.com/dune/gwiki.pl?morogram

3．ダウンロードしたファイルは zip 形式で圧縮されているので、それを解凍する（左ダブルクリックすればよい）。

4．解凍後、複数のファイルが生成されるが、それをまとめて一つのフォルダに移動しておく（後の作業が楽になるため）。場所は任意の所でよい。ここでは「morogram」フォルダを「C ドライブの直下に」作成した（C:¥morogram）。

5．「morogram」フォルダに、最初に作成したテキストファイルを移動しておく。

以上で、事前準備は終了である。これ以後の操作は、「コマンドプロンプト」と呼ばれるアプリケーションを起動し、その上で実行する。

5．2．2　実際の使い方

1．Windows の［スタートメニュー］→［すべてのプログラム］→［アクセサリ］→［コマンドプロンプト］を選択し、コマンドプロンプトを起動する。

2．コマンドプロンプトが起動したら、「morogram」フォルダに移動する。

3．「cd¥」と入力して Enter キーを押す。この操作により、C ドライブの最上位（ルート）に移動するはずである。

4．次に「cd morogram」と入力して Enter キーを押す。すると、先程作成したフォルダに移動するはずである。

5．移動したら、コマンドプロンプトに以下のように入力して、Enter キーを押す。

```
morogram-0.7.1yCJKT.exe --f=1 --g=2,2 論語.txt > 2gram 論語.txt
```

上の文字列は、morogram-0.7.1yCJKT.exe（極悪氏改造板 morogram 本体）に対し、「春秋.txt を読み込み、2gram で共起を頻度 1 以上で集計し、2gram 論語.txt に出力する」。と言う命令を出した事を意味する。

その命令を、Enter キーを押す事で実行するのである。

6．出力結果は、「頻度［水平タブ］文字列［水平タブ］gram 数」の形式を持つテキストファイルとなる（以下の例を参照）。

```
2 一人　　2
3 一以　　2
2 一則　　2
1 一匡　　2
1 一得　　2
3 一日　　2
```

「morogram」のオプション（付加機能）は以下の通りである。何れもいわゆる半角の英数字で入力し、各オプションの間は、必ず半角の空白で区切る事。

オプション	説　　明
--help	morogram のヘルプを表示
--f=n	最小頻度の指定。数値は半角で指定する。 無指定の場合「頻度 2 以上」が自動的に指定される。
--g=min,max	最小・最大グラム数の指定。数値は半角で指定。 例：1 〜 3gram →--g=1,3・2gram のみ→--g2,2。 無指定の場合「 1 〜256」が自動的に指定される。
--p	句読点を消去してから N-gram 単位の分割を実行する。
--BOM	Byte Order Mark を出力

また、極悪氏版の独自オプションは以下の通り。

オプション	説　　明
--f=n,m	頻度の範囲を指定する。 例：頻度 2 〜 5 まで→--=2,5
--c	アルファベットの大文字小文字を区別して共起を作成。
--w	文字ではなく、単語単位で共起を作成。
--I=sjis	読み込み元ファイルの文字コードを指定。
--O=sjis	出力先ファイルの文字コードを指定。
--V	morogram のバージョン＋指定可能な文字コード一覧を表示。

例：百人一首 . txt を対象に、いくつかのオプション付きで実行する場合の書式例

- 日本で一般的に使用される Shift_JIS エンコードで保存されたファイルとして読み込み、ngram 百人一首.txt というファイルに結果を書き出し。
- 分割単位は、2gram のみ。
- 頻度集計は 2-5gram

```
morogram-0.7.1yCJKT.exe --I=sjis --O=sjis --g=2,2 --f = 2,5 百人一首.txt > ngram 百人一首.txt
```

5．2．3　注意事項

作業中は、入力ファイルの 3 〜12倍のディスクスペースを利用するため、フロッピーディスクやメモリーカード等の上での実行は余り現実的ではない。できるだけハードディスク上で作業を実行する事。

morogram を実行するテキストファイルは、morogram の実行ファイルと同じフォルダに置い

ておいてもよいが、morogram が存在するフォルダに path を通しておけば、テキストファイルがどこにあっても実行可能となる。

サイズの大きいテキストファイルを対象としたり、大きい gram 数を指定して実行したりすると、実行終了までに非常に時間を要する能性が高い。従って、より短時間で作業を実行したい場合には、より高速な CPU・大容量のメモリーやハードディスクを搭載したコンピュータ環境を用意しておく事が望ましい。

5．3　morogram データの加工

ここでは、morogram のデータを目的に応じて様々に加工して、より情報として利用しやすくするための方法について説明する。

5．3．1　sorlf で並べ替える

上述のように、morogram の出力は「頻度［水平タブ］文字列［水平タブ］gram 数」の形式である。また、並び順については、Unicode 標準の文字表番号に指定された順序で配列されるため、そのままでは、「頻度上位（下位）順にデータを見たい」場合等に不便である。

そのためには、morogram の出力結果を指定した方法で並べ替える手段が必要となる。

並べ替え作業を実行するためのソフトウェアはいくつもあるが、ここでは、morogram の基本文字コードである UTF-8 形式に対応した sortl を使うことにする。

sortl は、益山健氏の開発にかかる並べ替え（ソート）専用の Windows 用プログラムである[(11)]。本ソフトは、morogram と同じく、コマンドプロンプト上で動作する（後述するように、bat ファイルを用いて一括作業を行う際に便利なため）。

sortl の特徴は、上述のように UTF-8 を含む多くの言語に対応している点と、並べ替えの指定方法が豊富な事が挙げられる[(12)]。使い方は以下の通り。

1．sortl は zip 形式でアーカイブ & 圧縮されているため、まずはそれを解凍する。
2．次に、解凍後生成されたファイルの中から、sortl.exe のみを morogram があるフォルダにコピーする（morogram と一緒に使う事を前提にする場合）。
3．次に、コマンドプロンプトを実行し、sortl.exe があるフォルダに移動する。

以下、並べ替えたい morogram 出力結果が、morogram と同じフォルダ（＝ sortl も同じ）にあると仮定して話を進める。

4．sortl で、morogram の構造を踏まえて出力結果を並び替える場合、以下の書式に従って入力すればよい。

●gram 数を基準：大きい gram 数から順に並べ替え

```
sortl -W U -n -r -t ¥t +2 -o ［出力ファイル］ ［入力ファイル］
```

●頻度数を基準：頻度の多い順に並べ替え

```
sortl -W U -n -r -t ¥t +0 -o ［出力ファイル］ ［入力ファイル］
```

その他のオプションは、以下の通り。

表 4 ：春秋』に見える諸侯の時代別一覧

オプション	説　　明
-W U	文字コードに UTF-8 を利用するという宣言。
-n	数字を昇順（数値の小さい順）で並び替える。
-r	数値を降順（数値の大きい順）で並べ替える。
-o	出力ファイルを指定。 morogram とは異なり、［出力ファイル］［入力ファイル］の順に記述する事に注意。
-t ¥t	並べ替える各項目の区切りが、水平タブである事を示す。
＋［数値］	並び替える基準部分を指定。 morogram では、「頻度（¥0)」「文字列（¥1)」「gram 数（¥2)」となる。[13]

5．最後に、「exit」と入力して Enter キーを押すと、コマンドプロンプトが終了する。

5．3．2　ngmerge. pl で NGSM データを生成

複数の N-gram 統計の結果を NGSM 方式にまとめるためのソフトとしては、ngmerge. pl（以下、ngmerge と略）がある。ngmerge は、NGSM ファイルを生成するための perl script 用スクリプトファイルであり、青山大学の近藤泰弘氏の開発にかかるものである。[14]

ngmerge の特徴としては、以下のものが挙げられる。

●複数の N-gram 統計の結果を融合して出力する。
●morogram のような UTF-8 のデータファイルでも結合可能。
●融合するファイル数の限界がない。

ngmerge. pl を実行するためには、あらかじめ perl script 実行環境を用意しておく事が必要となる。Windows の場合は、Windows 用 perl script 実行環境である Active Perl をインストールすればよいだろう。[15]

また、結合する morogram の出力結果は、あらかじめ「文字コード昇順に並べ替えておく」必

要がある。極悪氏版の場合、出力結果の標準は文字コード順なので問題ないが、もし、他の基準で並び直してあった場合は、上述の sortl を利用して並び替えておく事（各 morogram 出力ファイルは、上記 morogram フォルダにある事を前提に話を進める）。

使い方は以下の通り。

1．近藤氏の Web サイトから、ngmerge.pl をダウンロードする。公開されているファイルは perl script 形式なので、特に解凍作業等は必要ない[16]。
2．ngmerge. pl を morogram があるフォルダにコピーする。
3．コピー後、コマンドプロンプトを起動し、morogram フォルダに移動する。
4．ngmerge. pl の実行書式は以下の通り（出力ファイルは、元の結合するファイルと同じエンコードになる）。

```
perlngmerge.pl［一つ目の入力ファイル］［二つ目の入力ファイル］… > ［出力ファイル］
```

例：『論語』の学而第一（01.txt）・為政第二（02.txt）・八佾第三（03.txt）・里仁第四（04.txt）を対象に、2gram で共起頻度を収集し、それを ngmerge.pl でひとまとめにして出力する（ngmerge.txt）。

```
perl ngmerge.pl n01.txt n02.txt n03.txt n04.txt > ngmerge.txt
```

結果は、以下のように表示される。

```
子夏（n01:1 n02:1 n03:1 n04:0）
子張（n01:0 n02:2 n03:0 n04:0）
子曰（n01:15 n02:25 n03:20 n04:27）
子游（n01:0 n02:1 n03:0 n04:1）
子禽（n01:1 n02:0 n03:0 n04:0）
子貢（n01:4 n02:1 n03:1 n04:0）
```

ちなみに筆者は、このファイルを元に、Excel や統計用ソフト等で使用しやすいように、以下の形式に perl script を利用して置換したものを使用している。

共起	学而	為政	八佾	里仁
子夏	1	1	1	0
子張	0	2	0	0
子曰	5	25	20	27
子游	0	1	0	1
子禽	1	0	0	0
子貢	4	1	1	0

5．3．3　bat ファイルで一括処理

以上、morogram・sortl・ngmerge.pl の利用方法について説明した。

これらは何れもコマンドプロンプト上で実行する関係上、複数のファイルを処理する場合、一々同じ事をする必要がある。それが大変面倒な事は言うまでもない。そこで、一連の作業を楽にするために、「bat ファイル」を利用して半自動化する手順について説明する。

bat（バッチ）ファイルとは、MS-DOS の時代から使われている、あらかじめコンピュータに実行される操作や命令を記述したテキストファイルの事である。

bat ファイルを作っておけば、bat ファイルのアイコンを左ダブルクリックする事で、そこに記されている操作が一気に実行されるため、複数の同様の処理を行うのに重宝する。同種の機能を持ったソフトもあるが、今回は使用するソフトが何れも MS-DOS の流れをくむコマンドプロンプト下で実行されるものであるため、bat ファイルを用いている。

書き方は、以下の通り。

- ここでは、先程の『論語』の 4 ファイルを ngmerge でひとまとめにする手順を bat ファイルで一括処理する手順を記述した。
- 上から 4 行は、4 ファイルを対象に 2gram・頻度 1 以上で共起頻度を集計する命令である。
- 5 行目は、生成された 4 ファイル（n**.tyt）を ngmerge.pl でひとまとめにし、ngmerge.txt に出力するという命令を記述した。

```
morogram-0.7.1yCJKT. exe --f=1 --g=2,2 01.txt > n01.txt
morogram-0.7.1yCJKT. exe --f=1 --g=2,2 02.txt > n02.txt
morogram-0.7.1yCJKT. exe --f=1 --g=2,2 03.txt > n03.txt
morogram-0.7.1yCJKT. exe --f=1 --g=2,2 04.txt > n04.txt
perl ngmerge. pl n01.txt n02.txt n03.txt n04.txt > ngmerge.txt
```

1．命令が書けたら、bat ファイルを保存する。文字コードは Shift-jis エンコードを指定して保存する必要がある。また、拡張子は必ず「bat」でなければならない。

2．bat ファイルが生成されたら、後は、当該の bat ファイルを左クリックして実行すればよい。自動的にコマンドプロンプトが開いて命令が実行され、最後の命令が実行されたら自動的にコマンドプロンプトが閉じるはずである。

以上、morogram を利用して N-gram 統計をとり、sortl や ngmerge.pl を利用して、目的に応じた処理を行う手順を述べた。

本稿では紙幅の都合もあるので、ごく面的な説明に限ったが、より詳細を知りたい場合には、筆者の Web サイト「睡人亭」の下位ページ「N-gram モデルを利用したテキスト分析」を参照されたし（http://www.shuiren.org/chuden/teach/n-gram/index-j.html）。

註

（1）シャノンは、現代数学やコンピュータの歴史を語る上で不可欠な人物である。また、彼の情報理論については、シャノン1964を参照されたし。

（2）3gram では、「子曰學」「曰學而」等の文字の組み合わせが「共起」関係となる。

（3）これは、石井公成氏の命名にかかるものである。本来は、版本の系統比較等に N-gram 方式を用いるために考え出されたもので、筆者はそれを先秦文献相互比較のために用いている。

（4）2gram の有効性が高いのは、1gram の単漢字とは異なり、熟語が含まれるためである。それ以上に gram 数を増やすと（特に 4gram 以降）、用例が一致する表現が少なくなる（＝共起及びその頻度）。共起頻度が少ないと、統計的手法で用いる場合に問題が出てくるため、2gram（場合によっては 3gram も有効）が適当と判断した次第である。

（5）全文検索では、「全ての文字の組み合わせ」を対象とする N-gram 方式は、方法的に検索漏れが生じないというメリットがある。また OCR では、判別が難しい画像を文字として解析する際、とりあえず共起頻度の高い用例であると判断して文字にする事で、それなりに高い変換精度を達成する事が可能（例：『論語』で「子」の直後の文字が不鮮明であった場合、最も共起頻度の多い「曰」だと判断しておく事で、48％は正しい解析結果となる）。

（6）取り扱う文献全体で頻度が 1 の用例を全て捨ててしまうのも、同様の手法といえる。特に複数文献を比較する場合、頻度 1 の部分は他と比較しようがないからである。但し、この場合、頻度 1 が頻出する文献の情報を全て切り捨ててしまうため、「他の文献と共有する用例が少ない」という情報は保持されるが、「他の文献と異なる具体的な共起」についての情報を切り捨ててしまう事になる。

（7）多変量解析の一手法。対象となるデータ群を複数のグループに分類するために用いる。

（8）秋山2004a, b では、『戦国策』に含まれる説話群弁別と分析の手段として N-gram 方式を利用し、収集した共起の分析を行っている。

（9）これらの問題については、N-gram 方式を人文系の問題解決に用いようとした当初から指摘されていたが、これについても問題克服の試みが提示されている（谷本2001・師2001等）。

(10) もちろん「極悪氏」とは、インターネット上での「ハンドルネーム」である。

(11) sortl は、益山氏の Web サイトで公開されている。
http://www.asahi-net.or.jp/~ez3k-msym/archive/archive.htm

(12) 例えば、morogram の出力結果は、一行ごとに「頻度［水平タブ］文字列［水平タブ］gram 数」の形式で記述される。これは、区切り文字である［水平タブ］を媒介として、三つの部分に分かれる構造と言える（左から順に「¥0（頻度）」「¥1（文字列）」「¥2（gram 数）」となる）。sortl では、それぞれの部分を基準として並び替える事が可能である。

(13) 0 から番号が開始されるのは、sortl の仕様。

(14) 近藤氏の Web サイトの以下のページで公開されている。
http://klab.ri.aoyama.ac.jp/tool/

(15) 以下の URL に移動して、「Download ActivePerl」を左ダブルクリックしてページを移動し、画面に従って ActivePerl のインストールファイルを入手する。入手後は、ダウンロードしたファイルを左ダブルクリックしてインストラーを起動し、画面の指示に従ってインストール作業を進めればよい。
http://aspn.activestate.com/ASPN/Downloads/ActivePerl/

(16) ダウンロードの際、ブラウザによっては勝手に拡張子を「txt」等に書き換えてしまう場合がある。その場合は、拡張子を「pl」に変更しておく事。

参考文献

C. E. シャノン・W. ヴィーヴァー著／長谷川淳・井上光洋訳　1964『A Mathematical Theory of Communication（邦訳：コミュニケーションの数学的理論）』明治図書出版

秋山陽一郎　2002「『老子』傳奕本来源考―「項羽妾本」介在の検証」『漢字文献情報処理研究』 4
――――　2004a「劉向本戰國策が内包する先行説話群について」『立命館史学』25
――――　2004b「姚本戰國策考―劉向本旧態保存の是非と劉向以前本復元への展望」『中国古代史論叢』
石井公成　2001「N-gram 利用の可能性―仏教文献における異本比較と訳者・作者判定―」『漢字文献情報処理研究』 2　http://www.jaet.gr.jp/jj/2.html
――――　2002「仏教学における N-Gram の活用」東京大学東洋文化研究所附属東洋学研究情報センター編『明日の東洋学』 http://ricas.ioc.u-tokyo.ac.jp/pdf/nl008.pdf
井上　了　2006「中国古典への〈N-gram 分析〉応用に対する若干の疑問」『中国研究集刊』42
北　研二　1999　第 3 章「N グラムモデル」『言語と計算 4　確率的言語モデル』東京大学出版会
近藤みゆき　1999「平安時代和歌資料における特殊語彙抽出についての計量的研究と利用ツールの公開―古今和歌集の歌語と表現のジェンダー性について―」『科学研究費特定領域研究　人文科学とコンピュータ　研究成果報告書―コンピュータ支援による人文科学研究の推進―』
――――　2000「n グラム統計処理を用いた文字列分析による日本古典文学の研究―古今和歌集のことばの型と性差―」葉大学『人文研究』29
近藤泰弘・近藤みゆき　2001「N-gram の手法による言語テキストの分析方法―現代語対話表現の自動抽出に及ぶ―」『漢字文献情報処理研究』 2
谷本玲大　2001「曖昧検索性を持たせた N-gram サーチの手法―『新撰萬葉集』と菅原道真の詩の比較を例に―」『漢字文献情報処理研究』 2
Makoto NAGAO and Shinsuke MORI. 1994 “A New Method of N-gram Statistics for Large Number of n and Automatic Extraction of Words and Phrases from Large Text Data of Japanese”, In Proceedings of the 15th International Conference on Computational Linguistics, pp.611–615.
http://www-lab25.kuee.kyoto-u.ac.jp/member/mori/postscript/Coling94.ps
長尾　真　1996『自然言語処理』岩波講座『ソフトウェア科学』15　岩波書店
――――　1998『言語情報処理』岩波講座『言語の科学』 9　岩波書店
松本裕治　他　2004『単語と辞書』岩波講座『言語の科学』 3　岩波書店
師　茂樹　2001「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』 2
――――　2002「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』の異訳の比較を例に」京都大学大型計算機センター第69回研究セミナー『東洋学へのコンピュータ利用』予稿集
――――　2003「N グラムによる比較結果からの用例自動抽出―禅宗系の偽経を題材に」『東洋学へのコンピュータ利用第14回研究セミナー予稿集』2003
――――　2004「N グラムと文字データベースによる漢字仏教文献の分析」『情報処理学会研究報告』Vol. 2004, No.7
――――　2005「（研究ノート）NGSM 結果のばねモデルによる視覚化」『漢字文献情報処理研究』 5
http://jaet.gr.jp/books.html
山田崇仁　2001「『國語』韋昭注引系譜資料について― N-gram 統計解析法による分析―」『立命館史学』22
――――　2004a「歴史記録としての『春秋』― N-gram モデルと統計解析法による分析―」『中國古代史論叢』
――――　2004b「『孟子』の成書時期について― N-gram と統計的手法を利用した分析―」『立命館東洋史学』27
――――　2004c「中国戦国期の語彙量について― N-gram とユールの K 特性値を利用した分析―」『漢字文献情報処理研究』 5
――――　2004d「N-gram による先秦文献の分類」『漢字と情報』 8

———— 2004e「N-gram モデルを利用して先秦文献の成書時期を探る―『孫子』十三篇を事例として―」東京大学　東洋文化研究所附属東洋学研究情報センター「アジア研究情報 Gateway」http://asj.ioc.u-tokyo.ac.jp/html/034.html
———— 2005「『礼記』中庸篇の成書時期について― N-gram モデルを利用した分析―」『中国古代史論叢』続集
———— 2006「『周禮』の成書時期・地域について」『中国古代史論叢』三集