特集2

N-gramが開く世界

確率・統計的手法による新しいテキスト分析

近藤泰弘氏は、1999年、確率・統計的自然言語処理の分野で用いられていた **N-gram**を国語学に応用した発表を行ない、以後、氏と近藤みゆき氏は、この手法を国語学・国文学研究に適用して次々に優れた成果をあげられた。両氏の手法の豊かな可能性にいちはやく着目した漢字文献情報処理研究会では、同会のインターネット上の掲示板で、両氏をまじえてこの問題を熱心に話しあい、その過程で様々な利用法や便利なスクリプトが続々と開発されるに至った。

N-gramは、単なる検索を超え、文献そのものの分析に踏み込むことを可能にする強力で便利な補助手段である。今後のテキスト分析においては、**N-gram**をどのように使いこなしてゆくかが、重要な課題となろう。国語学・国文学・仏教・中国思想などにおける **N-gram**利用に関する諸論文と導入篇から成る本特集は、多様な分野での **N-gram** 活用の現状と可能性を紹介する世界最初の試みである。

＊N-gram については、n-gram、N-gram、N-Gram、Nグラム、nグラム、その他、様々な表記が内外で使われているが、本特集では、英語表記については N-gramを用いた

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—

近藤 泰弘 こんどう やすひろ
近藤 みゆき こんどう みゆき

はじめに

N-gram的な考え方は、従来から、言語処理、特に音声分析に広く用いられてきたが、近年、長尾真・森信介の両氏の開発された、任意の数の文字のN-gramをテキストから抽出するソフトウェアが発表されたことにより、語学・文学における応用に道が開かれるようになった[1]。筆者らも、このソフトウェアを用いて、平安時代の言語位相や引用などについての研究を行ってきた。しかし、この方法はまだ応用が開始されたばかりのものであり、言語テキストに文字単位のN-gram分析を用いることで、どのようなことが可能であるのかについては未だ充分に知られているとは言えない。そこで、本稿では、その可能性の範囲について概略を述べ、かつ、その一部の方法の新しい応用について述べてみたい。

N-gramの言語分析の特徴

長尾氏らによって開発されたN-gramの文字列分析の特徴は、これまで情報工学の中でN-gramといえば、単位の連続の出現確率などが主に問題になっていたのに対し、テキスト中から任意の長さの文字列をすべて抜き出してくることで、テキスト全体を総体的に見ることができるようになったことにある。きわめて人文科学に適した方法論と言える。

具体的に、例えば、次は、『源氏物語』『古今集』『後撰集』『拾遺集』に現れるすべての文字列をN-gram抽出ソフトウェアによって抜き出し、出現頻度とともに五十音順に対照したものの冒頭部分である。

```
あ ( genzi:13795 kokin:777 gosen:935 syui:908 )
あか( genzi:493  kokin:30  gosen:32  syui:50  )
あかざ      ( genzi:5  kokin:1  gosen:0  syui:1 )
あかざり    ( genzi:4  kokin:1  gosen:0  syui:1 )
```

```
あかざりし　( genzi:3 kokin:1 gosen:0 syui:1 )
あかし　( genzi:137 kokin:2 gosen:1 syui:8 )
あかして　( genzi:8 kokin:1 gosen:0 syui:1 )
あかしの　( genzi:45 kokin:1 gosen:0 syui:1 )
あかしのう　( genzi:8 kokin:1 gosen:0 syui:1 )
あかしのうら ( genzi:8 kokin:1 gosen:0 syui:1 )
あかす　( genzi:8 kokin:1 gosen:3 syui:3 )
あかすら　( genzi:1 kokin:1 gosen:0 syui:1 )
あかすらむ　( genzi:1 kokin:1 gosen:0 syui:0 )
あかず　( genzi:92 kokin:11 gosen:6 syui:6 )
あかずと　( genzi:1 kokin:1 gosen:2 syui:0 )
あかずも　( genzi:2 kokin:3 gosen:0 syui:0 )
あかずもあ　( genzi:2 kokin:2 gosen:0 syui:0 )
あかずもある ( genzi:2 kokin:2 gosen:0 syui:0 )
あかずもあるか( genzi:2 kokin:2 gosen:0 syui:0 )
あかずもあるかな ( genzi:2 kokin:2 gosen:0 syui:0 )
```

いわゆる単語ではなく、「あかずもあるかな」という独特の言い回しの文字列が、『源氏』に2例、『古今』に2例あることが、たちどころにわかるわけである。この手法によっては、その作品中のあらゆる文字列が切り出されてくるため、従来の単語索引では知ることのできなかった連語や言い回しのレベルでの傾向を、網羅的に知ることができる。このことは、人文科学研究にとっては、作品分析や言語分析における着眼点そのものを変え得る可能性を示唆する、画期的なことであると言えよう。

ここではすべて平仮名の形にした古典日本語を例にしたが、本特集の他の論考で示されているように、中国語などの漢字文献においてもこの方法は当然有効である。ただし、「文字」という言語単位の持つ言語学的な意味が言語によって異なるために、それぞれの言語ごとに1文字に対して、おおよそ次のような、異なった位置づけを与えておく必要がある。

- 日本語(仮名)　1モーラ(拍)
- 英語等　(ほぼ) 1音素
- 中国語　(ほぼ) 1形態素

したがって、文字のN-gramを採取して調査する場合に、日本語では、いわゆる五十音のモーラの連続であるところの、単語や複合語や言い回しを調査することができる。中国語・漢文では形態素や単語の連続を採取したこととなるのであり、単語や複合語の調査の他、慣用句やよく出現する定型句といったものの調査に適している。このように文字のN-gramといっても言語学的には異なったものを見ていることになるので、他の言語のN-gram分析と比較する場合には留意しなくてはならない。

また言語自身の持つ類型的(typological)な違いも分析に影響する。日本語や中国語は、形態素が連接して長い連語を形成したりするし、また主語・目的語・述語が比較的近い位置に存在しており、N-gramで捉えられる連続的な単位の分布の中で各種の言語学的な問題が容易に捉えられると考えられる。

さらに、連続的な分布という点だけを見るならば、日本語については、形態素解析(単語への分解)を行ってから、その単語の連続としてのN-gramの分析を行う方が処理の効率もよいように考えられるかもしれない。しかし、日本語の場合、そもそも形態素解析自体に問題があり、どのような単位に分解するかによって大きくその結果が異なってくる。例えば「食べさせられてしまっていたかもしれないんだろうね」(長大な付属語列)や、「我が国」(古典語の場合の複合形式)などの形式を統一的に単位に分解す

ることは容易ではない。この形態素解析の困難さは、最初から漢字という形態素単位に分解されている中国語や、文字表記の上で綴りがおおよそ形態素で分離されている英語などとは大きく異なった部分である。日本語の N-gram 分析においては、あえて、モーラや文字の連続として扱ってしまうことで、余分なものは入ってしまうものの、分解しにくい複合語語形も、もれなく分析を行うことができるというメリットもあるのである。

2種類のテキストの N-gram 対照

ところで、文字の N-gramをひとつのテキストから採取しても、多くの場合、そのテキストについてそこからすぐに結論を引き出すことはなかなか困難である。つまりかなり大量の文字列がそのままの形で現れてくるために、そのどこに着目してよいかがわからないのである。そこで、筆者らは、『古今集』内部の男性歌と女性歌[2]、『古今集』と『源氏物語』[3]というように、二種類の異なったテキストを対照し、それぞれの N-gram 文字列を集合演算し、次のようなことを述べた。

1. **差を見ることによってそれぞれのテキストの特徴を対比的に明らかにできる**
2. **重なり部分を見ることによって、それぞれのテキストの相互影響関係を調査することができる**

具体的には取り出した N-gram の文字列を長さを無視して次のように五十音順に並べたものをそれぞれの文献に対して作り、unix ツールの comm によって二つの N-gram 抽出文字列を相互比較することで行った。例えば、次の２つの歌の一部分を比較してみた場合を例にすると、下の文字列抜き出しのモデルのようになって、左右の合致した太字の箇所がこの部分での最長共通文字列となる。

をみなへしおほかるのべにやどりせばあやなくあだのなをやたちなむ（古今・229）

はなにあかでなにかへるらむをみなへしおほかるのべにねなましものを（古今・238）

（古今・229）	（古今・238）
を	を
をみ	をみ
をみな	をみな
をみなへ	をみなへ
をみなへし	をみなへし
をみなへしお	をみなへしお
をみなへしおほ	をみなへしおほ
をみなへしおほか	をみなへしおほか
をみなへしおほかる	をみなへしおほかる
をみなへしおほかるの	をみなへしおほかるの
をみなへしおほかるのべ	をみなへしおほかるのべ
をみなへしおほかるのべに	**をみなへしおほかるのべに**
をみなへしおほかるのべにや	をみなへしおほかるのべにね
をみなへしおほかるのべにやど	をみなへしおほかるのべにな

差や重なりとして抽出された文字列を、出現頻度数とあわせて比較検討することで、古今集論としては、男性歌人特有の表現を網羅的に抽出して歌語のジェンダー性を具体的に論ずることを実現し、また、『源氏物語』においては、『古今集』と『源氏物語』の言語の共通部分をすべて抽出することで、和歌表現と深い関係を持つとされる『源氏物語』の言語と文体に新たな視点を提供する手がかりを得ている。

3種類以上のテキストのN-gram対照

また、仏教学の漢文文献の研究の中で、石井公成・師茂樹両氏は、この相互対照を3つ以上の文献に対して行い、それを出現頻度付き表形式で表現することによって、漢文作品の諸本系統や作品ごとの単語の比較などを効率よく行えることを示した。石井氏はこれを NGSM (N-Gram based System for Multiple document comparison and analysis)と呼称しておられ[4]、本特集の中でもそれによって諸氏の研究が展開されている。近藤も、このNGSMの仕様による表を高速に作成するツール（ngmerge）を作成し公開しており[5]、本論文冒頭の対照表はそれによっている。このようにいくつかの文献の対比そのものについて、文字N-gramによる方法を使うことは、以上にあげたように、言語を問わず、言語の位相研究、引用研究、諸本対照研究などにとってきわめて有益である。今後、この手法やツールが一般化すれば、KWIC索引などと並んで、文学・語学・歴史・宗教等の分野の諸テキストについて必ず行わなくてはならない研究手法となると思われる。

テキスト対照による言語自体の性質の分析

ところで、N-gramを用いたテキストの対比という手法にはもうひとつ別の研究方法としての側面がある。それは、言語それ自体の分析においての有効性という側面である。上記で紹介した研究はいずれも作品研究の範疇に属するものであるが、同時に、N-gramを用いて、テキスト間の差などを見た場合、その差として抽出される部分には、自動的に濃縮された形で当該言語そのものの特徴が現れることがあるのである。これは、前述のように、語構成上、正規文法的傾向を強く持つ日本語の分析に、そもそもN-gram 文字列分析という手法が極めて適していることに由来すると考えられるが、とりわけ、言語的性格の異なる二つ以上のテキストを対比するとき、いわば対比という作業がフィルターとなって、膨大な数の文字列の中から、それぞれの言語テキストの有意の特徴が効率よく抽出されてくるのである。そしてその結果は、言語現象をめぐって、時にこれまで見えてこなかった、側面を見せてくれる場合がある。平安時代語については拙稿で複合語の抽出ということで述べたので[6]、以下ではこの手法を別の対象に応用してみることとする。

対話的表現の自動抽出

今回は対話的表現を抽出するために、次のような手法を用いた。まず書き言葉資料として、『日本経済新聞』CD-ROMの1994年の1月29日から31日までの三日分から1から9グラムまでのN-gramを取り出した。次に対話的な資料として ATR で作成された対話データベース（国際会議・旅行）を用いた[7]。こちらも同様に1から9グラムのN-gramを取り出した。参考までに対話データベースの原文の一部を次に示しておく。（原データはそれぞれ1.5メガバイトほどである）

> 質問者：はい、私このあいだ会議に申し込みまして、そのとき宿泊も一緒に頼んだんですけど、契約ホテルがいっぱいだったので、サンルート名古屋なら取れると言

われたんで、一応頼んだんですけど、これは会場からは遠いんでしょうか。

事務局：サンルート名古屋ですか、それは名古屋の何区にあるホテルなんでしょうか。私どもも、ちょっと契約ホテル以外だとよくわからないものですから。

このそれぞれから1から9までの長さの文字列をすべて取り出すと、日経新聞は、約285万行、対話データベースは、約270万行が取り出される。これを前記のngmergeで相互比較して対照表を作成した。対照表は約549万行のサイズになる。そのまま通覧することはとうてい不可能であるので、対照表のうち、日経新聞で一回も出現しないもので、かつATRデータベースに500回以上出現する文字列を抜き出すと次のようになる。

はい} [あの	(nikkei:0 atr:507)
まして、	(nikkei:0 atr:598)
ます。は	(nikkei:0 atr:1386)
ます。はい	(nikkei:0 atr:1357)
ます。はい、	(nikkei:0 atr:1066)
ますので。	(nikkei:0 atr:604)
よろしい	(nikkei:0 atr:831)
よろしいで	(nikkei:0 atr:551)
よろしくお	(nikkei:0 atr:538)
よろしくお願	(nikkei:0 atr:529)
よろしくお願い	(nikkei:0 atr:529)
りがとうご	(nikkei:0 atr:693)
りがとうござ	(nikkei:0 atr:693)
りがとうござい	(nikkei:0 atr:693)
りがとうございま	(nikkei:0 atr:693)
りますので	(nikkei:0 atr:656)
ろしい	(nikkei:0 atr:831)
ろしいで	(nikkei:0 atr:551)
ろしくお	(nikkei:0 atr:538)
ろしくお願	(nikkei:0 atr:529)
ろしくお願い	(nikkei:0 atr:529)
わかりま	(nikkei:0 atr:670)
わかりまし	(nikkei:0 atr:583)
わかりました	(nikkei:0 atr:582)
わかりました。	(nikkei:0 atr:552)
んですけ	(nikkei:0 atr:1244)
んですけれ	(nikkei:0 atr:843)
んですけれど	(nikkei:0 atr:843)

このように、頻度数付きの形でN-gramが採取されている場合、その頻度数の対比を相当に大きくすることで、きわめて濃縮された形でその対比を示すことが可能になるのである。

さて、ここに出現するものは、新聞に出現せず、会話に高頻度で出現することから、基本的に口語的あるいは対話的表現であると考えられる。その他の頻度の場合も考慮しつつ抽出したものは、テストケースとして9文字（9グラム）のものまでであるが、大きく分けて次のように分類できる。

1. 敬語表現　お願いいたします・させていただきます・なっております・お名前
2. 間投助詞・終助詞　っけ・さ・よ
3. 感動詞類　ああ・はい・ありがとう・失礼します
4. 接続助詞　ですが・けど
5. 接続詞　じゃあ・それでは・ですから
6. 特殊な語彙　さっき・こちら・そちら・このへん
7. 音韻変化による語形　てるんです・いただくっていう・いただくんで

8. 慣用句　て、大丈夫・と申します

以上の中で特に注目されるものは、6 番目にあげた特殊な語彙、特に、指示する言葉としてのいわゆるダイクシスに関わる語彙である。指示詞を初めとするダイクシスに関わる語彙は、話者と聞き手の相互関係の中で話者が時間・空間的な位置関係を指示するものであり、対話語の中に出現するのは自然なことである。しかし、その中にあっても「これ」「あれ」などは、例えば「あれも書きたいこれも書きたいという現場の考えが先行して」（1 月 29 日朝刊）というような間接引用的な文脈などを用意することができて書き言葉にもしばしば用いられる。「それ」も言うまでもなくよく使われる。これに対して「さっき」「そちら」のような近接的時間や空間を表すダイクシスの語はこのように文章語としては用いにくい。その理由はまだ明らかではないが、指示する空間の大きさや抽象度といったものと関係するものと思われる。対話という空間に用いられる語彙や文法を N-gram 分析によって自動的に抽出し明らかにするという方法は、きわめて有望であると思われ、今後の課題として行ってゆく予定である。

おわりに

以上、本稿では、N-gram 分析の応用が、日本語の文学・語学研究において、いかに新しい可能性を秘めたものであるか、いくつかの具体例をあげて述べてみた。これまでの稿者たちの手法は、集合演算によるテキスト比較という点に力点を置いたものであったが、それ以外にも研究者各自の問題意識の構え方によって、多くの展開が期待されるところである。

近年、現代語の形態素解析においてめざましい成果をあげている、文字クラスの N-gram による形態素解析の成果[9]によって、古典語の形態素解析の自動化を行うことなど、今後のこの方面の情報処理技術の応用にはきわめて広い展望が広がっていると思うのである。

注

[1]長尾真・森信介「大規模日本語テキストの n グラム統計の作り方と語句の自動抽出」（『自然言語処理』96-1,1993）

[2]近藤みゆき「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」（『特定領域研究「人文科学とコンピュータ」1998 年度研究成果報告書』及川昭文編、1998）、同「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―」（『千葉大学文学部紀要・人文研究』29・2000）

[3]近藤泰弘《文化資源》としてのデジタルテキスト―国語学と国文学の共通の課題として―」（『国語と国文学』77-11,2000）

[4] Ishii, Kosei. “Classifying the Genealogies of Variant Editions in the Chinese Buddhist Corpus”. 2001 EBTI International Conference, Seoul, Korea, May 2001 口頭発表

[5]http://klab.ri.aoyama.ac.jp/tool/

[6]近藤みゆき「n-gram 統計による語形の抽出と複合語―平安時代語の分析から―」（『日本語学』20-9,2001）

[7]ATR 対話データベース（国際会議・旅行）電話・キーボード（国際電気通信基礎技術研究所）

[8]国立国語研究所『日本語の指示詞』（1981）など参照。

[9]小田裕樹・森信介・北研二「文字クラスモデルによる日本語単語分割」（『自然言語処理』6-7,1999）など。