コンピュータを用いた統語法研究の方法

近藤泰弘（青山学院大学）

はじめに

日本語の文法記述について、数理的研究の方法を用いることについては、大きく分けて 2 つの方法があると思われる。ひとつは、何らかの統計的手法によってコーパスにおける用例の数を調査し、その分布から歴史的な変化や、位相上の異なり等を導き出すものである。また、この方法のヴァリエーションとして、用例の有無を、非文の判定に用いるものもある。母語であり、現代語である場合には、研究者の内省によって、いわゆる非文（非文法的な文）の判定をすることができるが、たとえ母語であっても古典語の場合など（たとえば、平安時代日本語）では、その判定は不可能である。ちなみに、日本語の古典語の研究で、コーパスでの用例の有無による非文判定の手法を意識的に用いた論文（古典語用例に＊マークがついているもの）は、近藤（1979）からではないかと思う。コーパスでの用例の有無を非文に繋げることには賛否があることはわかるが、ひとつの手法として今では広く使われているのも確かである。日本の英語学でコーパス言語学が非常に重要視されているのも、同じ理由による。内省が効かない言語を研究する場合には、コーパスの威力は大きい。

以上の方法は、従来、単語検索やその前後承接を使うことが多かった。近年では、単語の N-gram（N 個の単語連鎖）や文字の N-gram（N 個の文字連鎖）を用いて計測することも行われているが、本質的には同じである。また、コロケーションのレベルでのコーパスの統計調査を行うツール（たとえば、NINJAL-LWP など）もこの単語 N-gram 研究の範疇に入ると思われる。ちなみに、日本語学の古典語分野で、N-gram を用いた分析をおこなったものは、近藤（2001）が最初である。

また、もうひとつ言語研究の「数理的」な側面として、自然言語を形式言語として取り扱うという研究方法がある。この方面の研究は基本的にはチョムスキーに始まる。形式言語の文法を 0 型から 3 型までに分類し、人間の自然言語は基本的に 2 型の文脈自由文法を基本としたものであり、そこに「変形」が加わるというのが生成文法であった。生成文法はその後多くの変遷を経ているが、基本的には、文脈自由文法（句構造文法）の基本的な考え方は変わっていない。つまり、自然言語といえば 2 型（文脈自由文法・句構造文法）ということで研究されてきている。近年、工学部の自然言語処理では、この形式言語を基礎にした解析（句構造文法のパーザー）は勢いが弱くなり、機械学習や深層学習による自然言語処理が全盛となっているが、だからといって、何か特別の別な形式的な特性が明らかになったわけではない。

以上が、言語研究（文法研究）における「数理的」な研究の 2 つの側面である。本日の発表では、この 2 つの側面を少しずつ活用したような研究をご紹介してみたい。つまり、「数を数えて、言語の形式を論じる」研究である。

1　日本語の言語単位とコンピュータ処理

さて、コンピュータで言語単位の数を数える話となるが、日本語の文法記述の基礎的単位としては、文の成分である「文節」を調べるのが能率がいいだろう。文節の構成要素は、自立語と付属語、つまり単語であるが、しかし、よく知られているように単語の理論的な規定はむずかしい。本発表のように、国立国語研究所の「現代日本語書き言葉均衡コーパス」（BCCWJ）を使う場合を例にすると、そこには、「短単位」か「長単位」かという問題がある。短単位と長単位は、国語研の伝統的な単語の 2 種類（β 単位と α 単位）をもとにした考え方であり、例えば、「日本語」は、短単位では「日本＋語」だが、長単位では「日本語」である。通常の文法的な「単語」のイメージでは「日本語」の方が理解しやすいものである。詳しくは、小椋秀樹他（2011）を参照されたい。

また、そのアノテーション（情報付け）は、形態素解析を行ったソフトウェア（mecab）と辞書（UniDic）に依存しているため、品詞体系など、その制約の中でのことになるのは間違いない。従って、「文節」は、自立語（用言か体言）に付属語が接続するという明確な形態的な規則があるものの、BCCWJ コーパスで扱う場合には、短単位で扱う場合と、長単位で扱う場合との 2 種類を考える必要がある。

今回の発表では、特に、1 文節にいくつの単語が含まれるか（以下では、それを「文節長」と呼ぶ）の記述を中心に、短単位と長単位による文節長の計測について述べていきたい。また、それにより、言語単位のコンピュータによる計測がどのように文法記述に役立つかについて述べていく。また、日本語の歴史の中における文節の構造についても触れることにする。

なお、今回の研究にあたっては、BCCWJ のすべての用例が必要であるため、「中納言」を用いることはできない。DVD により配布されている BCCWJ の Ver1.1 および、中納言のデータ管理を行っている国語研の所内 SQL サーバーのデータを用いた。これは別の機会に論じたいが、「中納言」は近年ユーザーの増加で、反応の遅さやエラーの発生が目に付くようになってきた。確実な研究のためには、DVD のデータを元に、各自の研究室等で、自前で SQL サーバーにデータを格納して運用することが望まれる。SQL サーバーが使えない場合でも、SQLite などの Python の SQL 専用ライブラリモジュールで運用する方法もあり、それがもっとも簡便だろう。

2　日本語の文節の構造について

日本語の文節には、大きく分けて体言を自立語部分とするものと、用言を自立語部分とするものの 2 つに分かれる。（宮岡伯人（2002）は、後者を「用言構造体」と呼んでいる。宮岡（2015）では「体言複合体」という用語も用いるが、その場合は、助詞などは含まない。）なお、今回の発表では、副詞や接続詞などを核とする文節については扱わないことにする。

体言と用言、それぞれを核とする文節の基本は似ているが、若干の相違もある。まず基本的な形としては、両者とも自立語部分が単純用言・体言である場合と、複合用言・体言である場合があり、付属語部分には、体言では助詞が接続し、用言では、助動詞・助詞が接続する。

（体言文節）
〇新聞にだけは
〇国立国語研究所からは
（用言文節）
〇読んでしまっただろうか
〇読み始めないかもしれない

また、体言文節の場合は、自立語部分に、いわゆる「臨時一語」が来ることがある。
〇国会議事録公開情報開示が

今回の研究では、この「臨時一語」の分析は、すこし問題の種類が異なるために、触れていない。臨時一語を含む文節の分析については別の機会を持ちたい。

さらに、体言も用言も、いずれも助詞・助動詞部分にいわゆる「複合辞」が入る場合がある。
〇努力によっても
〇走らないかもしれない

「複合辞」の認定をどうするかは難しい問題であるが、今回の研究では、BCCWJ を用いているため、基本的にはその認定によることにする。つまり、短単位だけで計測する方法では、基本的には複合辞は無視されるため、「努力によっても」は「努力に」と「よっても」の 2 文節とされる。それに対して、長単位の場合は、複合辞は（すべて網羅されているわけではないが）多く「長単位」での 1 語となるため、「努力によっても」は 1 文節とされる。「走らないかもしれない」も 1 文節である。

ここでひとつ興味深い疑問が生じる。本研究では、日本語の文節長を計測するわけだが、その最大長はいくつであるかという疑問である。

先の例だと、

（短単位による計測）努力-に ／ よっ-て-も　（2 文節で、それぞれの長さは 2 および 3）
（長単位による計測）努力-によって-も　　（1 文節で、長さは 3）

となるが、最大文節長は、はたして、短単位によるものが大きいのか、長単位によるものが大きいのだろうか。

短単位は短く切るため、文節長が増える可能性があるが、途中で文節が切れることにより、その点では、文節長も短くなるだろう。逆に、長単位では、ひとつの単位が長いため、数は減るかもしれないが、複合辞の働きで、文節が長く続く傾向もあるかもしれない。

3 BCCWJ における文節長の計測

さて、実際に、文節長に注目して調査してみると次のようになる。まず、BCCWJ 全体の概数であるが、現在、中納言で運用されているデータで、短単位数で 1 億 2410 万件、長単位数で 1 億 187 万件であった。その中から、先に述べたような定義で、プログラム処理により文節を抽出したところ、短単位で 5600 万件、長単位で 4094 万件あった。

この抽出された文節を、それぞれ文節長によってソートして、上位に来るものを調べて見ると、極めて興味深いことがわかる。

まず短単位の計測で最長だったのは文節長 10 である（文節の内部の単語（短単位）の切れ目はハイフンで示す)。

○知り-たい-ちゅう-た-だけ-な-ん-じゃ-の-に-LBj9_00254
○すみ-ませ-ん-でし-た-だけ-な-ん-です-か-OC14_06765
○新-幹線-車-内-だっ-た-ん-です-が-ね,OY15_18903
○お-母-さん-だっ-た-ん-だ-よ-など-と,LBr0_00006

「新幹線」の体言文節などは、明らかに短単位の短さに助けられて文節長が長く計測されている例である。次は文節長 9 の例。

○出会え-なかっ-た-だけ-な-ん-です-よ-ね,OC09_05447
○病院-だっ-た-ん-だ-よ-と-か-です,OC12_04632

全体として 10 の例は数えるほどしかないが、9 の例以下は非常に多くなる。

次に長単位による計測結果を見てみよう。こちらの最大文節長は 11 であるが、1 個しか例文はない。

○思い-ます-よ-と-か-じゃない-の-か-なあ-と-か,LBq3_00082

次に文節長 10。やはり 10 例ほどしかない。最初の例は、複合辞によって引き延ばされていることがわかる。

○それ-まで-の-ことはなかっ-た-ことになっ-てしまい-ます-の-や,OB2X_00107
○偽計業務妨害-だ-の-な-んだ-の-という-の-に-は,OC14_06519

文節長 9 の例は、非常に多くなる。

○はっきりし-ませ-ん-です-ね-など-という-の-は,OB1X_00157
○更新中-な-だけ-だ-と-か-じゃない-です-か,OC01_03959
○圧迫さ-れ-てる-の-か-な-くらい-に-しか,OC09_11964
○交際し-て-なかっ-た-の-で-な-んだろう-と,OC11_00582
○考え-てい-ない-の-か-な-という-の-は,OM31_00007
○踏台-で-しかなかっ-た-のではない-か-と-まで-も,PB13_00280

内容もごく自然な語句であり、普通に発話される文と言ってよいだろう。また、当日の発表では、グラフなどの形で示すことにしたいが、圧倒的に用言文節の方が件数が多い。これは、体言文節が付属語部分が助詞だけによるのに対し、用言文節は、助動詞の助けを得て作られるからであるのは間違いない。最後の例は、体言文節であるが、「しかない」「のではない」などの複合辞がフル活用されて、長くなっている。

しかし、それにも係わらず、最長の長さが、体言も用言もほぼ同じであることは興味深い。また、このように、短単位による計測も、長単位による計測も、文節の最大長にはさほどの影響を及ぼさない点にも注目したい。

例えば、次の例（長単位）では「ている」が 1 単位であるために文節として繋がっているが、短単位だと「意識されて」で一旦文節は切れてしまうのである。このような例は、文節長を長くする方向に働くことは間違いないが、と言って、「ている」や「という」のせいで単純に 2 倍に長さが増えるということもないのである。

○意識さ-れ-てい-なかっ-た-から-な-のだ-ね,LBt8_00009

ここから、単に統語的な制約があるだけではなく、日本語にとって、ひとつの文節に盛り込める情報量に制約があると考えざるを得ない。

さて、このように現代日本語の文節長の最大は 11 あるいは 10 であることがわかったが、これは歴史的に見た場合にどのようになるのだろうか。これについては、先般、近藤（2020）でも若干触れたので少し重なる部分もあるが、次の節で見ておきたい。

4 文節長の歴史的変化

これについては、小田（2020）に言及があり、平安時代語では、理論的には 13 語の連接による文節が想定できるとされている。

○動詞-られ-ぬ-べかり-し-なり-けり-と-ばかり-を-だに-こそ-は

（きっと自然と…されるに違いなかったのだったとせめてそれだけでも）

また、この例文に見られるように、長くするために、断定の「なり」を挟み込むことが有効であると述べている。つまり「ける＋なり＋けり」のように、通常なら 1 回しか使えないテンスの「けり」を「なり」を介在させることで、2 回使うことが可能になるからである。現代語でも「歩いたのだった」のように「のだ」を介在させることで「た」を 2 回使うことができる。しかし、実際に、国語研究所の「日本語歴史コーパス」（CHJ）によって、平安時代語の短単位の最大文節長を計ってみると、最大は 7 となる。次がその例である。

○あはれ-な-めり-し-など-を-も,源氏
○いか-なら-む-とて-に-か-と,蜻蛉
○帰ら-まし-や-は-と-のみ-なん,源氏
○語り-つ-べから-む-を-がな-と,枕草
○言は-せ-たり-ける-なり-けり-と,蜻蛉
○置か-せ-たる-に-や-など-ぞ,源氏

古典語では、短単位と長単位の差は少ないため、長単位の例は、ここでは省略するが、いかにも長い文節がありそうな平安時代語も実は最長は 7 であり少ない。CHJ 全体を通して、最長のものは、明治時代に現れる。挙例の最初のものから、9, 10 であり、あとは 8 である。つまり、全体としては平安時代に 7 だったものが、ゆるやかに増加して近代に至っているようだ。

○小-なら-しめ-ざる-可から-ざる-が-如し-と-雖も.太陽
○諸國-の-如く-なら-ざる-可から-ざる-が-如し.太陽
○摩する-が-如く-なる-のみ-なら-ん-や,太陽
○高貴-なら-しめ-ざる-べから-ざれ-ば-なり.太陽
○精良-ナラ-シメ-サル-ヘカラ-サレ-ハ-ナリ,東洋

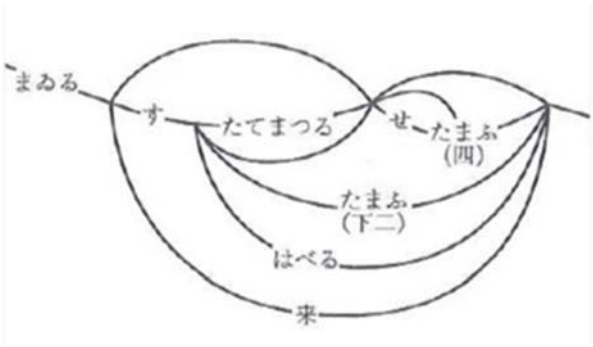
これら明治時代の用例を見ると、すべて「べし」を介在しており、前述「なり」と同様になっていることは、興味深い。古典の場合、小田の指摘のように、「ず＋べし＋ず」のように同じ成分を反復させることが文節長の長さに繋がっているようである。

5　文節形成の数理的分析

さて、文節の形成の数理的分析としては、水谷静夫（1974）に先駆的な研究がある。水谷が『国語学五つの発見再発見』で詳説しているように、本居春庭に始まる江戸時代の動詞活

用研究を、数理言語学の立場から考え直し、同書 3.3.4 節で「有限状態アクセプタによる春庭理論再構成」「国語で活用語に始まる部分連糸の結合の作りの良さ（合文法性）は、有限状態アクセプタでチェックできる」としている。

「有限状態アクセプタ」とは「有限オートマトン」とも言うが、ここで言うのは、次のように、順番に遷移する状態で、動詞や助動詞の連接を記述できるという数理的モデルである。（図は、古典語の敬語動詞のオートマトン）



既に、チョムスキーが明らかにしている通り、「有限オートマトン」で受理できる形の文法は、「正規文法」（3 型文法）であり、「文脈自由文法（句構造文法）」（2 型文法）よりも制約が強い。一例を挙げれば、句構造文法では、いわゆる入れ子型に間にどんどんと要素を埋め込むことが可能であるが、正規文法では、繰り返しで要素を織り込んで続けることは出来るが、中間に埋め込むことはできない。用言構造体の形成で例をあげると、動詞のあとに「食べ＋させ＋られ」のようにヴォイス要素が重なることはあるが、これはヴォイス範疇での単なる繰り返しである。この水谷の指摘は、その後、工学系の自然言語処理でも普通に継承されており、例えば、長尾真（1983）でも、日本語の動詞文節は「非決定性有限オートマトン」で記述できるとされている。（「非決定性」とは、ある状態から他の複数の状態に遷移する経路があることを意味する。）その後の、各種自然言語処理でも応用されている。

このことはあまりにも当たり前すぎるせいか、日本語の文法記述の中であまり重要視されていないように思われるが、このように、コンピューターで文節の姿を記述してくると、非常に重要視すべきことであると考えられる。

6　日本語文法の 2 つの分野

以上の様にみてくると、日本語の文法を、数理的に、形式言語という観点から見ると、

1　文節を組み合わせる、句構造文法（2 型文法）
2　文節を作り上げる、正規文法（3 型文法）

の 2 種類の異なった形式言語からできていることがわかる。今回は 2 について、そのできあがりの長さに注目したわけである。用言文節＝用言構造体の文法は、それがよりセンテンスに近い性質を持つ言語（すなわち抱合語ということになるが）では、文法そのものである。宮岡（2002・2015）の指摘のように、日本語の用言構造体はそれに類似した性格を持っている。したがって、日本語文法の正確な全面的な記述のためには、句構造文法の部分だけでなく、正規文法で記述できる文節の構造部分（通常の形態論的部分）にも注目しなくてはならないと考えられる。今回は、詳細には触れなかったが、文節長が長いものに注目すると含まれる「という」「ればいい」「かもしれない」などの複合辞は、もともと項を持った句構造や複文の構成要素である。それが、句構造文法の自由さを失って正規文法の枠に押し込められてしまったものと言える。「複合辞」を、形式言語的に、そとから規定すれば、そのように捉えることができるだろう。その記述のためには、本研究で示したように、現実の文節の構造の在り方をもとにした考察が必要だ。現代語で、母語であっても、内省はできない性質の「文法」であることに特徴がある。

なお、上記の現象には、古典語・現代語を通して、唯一の例外現象が存在する。それは、すでに水谷（1974）が指摘しているように、古典語の係り結びである。用言構造体の中で、「結び」の連体形・已然形だけは、遠く離れた句構造の影響を受けて変化する。また、動詞の中に割り込んで、かつ、動詞末を連体形等にする（「思ひぞする」等）ことなども、正規文法では記述できない。この極めて例外的な現象が現代語では消滅したのは理由のあることと思われる。

（参考文献）

小椋秀樹他（2001）『『現代日本語書き言葉均衡コーパス』形態論情報規程集第 4 版（上・下）』（国立国語研究所）

小田勝（2020）「『実例詳解古典文法総覧』補遺稿 49 回」（和泉書院ウェブサイト掲載）

近藤泰弘（1979）「構文上より見た係助詞「なむ」―「なむ」と「ぞ－や」との比較」（『国語と国文学』56-12）

近藤泰弘・みゆき（2001）「平安時代古典語古典文学研究のための N-gram を用いた解析手法」（『言語処理学会 2001 年度年次大会発表論文集』）

近藤泰弘（2020）「歴史に見た日本語の文節長について」（国立国語研究所・通時コーパスシンポジウム 2020）

水谷静夫（1974）『国語学五つの発見再発見』（創文社）

長尾真（1983）『言語工学』（昭晃堂）

宮岡伯人（2002）『語とはなにか　エスキモー語から日本語を見る』（三省堂）

宮岡伯人（2015）『語とはなにか・再考　日本語文法と「文字の陥穽」』（三省堂）