2018.9.29. 計量国語学会 第62回大会(於：京都教育大学) 予稿

漢文訓読をあらわす三つのモデルとチョムスキー階層

島野達雄

1.返り点の働きとリング表示

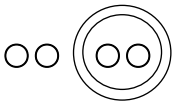
1-1 百聞不レ如二一見一(百聞は一見に如かず)は,⑴百聞と一見は漢字2字を続けて読み,⑵如は一見をさきに読み，不は如二一見一をさきに読む，という二つのルールにもとづいて読み下す．⑴のルールは漢字どうしを「順読」し，⑵は漢字1字をあとから「返読」する．

ここで漢字1字を丸カッコの対(つい)として()であらわし，順読する漢字2字は()()とならべ，返読する場合は，さきに読む漢字列を()で包むことにすると，百聞不レ如二一見一は，()()((()()))とあらわせる．このように返り点つき漢文を丸カッコであらわすことを，丸カッコの対の上部と下部をつなぐとリング(輪)になるので，漢文訓読の**リング表示**とよぶ(右図)．

【百聞不レ如二一見一のリング表示】

()()((()()))

○○((○○))



1-2 漢文訓読は歴史的な産物であり，レ点・一二点・上下点などの返り点は，自然言語である漢文(文言文)の語順を，自然言語である日本語の語順に置き換える役割をしている．返り点はいわば**自然メタ言語**である．

そのため有朋自遠方来には，有レ朋自二遠方一来(朋有り遠方より来たる)，有下朋自二遠方一来上(朋の遠方より来たる有り)の二通りの読み方がある．本稿では，有レ朋自二遠方一来は(())(()())()，有下朋自二遠方一来上は(()(()())())と別々のリング表示であらわす．

また，使子路問津焉には，現行の使二子路問一レ津焉，および歴史的な使役形特有の下点と二点を複合させた使下二子路一問上レ津焉の二通りの返り点のつけ方があるが，どちらも「子路をして津(しん)を問わしむ」という日本語になる．本稿では，このような**異点同順**の例は，どちらも(()()(()))()と同じリング表示であらわす．なお，黙字の焉は，日本語にはあらわれないが，リング表示には存在していることを注意しておく．

1-3 **レ点**は，原則としてレ点をはさむ前後2漢字の語順を入れ替える働きがある．以レ友輔レ仁(友を以て仁を輔く) のリング表示は(())(())，レ点が連続する不レ踰レ矩(矩を踰えず)は((())，匹夫不レ可レ奪レ志也(匹夫も志を奪うべからず)は()()((()))()になる．

原則として，と述べた理由は，たとえば百聞不レ如二一見一の不のように，如二一見一という「漢字と返り点が混じった列」をさきに読むケースがあるからである．

1-4 一二点(一二三…)，上下点(上(中)下)，甲乙点(甲乙丙…)，天地点(天地(人))などは**大返り**とよばれ，一・上・甲・天点がついた漢字を読んだあと，二・下・乙・地点がついた漢字1字を読む，という原則がある．

博学二於文一(博く文を学ぶ)は()(()()), 遠望二田間有二一馬一(遠く田間に一馬有るを望む)は()(()()(()())になる.

上下点の中に一二点がある不レ以二千里一称レ上也(千里を以て称せざる也)は(((())())())(), 甲乙点のなかに上下点, 一二点がある欲レ乙得レ下備二学徳一者レ上友甲レ之(学徳を備えし者を得て之を友とせんと欲す)は((((())())())(()))になる.

文末の「友甲レ之」には甲点とレ点の**複合返り点**がついており, 友と之の 2 漢字の語順を入れ替えたのち, 乙点のついた欲に返っている.

2.返り点だけを用いるモデル

2-1　前章で述べたレ点および一二点, 上下点などの返り点だけを用いる場合の, すべてのリング表示の集合{(), () (), (()), () () (), (()) (), …}を作り出す仕組みを, **形式言語学**の手法を用いて述べてみよう. (注：漢字数ごとのリング表示の数はカタラン数になる.)

その仕組みは 3 つの規則からできている.

規則 1　$S \to ab$

規則 2　$S \to aSb$

規則 3　$S \to SS$

矢印→は「書き換えることができる」と読む.

規則 1 は「Sはabに書き換えることができる」, 規則 2 は「SはaSbに書き換えることができる」, 規則 3 は「SはSSに書き換えることができる」と読む.

Sは**開始記号**とよび, この仕組みの出発点を意味するとともに, 変数の役割をしている.

たとえば, 規則 1 を使うと, abが導かれる. 規則 2 と規則 1 を使うと$aabb$が導かれる. 規則 3 と規則 1 を使うと$abab$が. 規則 1, 2, 3 を組み合わせてを使うと, $aabbab$が導かれる. 最後の例は, $S \Rightarrow SS \Rightarrow aSbab \Rightarrow aabbab$と導出される.

このようにして生まれる文字列(**語**(word)とよぶ)のaを左カッコ(　に置き換え, bを右カッコ)に置き換えるとリング表示になる.

規則 1, 2, 3 を繰り返し何度も適用すれば, すべてのリング表示を導くことができる.

2-2　形式言語学では, 集合$\{a, b\}$を**アルファベット**(**終端記号**ともいう), うえの 3 つのような書き換え規則をまとめて**文法**とよび, 文法によって生成されるすべての語(word)の集合を**言語**とよぶ.

ここでは, 規則 1, 2, 3 を文法G_0とし, G_0によって生成される言語を$L(G_0)$と書く.

$L(G_0)$は同じアルファベット$\{a, b\}$から次の文法G_0'によっても生成される.

規則$1'$　$S \to \varepsilon$

規則2　$S \to aSb$

規則3　$S \to SS$

規則$1'$のε(イプシロン)は**空語**(empty word). 「何も並べない」ことを意味している.

規則$1',2,,3$からも$L(G_0)$が生成される．つまり，一つの言語を生成する文法は，幾つも考えられる．同じ言語を生成する 2 つの文法を**等価**とよぶ．

3.返り点・再読をみとめるモデル

3-1　**再読**とは，未(いま)だ～ず，当(まさ)に～べし，のように，漢字を 2 度読むことをいう．この場合，リング表示を(　)の対ではなく［　］のような別の対にすればよい．

返り点だけを使い，同時に再読をおこなうことを認める場合は，

規則 1　$S \to ab$

規則 2　$S \to aSb$

規則 3　$S \to SS$

規則 4　$S \to pSq$

であらわせる．pとqを［　と　］に置き換えれば，リング表示となる．

規則 1, 2, 3, 4 を文法P_0とし，言語を$L(P_0)$と書く．アルファベットは$\{a,b,p,q\}$の 4 種類．

なお，再読が入れ子になる例は現実にはほとんどないが，規則 4 は猶レ未レ成(なお未だ成らざるがごとし)，つまりリング表示［［()］］が導出されることを認めている．

4.返り点・再読・竪点をみとめるモデル

4-1　これまで見たように，返り点だけを用いるときは漢字 1 字だけを返読する．軽蔑臣下を「臣下を軽蔑す」と訓読するには，軽二-蔑臣下一のように軽蔑の 2 漢字の中央に**竪点**(たててん．ハイフン，漢字連結記号，連続符号とも)を置く必要がある．

竪点を使うばあいのリング表示は，次のようにハイフンを使うのが妥当であろう．

竪点 1 個で 2 漢字をつなぐ，卑二-下之一(之を卑下す)は(-(())-)，患レ所二-以立一(立つ所以(ゆえん)を患(うれ)う)は((-(())-))，此非四吾所レ二-以居二-処子一也(此れ吾れ子を居処せしむる所以に非ざるなり)は()(()(-((-(())-))-))()．この例は竪点 1 個で 2 漢字をつないだ列が入れ子になって 2 度あらわれている．

竪点 2 個で 3 漢字をつなぐ奴二-僕-視之一(之を奴僕視す)は，(-(-(())-)-)．

竪点 3 個で 4 漢字をつなぐ馴二-致-服-習天下之心一(天下の心を馴致服習す)は，(-(-(-(()()()())-)-)-)．

4-2　返り点のほかに竪点 1 個の使用を認める場合は，文法G_0の規則 1, 2, 3 に規則 5 を加えた文法G_1で生成される．アルファベットは$\{a,b,t\}$の 3 種類．言うまでもなくtをハイフンに置き換えれば，リング表示になる．

規則 1　$S \to ab$

規則 2　$S \to aSb$

規則 3　$S \to SS$

規則 5　$S \to ataSbtb$

規則 5 は，「Sを(-(と)-)で囲んでもよい」ことを示している．

2 個以上の竪点で 3 漢字以上をつなげるときは，「竪点が 1 つあれば，もう 1 つ追加してもよい」ことを規定する規則 6 を追加する．

規則 6　$\mathrm{t}aSbt \to tataSbtbt$

規則 6 は，「taとbtのあいだにあるSは，$taSbt$に書き換えてもよい」ことを示しており，たとえば，漢字 1 字を竪点 2 個でつながれた漢字 3 字の熟語で返読する奴_=-僕-視之_は，規則 5,6 をつかって，$S \Longrightarrow ataSbtb \Longrightarrow atataSbtbtb \Longrightarrow atataabbtbtb$と導出できる．

竪点 3 個以上のときも同様に規則 5,6 で導出できる．

規則 1,2,3,5,6 の，竪点の無制限の使用を認める文法をG_Kとし，言語を$L(G_K)$とする．

なお，規則 6 の代わりに

規則v　$S \to \mathrm{t}aSbt$

とすることも考えられるが，これでは$S \Longrightarrow \mathrm{t}aSbt \Longrightarrow taabbt$ つまり-(())-というハイフンに始まりハイフンに終わる語(word)が導出されてしまう．また，変数X,Yを使って，

規則w　$S \to aXSYb$

規則x　$X \to XX$

規則x′　$X \to ta$

規則y　$Y \to YY$

規則y′　$Y \to bt$

を追加しても，X,Yの繰り返しの回数が同じとは決まっていないので，$S \Longrightarrow aXSYb \Longrightarrow aXXSYb \Longrightarrow atataSbtb$ というaとbの個数が異なる語(word)が導出されてしまう．

4-3　さて，規則 6 のかわりに別途、竪点 2 個の使用を認める，

規則 12　$S \to atataSbtbtb$

を設けることもできる．規則 1,2,3,5,12 を文法G_2とよぶ．

竪点 3 個の使用を認める場合は，さらに次の規則 13 を付け加え，文法G_3とする．

規則 13　$S \to atatataSbtbtbtb$

このようにして文法の列$G_0, G_1, G_2, \cdots$，および言語の列$L(G_0), L(G_1), L(G_2), \cdots$が定義できる．

ここで，竪点の使用をm個まで認める文法G_mが生成する言語$L(G_m)$は，mをどんどん大きくすると，「竪点の使用に制限のない，すべてのリング表示」にどんどん近づく．つまり$L(G_m)$はどんどん$L(G_K)$に近づき，文法G_∞と文法G_Kは等価と言える．

4-4　むろん，返り点・竪点に加えて再読を認めるときは，返り点の規則 1,2,3 に再読の規則 4 を加え，さらに竪点の使用を認める規則 5,6 を加える．

規則 1,2,3,4,5,6 を文法P_Kとし，言語を$L(P_K)$とする．

また，規則 12,13…を追加してゆき，m個までの竪点と再読を認める文法をP_m，言語を$L(P_m)$と書く．mをどんどん大きくすると，$L(P_m)$はどんどん$L(P_K)$に近づき，文法P_∞と文法P_Kは等価と言える．

5.チョムスキー階層とオートマトン

5-1　ノーム・チョムスキーは，1956 年に Three Models for the Description of Language（言語をあらわす 3 つのモデル），1959 年に On Certain Formal Properties of Grammars（文法に関するある種の形式的性質について）と題した，形式文法を幾つかの型に分類する論文を発表した．現在ではこの分類を**チョムスキー階層**とよんでいる．

形式言語学における形式文法は，**非終端記号**（開始記号Sをふくむ，変数の役割をする記号），**終端記号**（アルファベット．空語のεをふくむ）の二種の記号と，これらの記号列の有限個の**書き換え規則**（生成規則），の三つにより構成される．

チョムスキー以来の伝統にしたがい，非終端記号を$A,B,\cdots$の大文字，終端記号を$a,b,\cdots$の小文字であらわし，非終端記号と終端記号が混在する（片方だけでも良い．また同じ記号が重複しても良い）記号列をギリシア文字のα,β,…であらわす（ここでは混在列とよぶ）．

チョムスキー階層は次の四つに分かれている．

(1) 書き換え規則の矢印の左辺が常に非終端記号ひとつであり，右辺がa，およびaBまたはBa，という形の，最も厳しい条件の書き換え規則だけを認める文法を**正規文法**，生成される言語を**正規言語**とよぶ．正則文法，正則言語とよぶこともある．

(2) 左辺が非終端記号ひとつで，右辺は混在列，という条件だけの文法を**文脈自由文法**，言語を**文脈自由言語**とよぶ．返り点だけの文法G_0，返り点と再読の文法P_0，および返り点と竪点m個までを認める文法G_m，返り点と再読と竪点m個までを認める文法P_mは，この文脈自由文法に属す．これらの規則の左辺は常に開始記号Sだけになっている．

(3) αAβ → αγβという形の規則をもつ文法は，書き換えが混在列αと混在列βという前後の文脈に依存するという意味から，**文脈依存文法**とよび，言語を**文脈依存言語**とよぶ．返り点と無制限の竪点を認める文法G_K，返り点と再読と無制限の竪点を認める文法P_Kは，規則 6　$taSbt \to tataSbtbt$　がαAβ → αγβ の形をしており，この文脈依存文法に属する．

(4) 最後に，規則にどんな制約もおかない，もっとも緩（ゆる）やかな文法を**句構造文法**，言語を**句構造言語**とよぶ．無制限文法，帰納的可算言語とよぶこともある．

5-2　これらの形式言語は，それぞれ上位の言語の真部分集合であることが知られている．すなわち，

正規言語⊊文脈自由言語⊊文脈依存言語⊊句構造言語

となっており，正規言語は文脈自由言語であり文脈依存言語であり句構造言語でもある．

また，文脈自由言語でない文脈依存言語が存在するが，本稿では，G_K, P_kが文脈自由文法でなく，$L(G_K), L(P_K)$が文脈自由言語でない証明はおこなっていない．

5-3　正規言語は**有限オートマトン**に，言語$L(G_0), L(P_0), L(G_m), L(P_m)$が属する文脈自由言語は**プッシュダウンオートマトン**に，言語$L(G_K), L(P_K)$が属する文脈依存言語は**線形拘束オートマトン**に，句構造言語は**チューリングマシン**に受理されることが知られている．

ここでのオートマトンとは，読み取った語(word) を内部で処理し，受理するか拒否するか，判定できる「仮想機械」をいう．たとえば，$aabb, abaaabbb$など$L(G_0)$に属する「正しくカッコの付けられた語(word)」を受理し，$aab, baab$など$L(G_0)$に属さない語を拒否するオートマトンは，「$L(G_0)$を受理する」という．

5-4　最後に竪点でつながれた漢字列をまとめて漢字 1 字とみなし，卑⚌下之⚌(之を卑下す)を$(-(()))$つまり$ataabb$とあらわすモデル(**偏リング表示**とよぶ)の文法を考えてみよう．

変数Tを導入し，

規則 7　$S \to aTSb$

規則 8　$T \to ta$

規則 9　$T \to TT$

としたとき，規則 1,2,3,7,8,9 の文法G_Tは，G_Kと同様に無制限に竪点の使用を認めているが，矢印の左辺がすべて非終端記号ひとつだから，G_Kと異なり文脈自由文法になる．

再読をふくむ規則 1,2,3,4,7,8,9 の文法P_Tも同様で，言語$L(G_T)$および言語$L(P_T)$は文脈自由言語になり，プッシュダウンオートマトンによって受理される．

6.まとめ

6-1　本稿では，返り点・再読・竪点をふくむ漢文が形式文法であらわせ，生成される形式言語がチョムスキー階層に対応し，オートマトンに認識されることを駆け足で説明した．

今後の**漢文教育**を**情報科学**に結びつける手段として，この漢文訓読の形式文法，形式言語理論が活用されることを期待する．

5-4 で示したように，漢文訓読を対象とする**数理漢文学**には，情報科学の一分野(具体例)として発展する余地が大いにある．

参考文献

(1) 『これならわかる返り点―入門から応用まで―』古田島洋介，新典社

(2) 『漢文訓読入門』古田島洋介・湯城吉信，明治書院，2011

(3) 『計算論とオートマトン理論』A.サローマ，野崎昭弘・町田元・山崎秀記・横森貴共訳，サイエンス社，1988

(4) 『オートマトン・言語理論入門』大川知・広瀬貞樹・山本博章，共立出版，2012