2019.9.21.計量国語学会第 63 回大会（於　国立国語研究所）

漢文の複雑度について

島野達雄（関西学院大学）古田島洋介（明星大学）
湯城吉信（大東文化大学）田村誠（大阪産業大学）

四書五経，左国史漢，唐詩宋詞などの中国古典文は，長らく東アジア漢字文化圏の普遍言語としての地位を保ち，豊富な研究教育資源となっている．我が国では返り点や竪点（たてん．漢字連結記号）を用いる独自の訓読法が発達し，中国古典文とあわせて訓読後の書き下し文をも**漢文**と呼ぶことがある．本稿では，これら二つの表記言語間の語順変換にもとづいて**複雑度**を定義し，著者や成立年代の統計的推定のほか，AI による古典籍の自動テキスト生成や翻訳などの実用に供する．

１．返読する漢字数＝交点数

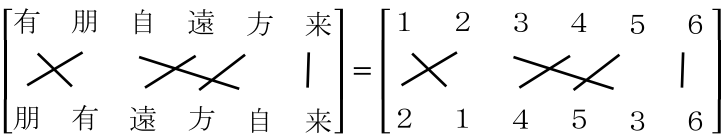
1-1.有朋自遠方来の読み方―返り点だけを使う場合

たとえば，中国語の「有朋自遠方来」に対して，

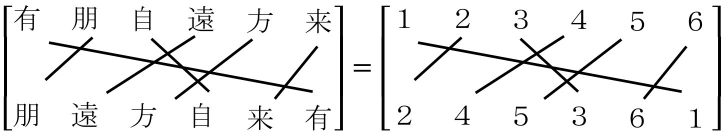
①有$_{レ}$朋自$_{二}$遠方$_{一}$来（朋有り，遠方より来たる．漢字だけを並べると朋有遠方自来）

②有$_{下}$朋自$_{二}$遠方$_{一}$来$_{上}$（朋の遠方より来たる有り．漢字だけを並べると朋遠方自来有）

の二通りの読み方がある．もとの有朋自遠方来に順に 123456 と番号を付けると，①の漢字列は 214536 になる．これを有朋自遠方来を朋有遠方自来に並べ替える K 順列（K は Kanbun の K）とよぶ．ここで，左辺の上段に中国語，下段に日本語漢字列，右辺の上段に 123456 の数字，下段に K 順列を置き，上段と下段の同じ漢字，同じ数字どうしを線分で結ぶと，次のようになる．

①


こうして生じる交点の数 3 を K 順列 214536 の交点数とよぶ．①は有より先に「朋」，自より先に「遠方」の合計 3 字を読み,交点数 3 は返読する漢字数 3 に等しい．同様に②も，

②


となり，有より先に「朋自遠方来」の 5 字，自より先に「遠方」の 2 字，合計 7 字を返読し，やはり交点数 7 が返読する漢字数 7 と一致する．

1-2.欲平治天下の読み方―返り点・竪点を使う場合

③欲$_{三}$平$_{二}$―治天下$_{一}$（天下を平治せんと欲す）の K 順列は 45231，交点数は 8 となる．

③ $\begin{bmatrix} 欲 & 平 & 治 & 天 & 下 \\ 天 & 下 & 平 & 治 & 欲 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 4 & 5 & 2 & 3 & 1 \end{bmatrix}$

ここでは，先に読む天－天，下－下の 2 線分を，後に読む竪点でつながれた平－平および治－治の平行線が横切って交点 4 つを生じている．つまり，先に読む天と下の 2 字に，竪点でつながれた平と治の漢字数 2 を掛けた$2 \times 2 = 4$が，返読する漢字数になっている．さらに欲－欲が漢字 4 字（平治天下）の線分を横切って交点 4 を生じ，合計$4 + 4 = 8$が K 順列 45231 の交点数となる．

２． 重要な注意事項

2-1.交点数に関する和の法則

①有㆑朋自㆓遠方㆒来（朋有り，遠方より来たる．K 順列 214536）を，有㆑朋（朋有り）と自㆓遠方㆒来（遠方より来たる）の二つの文に分け，それぞれの K 順列を 21 と 2314 とすることもできる．このとき，二つの交点数 1 と 2 の和 3 は，もとの K 順列 214536 の交点数 3 に等しい．この和の法則は，どんな長文でも成り立つ．

2-2.交点数に関する積の法則

1-2.で述べたように, $m - 1$個の竪点でつながれたm字の漢字がp字の漢字を返読する場合，交点数は$m \times p$となる．竪点をもつ漢字列どうしが入れ子になっている場合でも，積の法則が成り立つ．

2-3.三本以上の線分は一点で交わらない.

交点数を図示するとき，三本以上の線分は一点で交わらないように描くこととする．数学的な定義は 3-4.に後述する．

2-4.交点数の最小値と最大値

④国破山河在（国破れて山河在り）のように，中国語と日本語の漢字列がまったく同じとき，K 順列は 12345 となり，交点数は最小値 0 をとる．⑤欲㆑不㆑勝㆑簪（簪（しん）に勝（た）えざらんと欲す）の K 順列は 1234 の逆順の 4321 になり，1-1 と 2-2 の 2 線分で交点 1 を生じ，3-3 が 1-1 と 2-2 の 2 線分と交わることで交点 2 を生じ，4-4 が 1-1，2-2，3-3 の 3 線分と交わることで交点 3 を生じ，交点数は都合$1 + 2 + 3 = 6$となる．

同じ論法で，n 字の漢字列に n－1 個のレ点が付いているとき，K 順列は逆順となり，交点数は最大値$1 + 2 + \cdots + (n - 1) = n(n - 1)/2$をとる．

３． 数学的な補足

3-1.順列および順列の積の定義

一般に n 次の**順列**とは 1,2,⋯,n の n 個の数字を並べたものをいう．それぞれの数字を順列の**要素**とよぶ．4字の漢字列 ABCD を CDBA に並べ替えるとき，A は 4 番目に，B は 3 番目に，C は 1 番目に，D は 2 番目に移動する．これを ABCD を CDBA に並べ替える順

列 4312 とよび，$\mathrm{ABCD} \xrightarrow{4312} \mathrm{CDBA}$とあらわす．

続いて CDBA を DCAB に並べ替える（C は 2 番目に, D は 1 番目に, B は 4 番目に, A は 3 番目に移動する）順列 2143 は，$\mathrm{CDBA} \xrightarrow{2143} \mathrm{DCAB}$とあらわせる．二つの並べ替えをまとめて，

$$\mathrm{ABCD} \xrightarrow{4312} \mathrm{CDBA} \xrightarrow{2143} \mathrm{DCAB}$$

とあらわせるが，順列 4312 と順列 2143 の**積**を，最初の ABCD を最後の DCAB に並べ替える順列 3421 と定義し，4312*2143=3421 と書くことにする．

3-2.単位順列と逆順列

一般に 1 から n まで数字が順に並んだ n 次の順列ε_n =12⋯n を n 次の**単位順列**とよぶ．次数が明らかなときは，単にεだけであらわす．単位順列はどんな漢字列に対しても漢字を並べ替えないので，任意の n 次順列σについて，$\varepsilon * \sigma = \sigma * \varepsilon = \sigma$が成り立つ．

ここで, n 次順列σについて, $x * \sigma = \sigma * x = \varepsilon$ となる n 次順列xをσの**逆順列**とよび, σ^{-1}と書く．すなわち，$\sigma^{-1} * \sigma = \sigma * \sigma^{-1} = \varepsilon$ が成り立つ．ここから$(\sigma^{-1})^{-1} = \sigma$ も言える．

たとえば，中国語の有朋自遠方来を朋有遠方自来に並べ替える順列 215346 に対して，

$$\text{有朋自遠方来} \xrightarrow{215346} \text{朋有遠方自来} \xrightarrow{214536} \text{有朋自遠方来}$$

すなわち，215346*214536=123456 が成り立つので，215346^{-1}=214536 となる．①の K 順列 214536 は有朋自遠方来を朋有遠方自来に並べ替える順列 215346 の逆順列である．

一般に, K 順列は, 中国語の漢字列を日本語の漢字列に並べ替える順列の逆順列を指す．

$\mathrm{n} = 1,2,3$のとき, n 次の K 順列の集合は n 次の順列全体の集合に一致するが, $\mathrm{n} \geq 4$では, K 順列の集合は順列全体の集合の一部分になる．たとえば$\mathrm{n} = 7$のとき，順列全体は5040個あるが，K 順列は 948 個に過ぎない．

3-3.二つの順列の距離

二つの n 次の順列$\alpha = e_1 e_2 \cdots e_n$, $\beta = f_1 f_2 \cdots f_n$の**順列間距離**$d(\alpha, \beta)$を次のように定義する[i]．

αのi番目の要素（数字）e_iとβのj番目の要素（数字）f_jが一致したとする．このとき，e_iの右側にある要素の集合$\{e_{i+1}, e_{i+2}, \cdots, e_n\}$と$f_j$の左側にある要素の集合$\{f_1, f_2, \cdots, f_{j-1}\}$の共通集合の要素の数を$d(\alpha, \beta, e_i)$とする．すなわち，

$$d(\alpha, \beta, e_i) = \left|\{e_{i+1}, e_{i+2}, \cdots, e_n\} \cap \{f_1, f_2, \cdots, f_{j-1}\}\right|$$

とする．これは，αを上段に，βを下段に置き，要素$e_i = f_j$について，線分$e_i - f_j$を横切って交点を生じる上段と下段の同じ要素どうしを結んだ線分の数をあらわしている．

たとえば，$\alpha = 34521$, $\beta = 21453$のとき，$d(\alpha, \beta, 3) = |\{4,5,2,1\}| = 4, d(\alpha, \beta, 4) = |\{2,1\}| = 2$, $d(\alpha, \beta, 5) = |\{2,1\}| = 2, d(\alpha, \beta, 2) = 0, d(\alpha, \beta, 1) = 0$となる．これらの合計$\sum_{i=1}^{n} d(\alpha, \beta, i)$を二つの n 次順列α，βの順列間距離とよび，$d(\alpha, \beta)$と書く．

3-4.順列の交点数の定義

とくに n 次の順列ρと単位順列εの順列間距離$d(\rho, \varepsilon)$を順列ρの**交点数**とよぶ．

順列間距離の定義から， n 本の線分から 2 本を選ぶ組合せの数 $_nC_2 = n(n-1)/2$が互いに交わる n 本の線分から生じる交点数になる．これが交点数を図示するとき，三本以上の

線分は一点で交わらないように描く理由．n 字の漢字列の交点数の最大値も ${}_nC_2$ になる．

また，$d(\rho,\varepsilon)=d(\varepsilon,\rho)=d(\rho^{-1},\varepsilon)$が成り立つ．後半の$d(\varepsilon,\rho)=d(\rho^{-1},\varepsilon)$は，$\varepsilon$ =12…n をρ = $e_1e_2\cdots e_n$の数字に並べ替える順列（K 順列）はρ^{-1}になるので，上段に 12…n，下段に$e_1e_2\cdots e_n$を置いた交点数を求める図式において，e_1を 1 に，e_2を 2 に，…e_nを n に「書きかえる」と，上段がρ^{-1}に，下段が 12…n になることからわかる．

3-5.順列の分割

n 次の順列$\alpha=e_1e_2\cdots e_n$の要素e_kの右側にある要素がすべてe_kより大きいとき，$\alpha=e_1e_2\cdots e_k/e_{k+1}\cdots e_n$のように分割して，交点数を「/」の左側と「/」の右側で別々に求め，それらの和をαの交点数としてもよい（2-1.で述べた交点数の和の法則）．

たとえば有朋自遠方来を朋有遠方自来に並べ替える K 順列 214536 は，21/453/6 と分割でき（21，231，1,の三つの K 順列としてもよい），21 の部分の交点数 1，453 の部分の交点数 2，6 の部分の交点数 0 だから，合計して交点数$1+2+0=3$となる．

４． 交点数の応用

4-1.返読率の計算

K 順列から求めた交点数を漢字数で割った値を**返読率**とよぶ．

①有レ朋自二遠方一来（朋有り，遠方より来たる）の場合，交点数は 3，漢字数は 6 だから，返読率は 0.5 となる．②有下朋自二遠方一来上（朋の遠方より来たる有り）の場合，交点数 7，漢字数 6 から，返読率は 1.167 となる．②は，上下点のなかに一二点がある**入れ子構造**をしているので，返読率が①より大きくなり，①より複雑と言ってよいだろう．

⑥伝レ不レ習乎（習はざるを伝ふるか）の交点数は①と同じ 3 だが，漢字数は 4 なので，返読率は 0.75 となり，①の 0.5 より大きく，②の 1.167 より小さくなる．返読率は，漢字数が同じ場合はもちろん，字数の異なる場合の**複雑度**を比較するときの目安の一つとなる．

n 字の交点数の最大値は n(n－1)/2 だから，n 字の返読率の最大値は(n－1)/2 になる．

4-2.杜甫と李白の返読率

荻生徂徠の門人，服部南郭が訓点・解説をほどこした『唐詩選国字解』にある杜甫と李白の，五言の古詩・律詩・排律・絶句，および七言の古詩・律詩・絶句の返読率を示す．

4-2.杜甫と李白の返読率

杜甫	首数	句数	漢字数	交点数	返読率	李白	首数	句数	漢字数	交点数	返読率
五言	23	252	1260	301	0.239	五言	13	94	460	148	0.322
七言	28	306	2123	618	0.291	七言	20	94	658	188	0.286
総計	51	558	3383	919	0.272	総計	33	188	1118	336	0.301

4-3.唐詩選の七言詩

唐詩選の七言の詩だけに限っていえば，K 順列が単位順列ε_7 =1234567，つまり訓読しても中国語の語順のままで交点数が 0 のケースは，杜甫，李白とも出現回数がもっとも多い．

杜甫の七言詩は全 306 句中 96 句(31.4%)がε_7で, 出現回数上位五種で全 306 句の 63.4%を占め，上位五種をあわせた返読率は 0.121 となっている．

李白の七言詩は全 94 句中 30 句（31.9%）がε_7で，出現回数上位五種で全 94 句の 63.8%を占め，上位五種をあわせた返読率は 0.181 となっている．

4-4.論語の諸家点

論語には，①有㆑朋自㆓遠方㆒来，②有㆘朋自㆓遠方㆒来㆖以外にも，同じ原文に異なった返り点・竪点をつけるときがある．以下に，清原家本（正和 4 年，論語集解，東洋文庫藏），道春点（林羅山訓点，天保改正四書集注，国会図書館藏），渋沢栄一（大正 12-14 年，論語講義），岩波文庫（昭和 38 年，金谷治訳注）の四本を素材として，同文異読の五例の K 順列と交点数を示す．

4-4.論語四本の K 順列－交点数

篇名	原文	返り点表示	K 順列－交点数	清原	道春	渋沢	岩波
学而	有朋自遠方来	有㆑朋自㆓遠方㆒来	214536－3				○
		有㆘朋自㆓遠方㆒来㆖	245361－7	○	○	○	
学而	礼之用和為貴	礼之用㆑和為㆑貴	124365－2	○		○	
		礼之用和為㆑貴	123465－1		○		○
里仁	不以其道得之	不㆘以㆓其道㆒得㆖㆑之	342651－8	○		○	○
		不㆑以㆓其道㆒得㆑之	342165－6		○		
子路	言不可以若是其幾也	言不㆑可㆓以若㆒㆑是其幾也	146532789－8	○			○
		言不㆑可㆓以若㆑是其幾㆒也	146578329－12		○	○	
堯曰	所重民食喪祭	所㆑重民食喪祭	213456－1	○		○	○
		所㆑重㆑民食喪祭	321456－3		○		
			交点数合計	26	29	30	21

原文の総漢字数は 33 字だから，返読率は清原家本で 0.788，道春点で 0.879，渋沢栄一で 0.909, 岩波文庫で 0.636 となり，複雑度が時代とともにどのように変化したかがわかる．

５．K 順列・交点数の一般化

5-1.再読文字への拡張

同じ漢字を二度読む再読文字のときでも交点数を定義できる．⑦有㆓未㆑招魂㆒（未だ招かざる魂有り．漢字だけを並べると未招未魂有）は，次のように交点数を求めることができる．

$$\begin{bmatrix} 有 & 未 & 招 & 魂 & \\ 未 & 招 & 未 & 魂 & 有 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 & 4 & \\ 2 & 3 & 2 & 4 & 1 \end{bmatrix}$$

この例では，上段の「未」から出ている 2 線分が交点 1 を生じ，その他の交点 5 とあわ

せて交点数 6 となっている．

ふつう再読文字には，レ点や二点，下点などの返り点が後続する．最初に再読文字を読むときは，この後続する返り点を無視する．二度目に読むときは，返り点の指示にしたがって読む．この例では，未のレ点を取り払った「有㆓未招魂㆒」から，最初に未を「いまだ」と読み，そのあと「有㆓未㆑招魂㆒」とレ点の指示にしたがい，未を「ざる」と読んでいる．

再読現象とは，n 次 K 順列$e_1e_2\cdots e_n$において，要素e_kの左側にあるi個の要素がすべてe_kより大きいとき，つまりi個の要素を「返読」するとき，それらの左側に「初読」に相当する 2 個目のe_kを置くことを意味する．

このように同じ要素が二度あらわれることもある長さ m の n 次 K 順列を，大きさ m の n 次 **S 拡張順列**（S は Saidoku の S）とよぶ．

正確には，S 拡張順列は K 順列を拡張した概念だが，同じ数字が二度登場することもあるので順列とは言えない．なお，7 次の S 拡張順列は，全部で 6078 個ある．

5-2.S 拡張順列の交点数と返読率

n 次 S 拡張順列αと n 次順列βに関して，順列や K 順列と同様に三本以上の線分は一点で交わらないようにして，**S 拡張順列間距離**$d(\alpha,\beta)$を「交点の数」で定義する[ii]．また，n 次の単位順列ε_n =12⋯n との S 拡張順列間距離$d(\varepsilon_n,\alpha)$を S 拡張順列αの**交点数**とよぶ．この交点数を n で割った値を S 拡張順列αの**返読率**とよぶ．

⑦有㆓未㆑招魂㆒（未だ招かざる魂有り）では，大きさ 5 の 4 次 S 拡張順列 23241 の交点数 6 を原文の漢字数 4 で割り，返読率 1.5 を得る．

5-3.二か国語間の翻訳

中国語の「仁」を英語では「true virtue」と翻訳するように，二つの言語間の翻訳では，対応する単語（語彙）および（句読点などの）記号の数が増減する．

ここで，S 拡張順列の概念を更に拡張し，1,2,⋯,n の n 個の数字を，それぞれ重複を許して全部で m 個並べた数字列を大きさmの n 次 **Q 拡張順列**とよぶ．大きさ m の n 次 Q 拡張順列$\alpha_{n,m}$と大きさ q の p 次 Q 拡張順列$\beta_{p,q}$を考え，S 拡張順列間距離と同じように，同じ数字（数字 1 から，数字 n と数字 p の小さいほうまで）どうしを線分でつなぐと，$\alpha_{n,m}$と$\beta_{p,q}$の**交点数**が求められる．$\alpha_{n,m}$を A 言語の単語・記号列（文），$\beta_{p,q}$を A 言語を B 言語に翻訳した単語・記号列（文）と読みかえると，得られた交点数は，A 言語を B 言語に翻訳したときに生じる複雑さを示す**特徴量**のひとつと言える．

漢文訓読の返り点，竪点，左右の傍訓などの添え仮名は，中国語と日本語という二つの言語の橋渡しをする**メタ言語**であり，歴史的産物という意味から**自然メタ言語**として捉えることができる．世界でも類を見ない 1000 年以上にわたる漢文訓読の伝統を，数学的に定式化し，**情報科学**として進展させることは，目下の急務となっている．

[i] 順列間距離が距離の 3 条件を満たすことは 2019 年 11 月の数学史研究発表会の要旨「漢文訓読の複雑度について―歴史情報学の試み―」の「３．数学的な補足(3)順列間距離の性質」で証明している．

[ii] 再読文字がひとつの場合，n 次Ｓ拡張順列αの再読にあたる要素をｎ+1に書き換えると，Ｓ拡張間距離はｎ+1 次の順列間距離になる．再読文字が複数個あるときも同様．