《短　信》

ネットワーク法における「共有度」と「距離」の定義と計算について

——伊藤隆氏の≪短信≫に答える——

柴田　武
熊谷　康雄

「国語学」第150集に発表した「ネットワーク法における地点間の言語的類似の新しいとらえかたと処理のしかた——言語的特徴による地域分割のためのネットワーク法II——」（以下，柴田・熊谷(1987)と略す）に対して，「国語学」第151集に，伊藤隆氏による「類似度・共有度・距離——柴田・熊谷のネットワーク法との関連で——」（以下，伊藤(1987)と略す）と題する≪短信≫が掲載された。柴田・熊谷(1987)の中では，地点間の言語的な類似の度合を測る尺度として，2地点間の言語的特徴の共有度というものに加えて，2つの地点がそれぞれ示す共有度の値の分布パターン相互間の距離というものを用いることを試みた。伊藤(1987)では，この「共有度」と「距離」という尺度について多少の問題があるとして，その検討がなされている。

伊藤(1987)では，地点ごとのデータの平均が0，分散が1になるように「データの標準化」をすべきことを示唆し，結論として「データを標準化した場合には，言語的類似の程度を測る尺度として「共有度」もしくは「距離」の代わりに，(標準化していない)もとのデータについて計算された相関係数を用いても実質的には大差ないことになる」と述べている。

しかし，この議論は，残念ながら，ネットワーク法において「共有度」や「距離」を用いた意図や定義についての誤解から出発しているもののように考えられる。そこで，以下では，伊藤(1987)に，基本的にどのような誤解があるのかについて述べる。

伊藤(1987)では，「相関係数」とネットワーク法の「共有度」および「距離」が比較されている。「相関係数」を使った計量的方言区画の試みはあるが，今，ここで問題とするのは，相関係数一般ではなくて，伊藤(1987)における「相関係数」の用い方についてである。

まず，ネットワーク法の入力データというのは，言語地図を描いて，意味のある分布を見せた特徴を項目として，それぞれの調査地点について，項目ごとに，その特徴を持つか否かを記録したものである。この種のデータを処理するのに，よく行なわれている方法であるが，ある特徴を持っているということを1で，持っていないということを0で表わすという約束をした。こういう性格のデータ（名目尺度のデータ）については，データの平均とか分散というようなことはありえない。伊藤(1987)では，この1，0を数値と考えて，平均や分散の計算をし，地点間の「相関係数」を求めている。

伊藤(1987)が結論として示している「共有度」「距離」と「相関係数」との間の関係は，「データの標準化」という操作をすることによって得られている。しかし，ネットワーク法

では，「共有度」は，共有度 NC_{ij} というのは地点 i と地点 j とが共有する言語的な特徴の数のことであると定義されている。これは，2つの地点が共通して持っている特徴が多ければ，言語的により似ているものと考えるというものである。そして，われわれの「共有度」の定義式というのは，上に述べたようなネットワーク法の入力データから計算によって「共有度」が求められるように定めたものである。「共有度」の定義式は，データが上述の約束にしたがって表わされているときにのみ，正しく働くものである。しかし，伊藤(1987)には，この点について誤解があるようである。

「共有度」の定義式では，任意の2地点 i と j との間の「共有度」を NC_{ij} とおいて，地点 i が項目 k の特徴を持つならば $L_{ki}=1$ とし，地点 i が項目 k の特徴を持たないならば $L_{ki}=0$ とする。項目数を m とすると，「共有度」は，次の式で定義される。

$$NC_{ij}=\sum_{k=1}^{m} L_{ki}\cdot L_{kj}$$

特徴を持てば1，持たなければ0というようにしてデータを表現しておくと，2つの地点が両方とも特徴を持てば1，特徴を持たなければ0ということが掛け算によって表現できる。このようにして，ある特徴について，ある2つの地点がその特徴を共有しているか否かを示すことができる。こうして，全特徴にわたって，いくつ特徴を共有しているかを数えるには，この各特徴について行なった掛け算の結果を足せばよいということになる。例えば，伊藤(1987)の表1の例として出ているペアAについて言えば，「共有度」の計算は次のようにできる。

$(1\times1)+(1\times0)+(0\times0)+(0\times0)=1$

さて，伊藤(1987)は，式(3)において「共有度」と「相関係数」との関係を導いている。これは単なる式の変形ですむ。しかし，次の段階で，「データの標準化」をして，式(5)を導いても，これは「共有度」に関するものとしての意味はない。個々の地点ごとにデータの平均を0，分散を1に「標準化」するというようなことをして，2地点のデータについて積和を取っても，これは「共有度」を計算したことにはならない。

伊藤(1987)が NC'_{ij} としているものは，「共有度」とはまったく別の量ということになる。したがって，「距離」だとしている DC'_{ij} も意味がない。NC_{ij} と NC'_{ij} とで，これらに共通して保たれているのは，2つの変数を掛けて足しているということである。伊藤(1987)は「共有度」をこのようなものと理解しているようである。しかし，繰り返しになるが，「共有度」の定義式は2つの変数を掛けて足す（積和）というかたちをとってはいるが，このこと自体がデータの表わし方とは独立に，それ自身として「共有度」の定義としての意味を持っているわけではないのである。

ネットワーク法では，まず「共有度」というものを得てから，これを処理しているのであった。これに対して，伊藤(1987)の示しているのは，「共有度」ではない別の尺度で地点間の類似関係を測ろうとしたものだと理解すべきようである。(*) 伊藤(1987)は，式(5)と式

(6)を使って,「相関係数」と「共有度」「距離」との関係を示そうとしながら,実際には別のものについて示しているのである。

以上をまとめると,伊藤(1987)が示している「共有度」,「距離」,「相関係数」の間の関係についての結論は,ネットワーク法の「共有度」や「距離」については成立しない。

(*)伊藤(1987)では「関連性のひとつの尺度である相関係数」というような表現がある。したがって,「相関係数」を関連性を示すものとして使おうとしているかに思われる。そうだとすれば,「相関係数」の値の解釈は通常の意味に従うのだろうか。このとき,伊藤(1987)の「相関係数」は,地点間の言語的類似をとらえるものとして,好ましい形では働いていないと思われる。伊藤(1987)の「相関係数」というのは,2地点 i と j の間で,一方の地点が特徴を示すか示さないかということと,もう一方の地点が特徴を示すか示さないかということとの間にどのくらい関連があるかということを示すものとなっている。

したがって,たとえば,伊藤(1987)の表1のペアBのように,1,0の可能な組み合わせのすべてが同じ数だけ出てくるような場合は「相関係数」は0である。一方の地点が特徴を示すときに,もう一方の地点が特徴を示す場合も示さない場合も同じ割り合いでおこっており,また,一方の地点が特徴を示さないときに,もう一方の地点が特徴を示す場合も示さない場合も同じ割り合いでおこっているということである。しかし,このような意味で関連がないということは,言語的な特徴をどれだけ共通に持っているかということから地点間の言語的類似を見ていこうとする観点からは,相互の類似が0であるということにはならない。

――しばた　東京大学名誉教授,くまがい　埼玉大学研究生――

(昭63年2月16日　受理)