通時コーパスと言語空間論[*1]

山元啓史（東京工業大学/カリフォルニア大学サンディエゴ校）
田中牧郎（国立国語研究所言語資源研究系）
近藤泰弘（青山学院大学/国立国語研究所言語資源研究系）

Diachronic Corpus and Linguistic Space

Hilofumi Yamamoto (Tokyo Institute of Technology/University of California, San Diego)
Makiro Tanaka (Dept. Corpus Studies, NINJAL)
Yasu-Hiro Kondo (Aoyama Gakuin University/Dept. Corpus Studies, NINJAL)

1 はじめに

国立国語研究所共同研究基幹型プロジェクト「通時コーパスの設計」[*2]は古代から近世までのいくつかの時点における代表的資料により、「通時コーパス」のモデルを試作するものである。そこでは、主に１）資料の選定、２）古典本文の電子化と情報（異文・原文表記・異体字・引用・文体など）付与の問題、３）各時代・各文体に対応した形態素解析などの実務的あるいは技術的問題を扱ってきた。

本稿は、同プロジェクトを通して培われた基本的な概念、共時態と通時態、記述言語学とコーパスの関係、可能性と作業モデル、適応例、通時コーパスを活用する上で、今後、必要性が予想される研究領域について議論する。

2 言語の記述と言語の空間

コーパス開発には、テキスト収集、著作権、電子化、提示情報の決定などの実務的な作業が多く、研究アプローチに関する議論が後回しにされやすい。コーパスの開発は手段であり、それを用いて言語の普遍的形式を探求することが目的である。しかし、コーパス言語学においてはその理論的背景となる「言語の記述」「言語の普遍性」「言語の空間」「共時態と通時態」などの点について、あまり議論されてこなかった。この機会を利用し、これらについて検討し、考え方を整理しておきたい。

2.1 共時態と通時態

Saussure (1983) によれば、通時態とは時間の流れにしたがって、変化していく言語のありさまであり、共時態とは言語の一定時期におけるありさまである。つまり、言語をある時の点と見るか、ある時とある時をつなぐ線と見るかということである[*3]。数理的に整

[*1] 本研究は国立国語研究所共同研究プロジェクト基幹型「通時コーパスの設計」（代表者：近藤泰弘）、および、科学研究費基盤研究C「和歌形態素解析用辞書開発のための用語連接規則に関する基礎研究」（代表者：山元啓史）の助成を得た。

[*2] http://historicalcorpus.jp/

[*3] したがって、現代語のみを取り上げた研究であっても、経年的な変化を問題とした分析であれば、たとえ、時間の幅が短くても、それは通時的分析となる。時間的な変化を無視し、時間的な変化はないものとするならば、共時的分析となる。日本語の研究においては、一時代一言語を分析の対象とする共時的分析を行い、それらの記述をもとに、時代あるいは言語の間を紡ぐ通時的研究が進められる (服部, 1980, p. 249) とい

理すれば、共時態を示す点の集まりが連続して線形に見える時、それは通時態と考えられる。ひとつひとつの点は静的である一方、その連続した線が見せる軌跡（線の内容）は動的である。この動的なありさまを見るには、共時態の層を幾層にも並べ、各層の差分をとり、その差分を層間の変化量として分析するのである。その際のポイントとしては、体系の差を強調するだけでなく、体系の背後にある共通の原理をも抽出することが重要である（図 1）。こうすることにより、従来、現代語だけあるいは古典語だけを分析していたのでは見えなかったものが見えてくるものと考える (近藤, 2000, p. 80)。

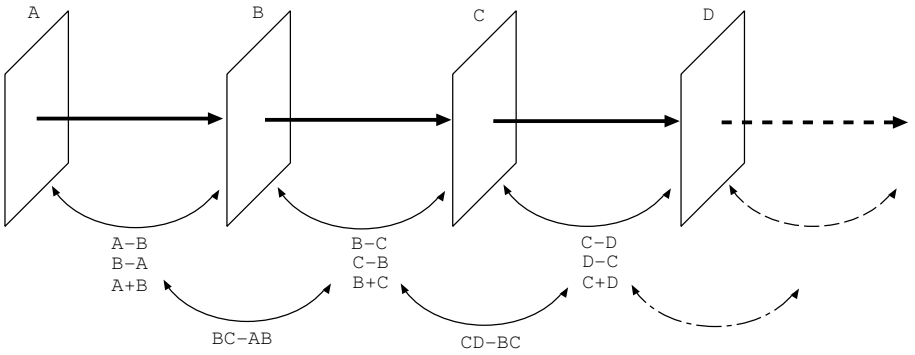

図 1　共時態の各層から差分をとる：たとえば、ＡＢＣＤは時間軸に並べられた任意の資料。差分をとるだけでなく、両者の体系に共通の原理を抽出し、その抽出したものをさらに隣接の抽出したものと比較していく。

2.2　通時コーパスと言語の記述

われわれの脳裏に潜在する langue（言語構造）は、直接観察できず、言語処理の結果として出力された音声や文字列、parole（言語現象）によってしか分析ができない (Saussure, 1983)。そのため、言語現象からさまざまな推論を働かせて、分析する必要がある上、古典語の場合には、その当時の注釈書や古辞書など多角的な資料を利用して分析を進めなければならない。また、現代人には古典語についての内省がなく、その分析には現代語からの推論を利用するため、現代に生きる観察者の知識に依存する要素も多く、観察者の間に差違が認められ、どの基準によれば普遍的な姿としてとらえられるのか、なかなか決定できない[*4]。

一方、コーパスとコンピュータ処理によれば、従来、研究者がテキストを主観的に見て観察していた研究方法から、あらかじめ研究者が設定した客観的なルール（仮説）がテキスト全体にわたって（網羅的に）当てはまるかどうかを徹底的に調べ上げる方法にかわる。これは見ればわかるテキストを、あえて見ないことにより、現代のわれわれには通常認知できないタイプのデータの構造性や規則性を厳しく探り出すことができる。それにより、内省に代わる感知の機構を手に入れることができるのである (近藤, 2001, p. 35)[*5]。

われているが、服部 (1980, p. 230) によると、「上代から現在に至るまでの日本語の変遷史の研究は組織的に行われてきていない」とある。おそらく現時点においてもその事情は同じであろう。

[*4] 服部 (1980, p. 249) は「我々の言語活動は、我々の脳裏に潜在すると推定される langue に支配されているために、或種の特徴がそこに繰り返し現れるものと考えられる。しかしながら、このような langue は、現在の所外部観察することができず、内部観察もほとんど不可能である」と述べている。

[*5] たとえば、近藤 (2000, pp. 301–11) は大量言語処理による観察を通して、内省だけではとらえにくい「のが」「ことが」によって示される名詞節の性質を明らかにし、その記述に成功している。

テキスト自身は静的な、ある時点で表現された言語の現象である。動的な姿をとらえるには、任意のテキストを多重に比較し、その変化量を分析しなければならない。そこで、比較の計画が重要になる。通時コーパスでいえば、図 1 に示すように作品 A〜D を連続したものとしてとらえる仕組みが必要となる。語彙（語種）の場合なら、A–B, B–C, C–D のように互いの 2 者間をとりもつシソーラス（ブリッジシソーラス）を作成し、シソーラス間の差分をとり、変化量とする。文法を明らかにしたいなら、シソーラスの代わりに連接の出現パターンの一覧表（ブリッジテーブル）を作成し、テーブル間の差分を分析することになろう[*6]。いずれの場合も、いったん共時態の間をつなぎ通した上で、時間軸でとらえた動的なマトマリを記述していく[*7]。こうすれば、それぞれの要素は目で見てわかる状態となり、どの段階でどのような要素が変化したのかが考察できよう。

さて、時間軸を紡ぎ、内省を網羅的大量処理で補完することによって準備ができたら、つぎに必要なことは何だろう。おそらく結果をどのように出力するかということであろう。

2.3 言語空間論

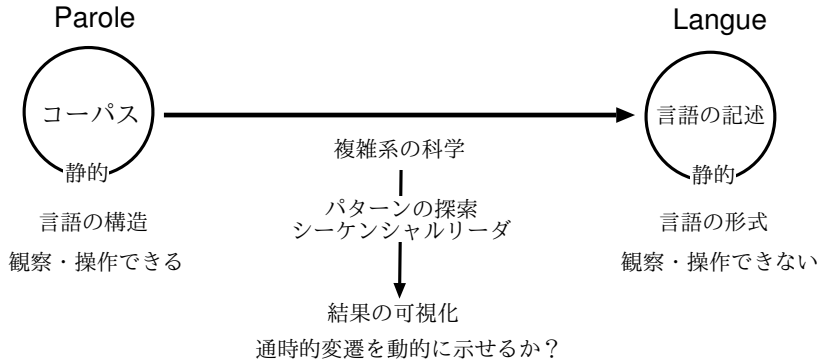

図 2　コーパスと記述、langue と parole: 一般的に記述されたものは静的ではあるが、言語の存在自体は常に変わりつづける動的なものである。その動的な記述はどうすればできるのであろうか。言語の要素はさまざまなものからなり、コーパスにて観察できる姿は複雑多岐にわたる要素が絡み合った現象である。

言語が固定的に見えるのは資料の上だけであって、実際の言語は絶えず揺れ動いている。このような動的なふるまいを明らかにするには、静的な複数の資料を比較し、その差分を分析することによって、動的な変化を見ることは前に述べた。データから得られる動的な変化を分析することにより、過去から現在に至る普遍性だけでなく、さらに将来への予測も期待できる。

言語は時間によって変化するだけでなく、時間の経過につれてヴァリエーションが生じ、それが地域に伝播することによって、言語の地域差が生じる。「地域」は地理的な一要因に過ぎず、文化や産業などさまざまな場面においてもそれぞれのヴァリエーションが生じる。「時間」とともに「空間」が成長し、その中で言語は揺れ動き、変化しつづける。言語の分析として、従来の静的な記述だけでなく、動的な振る舞い（ある部分は枝分かれ

[*6] 後でも述べるが、作品ごとで形態素解析辞書を作成し、その辞書差分をとってもよい。しかし、形態素解析辞書による場合は、隣接パターンに限られるだろう。

[*7] もう少し細かくいえば、資料間の差分がを最も小さくなるようにするために、資料の多重比較を行い、最も近く隣接する資料の間を線で結んでみて、だいたいなめらかな線となるようにモデルを作るのである。

をし、代謝を繰り返し、一部は成長し、またある部分は衰退していく様子）が記述でき、近い将来の言語のふるまいも動的に記述できるはずである。

言語の科学性に関する議論はほとんどの言語学の入門書にあるが、そのいくつかは言語を生物の営みに喩えている。言語の進化や普遍性を考えるために、ここでは生物学での成果を利用することを考える。たとえば、Dong and Searls (1994) は言語学の句構造規則を用いると、遠く離れたところであっても、ある 2 組の DNA のパターンが引き合い、ある細胞を構成する要素になると説明している。

2.4　シーケンシャルリーダ、コーパスロボット

生物学（遺伝子）では、DNA の 4 つのアミノ酸の配列の並びを調べるコンピュータプログラムが数多く公開されている (Dong and Searls, 1994)。これと同じ原理で、コーパスの文字列に見られる言語パターンを調べるコンピュータプログラムを考えてみる。これは研究者の仮説にもとづいて、コーパスの文字列を行き来しながら、何回でも瞬時に仮説を検証することができる機械（コーパスロボット）である[*8]。

開発途上あるいは更新中のコーパスに対応させるには、コーパスの進化に応じた動的なパーシングシステムを考慮しておけばよいだろう。その理論の構築および内部仕様の決定は重要かつ慎重に行われるべきであろうが、同時に実に楽しくなりそうな仕事でもある。コーパスの完成を待たなくても、テキストが質・量ともに充実するにしたがって、コーパスロボットを使って、仮説を立てては何回でも瞬時に検証する仕組みができるならば、言語研究者にとって、それはパワフルで魅力的なものになるにちがいない。次節では、そのための差分の要素（あるいは微分？の要素）はどのように整理すればよいかについて考えてみたい。

3　差分の方法

さまざまな局面で差分を抽出する方法が考えられようが、ここでは任意 2 つのテキストの間の差分をとる方法について簡単に述べる。従来のテキストの比較における問題点のひとつは、テキスト間の内容的な異なり（話の内容）と、言語的な異なり（用語の流行り廃り）を理論的に区別できていないことである。たとえば、ある原作に対する 2 つの翻訳間や原作と現代語訳の間の比較分析を行った論文には、調べたいことが言語の変化にあるのか、翻訳の過程でそうせざるをえなくなったのか、が曖昧になることが多い。これは分析を始める前から研究の枠組みとして区別されていないのである。

田中 (2011) は今昔物語集とその典拠との対応（日本霊異記／宇治拾遺物語）の中で系列比較モデルによる漢語と和語の比較分析を行っている。中国の仏典（法苑珠林：漢文）と今昔物語（和漢混交）を比較すると、漢語がそのまま取り入れられていたり、まったく受け入れられずに捨てられていたり、少々形を変えて、取り入れられていたりなどして、最終的に今昔という形でまとまる。一方、今昔と宇治拾遺物語（ほぼ同じストーリーで和文体）を比較すると、和語がそのまま取り入れられていたり、まったく受け入れられずに捨てられていたり、少々形を変えて、取り入れられていたりなどして、これまた結果的に今昔としてまとめられる。いずれの場合も、2 作品の差から、1）言語の変化により今昔では適宜変更が加えられたもの、2）翻訳者が何らかの基準でことばを取捨選択したもの（翻訳態度）の 2 種が分類できる。従来の比較研究では、このような言語変化による要因

[*8] パターンの探索の点でいえば、DNA の研究がアイデアの発端となるかもしれないが、文字列の数理文法の点でいえば、水谷 (1982, 1983, 2005) などの研究によるところが大きい。

と翻訳者の取捨選択による要因の区別が研究の計画時から一貫しておらず、言語変化のつもりで取り出したデータの中に明らかに翻訳者の操作によるものが紛れ込んでしまうことがあった。上記を研究の枠組みとして、弁別・整理できるように構成したものが系列比較モデルである[*9]。

図 3 は、任意点の時間軸上にある資料を比較する方法を示したモデルである。説明の都合上、任意の 2 点間の違いに限って説明するが、分析の対象は 2 点に限らなくてもよい。A と A' は同じ系列[*10]の言語資料である。A が発生した時を t_1、A' が発生した時を t_2 とする。A と A' の関係は、ある時代の源氏物語の写本とそののちの時代の同作品の写本としてもよいし、1990 年代のプロ野球の実況中継録音と 2000 年代のそれとしてもよい。対応の程度は A と A' の内容をどう捉えるかによる。

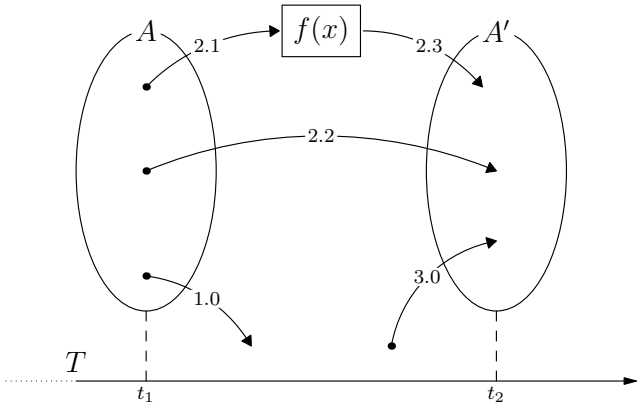

図 3　系列比較のための変遷要素の差分モデル: A は t_1 の時に発生した、あるまとまりを持った内容、A' は t_2 時に発生した、A に対応するまとまりを持った内容。T は時間軸。$f(x)$ は A の任意の要素 x を A' の要素とするための関数。

A に含まれる要素が A' に含まれないことがある。これを 1 類 (1.0) と呼ぶことにする。逆に、A に含まれない要素が A' に含まれることがある。これを 3 類 (3.0) と呼ぶことにする。A にも A' にも含まれる要素がある。これを 2 類と呼ぶことにする。2 類にはまったく同じ要素が A と A' に含まれるものと、A と A' で対応する要素に多少の変動が認められるものがある。前者を 2.2 系とし、後者のうち、A に見られるものを 2.1 系、A' に見られるものを 2.3 系とする。A から A' への時間経過において、取り除かれる要素（1.0)、継続されるが変換の必要な要素（2.1)、そのまま継続される要素（2.2)、継続される際に変換された要素（2.3)、新たに加えられた要素（3.0）の 5 区分で要素の分類を行う。

$$A = \{1.0, 2.1, 2.2\} \tag{1}$$
$$A' = \{2.2, 2.3, 3.0\} \tag{2}$$

A に含まれる要素は、1.0, 2.1, 2.2 の 3 種類、A' に含まれる要素は、2.2, 2.3, 3.0 の 3 種類である。t_1 と t_2 が限りなく近い場合、t_1 と t_2 は共時の資料として捉えてもよい。その場合には、2.x（同じ語、類を同じくする語）が主となり、1.0 や 3.0 の要素が減少する

[*9] 田中 (2011) の例は時間軸というよりも翻訳間という要素で説明した方がよいので、むしろ付録に示す図 4 にしたがうべきである。ただし、これにはまったく時間軸の要素がないわけではない。

[*10] 幅がある概念としておく。狭くは同一内容、原文と翻訳、広くは同じジャンルとする。後に説明する共時モデルにも関わってくる。

が、同一テキストの物理的なコピーや同一内容の録音資料のダビングでもない限り、それらがなくなるわけではない。共時と捉えて分析する資料であったとしても、別々に発生したテキストであれば、時間のずれ（ごく微量な時間経過を伴う要素あるいは単なる言い換え）がある。一方、t_1 と t_2 が限りなく離れている場合には、2.2 が少なくなり、1.0 や 3.0 の要素が増加することが予想されよう。言語の変遷をぜひ見てみたいと思う研究者にとっては、２類 (2.1, 2.2, 2.3) の各要素の分布は注目に値することだろう。

この方法の問題点としては、今、分析している要素が、次の時代に用いられなくなった要素 (1.0) なのか、あるいは何らかの変換が施されて引き続き用いられている要素 (2.1, 2.3) なのか、この両者の判別がつきにくくなることである。しかしながら、表記や読みに関わるマイナーな違いのみを２類の 2.1 あるいは 2.3 とするような厳しい基準であったとしても、この方法で、新たにわかることは多いのではないかと考えている。本稿では、共時的な比較については詳説はしないが、付録にて通時軸を共時軸に置き換えたモデル（図 4)、共時軸を横軸にして通時と共時を一緒に示したモデル（図 5）も紹介しておく。

この系列比較モデルにしたがって、語の弁別に十分利用できるシソーラスを用いたコーパスロボット（任意の t_1 と t_2 における語彙を５区分に自動的に分類するマシン）が作成できれば、近い将来、言語変化の諸相を明らかにしてくれるのではないかと期待している[*11]。

4　今後の研究領域

ここまでの方法を実現するには、必要不可欠な課題がいくつか考えられる。第 1 に処理の単位を柔軟に考えることである。単語の定義は永遠の課題であり、未だ作業的便宜として取り扱われており、確かな理論に基づいて行われているものではない。しかし、これは日本語の問題だけでなくどの言語においても問題とされており、言語学全般に関わる問題である[*12]。語には長いもの短いものさまざまがあるが、短い語は不安定で、長い形で用いられる傾向がある。たとえば、和語については「和語を語形の長さの面からみると、まず『目、葉、荷』のような一拍の語は短すぎて安定しにくく、『はっぱ』『にもつ』のように長い形になって落ち着くばあいもある (西尾, 2002, p. 80)」ということである。また、短い語は多義であるが、長くなればなるほど意味が限定される傾向がある。このことから、おそらく文脈に即してノビチヂミする機構を何らかの方法で開発しなければならないのであろう。しかし、現在のところその有力な手立てはない。

第 2 に形態素解析辞書を一度は資料（作品）別に作って、その資料にとって最も効率のよい辞書を作成し、作品毎の差分・共通を割り出してみることである。それぞれの辞書の連接確率を読むことによって、syntagmatic な側面が、またそれぞれの辞書の語彙差分を読むことによって、paradigmatic な側面が、通時的に把握できるのではないかと考えている。

第 3 に資料（作品）の間を取り持つシソーラスの整備が必要である。系列比較モデルにおいて最も必要なのはシソーラスである。あるトークンが他のトークンと同じであるかどうかを認定するための語彙表が必要なのである[*13]。残念なことに、現状のシソーラスで

[*11] A と A' を系列を同じくする内容を持つものとしたが、共時における系列を異にする A と A' とを比較する際も、上記５区分で系列の異なりを分析することができよう。

[*12] 宮島 (1994, p. 113) は「日本語の語い調査でいちばんこまることは、『単語』という単位が確立していないことである」と述べている。

[*13] 実際には同じトークンであるかどうかを認定することを目標にすると失敗する気がする。生物の世界でも

は、同じ時代の任意の 2 作品（共時的資料）を比べるにも、異なる時代の 2 作品（通時的資料）を比べるにも、単語の長さがまちまちであったり、表記がさまざま（漢字仮名、送り仮名）であったり、上位概念の単語（たとえば花）で、下位概念の単語（たとえば、桜）を示していたりしていて、なかなか 2 作品（資料）の比較が行えない。しかも、単語は時の経過とともに意味を変えることがしばしばあり (Lyons, 1987, p. 212)、作業は困難を極める。意味を決定した符号（たとえば、従来の注釈書で見られるような詳細な意味記述）でデータを処理してしまったら、その符号によってすべての研究の結果は支配されてしまう。したがって、意味的にもきわめて中立的なシソーラスを考案しなければならない。

5　おわりに

本稿では、通時コーパスプロジェクトを進める上での基本的な概念の整理とコーパス言語学の枠組みについて議論し、いくつかのモデルを提案した。しかし、考えれば考えるほど、なさねばならないことは積みあがるばかりである。本稿で取り上げたもの以外にも、音韻、文字、語種など、さまざまな問題があるが、それらに関する通時コーパス流のアプローチは別の機会に考えさせていただきたい。

参考文献

Dong, Shan and David B. Searls (1994) “Gene Structure Prediction by Linguistic Methods”, *Genomics*, Vol. 23, pp. 540–551.

服部四郎 (1980)『言語の本質と機能』，第 1 巻，日本の言語学，第 1 章，大修館書店．

近藤泰弘 (2000)『日本語記述文法の理論』，ひつじ書房，東京．

——— (2001)「コンピュータによる文学語学研究にできること—古典語の「内省」を求めて—」,『文学・語学』，第 171 巻，34–43 頁．特集平成 13 年度夏季大会シンポジウム．

Lyons, John (1987)『Language and Linguistics (言語と言語学)』，岩波書店，第 7 版．

宮島達夫 (1994)『語彙論研究』，麦書房，東京．

水谷静夫 (1982)『数理言語学 (現代数学レクチャーズ D- 3)』，培風館．

——— (1983)『語彙』，第 2 巻，朝倉日本語新講座，朝倉書店，第 1 版．

——— (2005)『言語と数学 POD 版』，森北出版，第 POD 版．

西尾寅弥 (2002)「語種」,『語彙・意味』，第 4 巻，朝倉日本語講座，朝倉書店，第 1 版，79–109 頁．

Saussure, Ferdinand de (1983) *Course in general linguistics...*: McGraw-Hill. tr. of Cours de linguistique generale. from the French by Bally, Charles and Sechehaye, Albert.

田中牧郎 (2011)「平安時代末期における語彙の文体的変異: 同文説話の語彙比較を通して」．第 99 回国語語彙史研究会, 於: 大阪大学．

人間の世界でもそうであるようにまったく同じものは 2 つとして存在しないのであり、厳格に分析するとどこかで違うと判定されてしまうだろう。むしろ、任意 2 つのトークンが同じ類に属するかどうか判別する仕組みを作るべきである。

付録

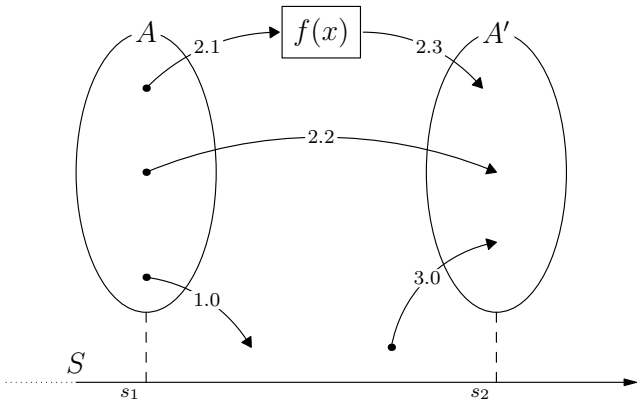

図 4　系列比較モデル（共時）: 通時のモデルの時間軸 T を共時軸 S にしただけである。ただし、T は時間しか表さないが、共時軸 S は、同じ時に発生した同じテキストの異なる言い方や文化、翻訳、方言など、さまざまな場合が考えられる。

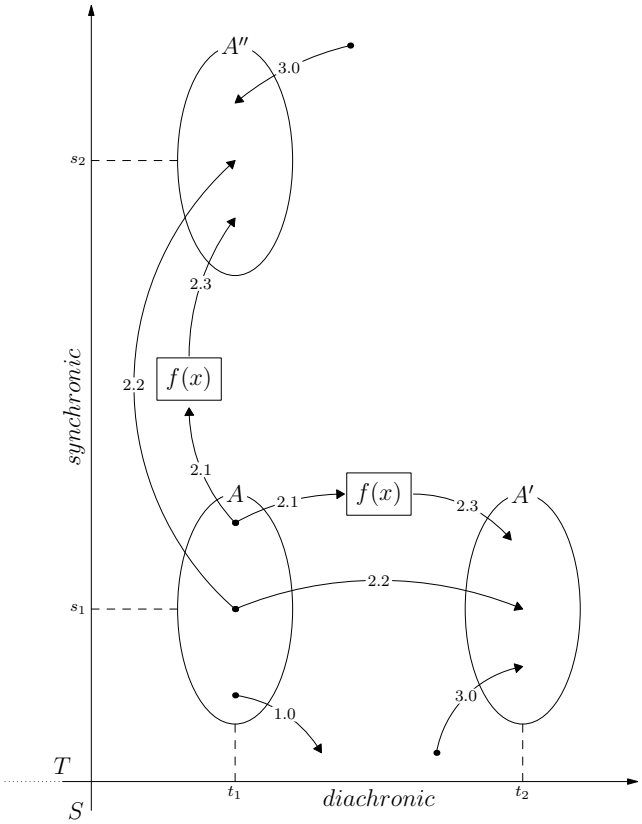

図 5　系列比較モデル（共時／通時）: 縦軸が共時 (synchronic)、横軸が通時 (diachronic)。共時と考えられる関係であっても時間の幅を持つ要素が含まれることもある。