Vol.2009-CH-82 No.10
2009/5/23

言語テクストのデジタルアーカイブズについて

白　須　裕　之[†1]

本稿は言葉によるテクストの意味作用や伝達について、記号論的に再考することで、従来から行なわれてきた全文テクストデータベースや画像データベースではない、テクストの新しいデジタルアーカイブズの可能性を探ることを目的とする。テクストを表現する枠組みとして、アーカイブズされる対象を明確化すること、表現媒体の移行をモデル化すること、及び非言語的な知識を分離すること等が考慮される。

A Semiotic Understanding of Digital Archives of Texts

Shirasu Hiroyuki[†1]

The increasing need for digital archives requires theoretical understanding of their semantics in the information-oriented society. In some papers we discussed symbolic functions of archives in order to present the unified theoretical foundations of archives and digital archives. In this paper, we discuss semiologies for significations and communications of natural languages, in order to present a linguistic model on constructing digital archives of texts.

1. はじめに

現代の情報技術の発展にともなって、従来のアーカイブズにおいても電子化の方法が検討され、また、アーカイブズとは異なった経緯で始まったデジタルアーカイブズにおいても、そのアーカイブズとの乖離が概念的に見直され、アーカイブズの概念や理念とデジタルアーカイブズの技術の融合、及びその捉え直しが始まったばかりと言えよう。

一方、人文学研究へのデジタルアーカイブズの応用研究も端緒に就いたばかりである。従来の紙媒体に頼った研究から、情報技術による電子媒体への研究基盤の移行の可能性は、単なる媒体の変化のみではなく、知識の表現法や研究方法の大規模な革新をも生み出そうとしている。このような状況にあって、人文学はその対象と研究方法の明確化の必要性を突き付けられていると言えよう。

本稿は以上の背景に基づき、言語によるテクストを研究基盤とする人文学において、そのデジタルアーカイブズをどのように捉えるかについて議論する。従来、テクスト表現には、全文テクストデータベースや画像データベースが利用されてきた訳であるが、アーカイブズの対象は何かということを再検討することで、新たにテクスト表現の枠組みを提出しようと思う。また、人文情報学固有の研究対象についても何らかの示唆が得られればと思う。

言葉によるテクストのデジタルアーカイブズ (これ以降、テクストアーカイブズと呼ぶ) について考える前提として、文献[4] で述べたアーカイブズのテーゼから始める。文献[4] では、N. Goodman の記号体系の概念を援用して、アーカイブズの活動とは「表記的体系」を求めることであるとした。これは簡単に述べると、アーカイブズの活動を、アーカイブズしようとしている対象をより良く表現するような記号体系を求めること、に帰着させるということである。「より良い表現」を「表記的体系」という概念として提出したのであるが、これについてはここでは詳しく述べない。但し、言語的な体系は「表記的」でないことは指摘しておこう。従って、アーカイブズしようとする対象の言語学的、記号論的な意味をもう一度検討する必要がある。

本稿の構成は以下の通りである。まず記号論的な道具を導入し、テクストアーカイブズの構築において、その前提条件として検討すべき要因を洗い出す。これらの検討要因の内で特に問題となるテクストアーカイブズの対象について議論するために、テクスト分析の理論によるアプローチを試みる。

2. 記号論的な前提

最初にテクストアーカイブズの構築がどのようなものであるかを考察するために、本節で記号論的な枠組みを設定しよう。記号論は文化現象の様々な側面を理解するための学問分野であり、またテクストの概念も非常に広いものであるが、ここでは言語によるテクストに話しを限定しよう[⋆1]。

†1 東京大学大学院人文社会系研究科 次世代人文学開発センター
The Center for Evolving Humanities, Graduate School of Humanities and Sociology, The University of Tokyo

⋆1 「テクスト」という概念は世界の見方にも関わり、また、これによってテクストアーカイブズの理論的扱いも異なるであろう。後に見るように、本稿では節 5 で述べる言語学的なテクスト概念を使用する。

Vol.2009-CH-82 No.10
2009/5/23

対象である文献としてのテクストは自然言語で書かれているため、自然言語を理解できるような理論的基盤がまず必要である。また、その言語テクストとしてのテクストアーカイブズを計算機で扱うのであるから、自然言語より広い記号現象を扱える理論的基盤が必要である。但し、その記号現象が余り広すぎては細かな議論ができなくなってしまう。この点に関して、文献[13] は「コミュニケーションの記号学と意味作用の記号学」の差を強調する。ここでは「コミュニケーションの記号学」を第二の理論的基盤として設定する。

2.1 言語学的な前提

テクストアーカイブズを言語学的な対象として扱うために、まず言語記号についての前提を確認しておこう[⋆1]。従来、テクストアーカイブズの対象が自明視されている中で、後に議論するようにその対象を明確化しようという営為は、本研究の理論的な基盤を与える。

2.1.1 言 語 記 号

言語へと接近する方法として、文献[11] は言語についての全ての考察、及び全ての言語理論が対決する問題、即ち「音的ないし書記的な質料性が意味を伝達するのだが、この事情をどのように説明するか、という問い」から始める。

言語はコミュニケーションの道具として使われるのであるから、何か伝えたい事柄 ⟨意味されるもの⟩ があり、音や書記等の媒体 ⟨意味するもの⟩ を使用して伝達を実現する。これらの ⟨意味するもの⟩ と ⟨意味されるもの⟩ の二つの面は、どちらも言葉が出現する以前には、不定形な漠然とした連続体であり、事前にはっきりした区別がある訳ではない。「記号」が出現するまでは何一つ画定された現実領域はない[⋆2]。言語は言葉の ⟨意味するもの⟩ と ⟨意味されるもの⟩ を同時に画定することによって、二つの世界の分節化を行なう[⋆3]。

Saussure はこのような機能を「記号」signe という概念として提出している[⋆4]。「記号」は二つの面、即ち意味するものとしての「記号表現」signifiant と意味されるものとしての「記号内容」signifié を持つが、これらは分離不可能であって、「記号」の部分をなす訳ではない。却って「記号」はこれらの二つの面を持つ二重の存在である。しかし、「記号」はそれ自身、単独で立ち現れる実体ではなく、その他の「記号」との関係性によって存在を保つような存在である。この関係性については次の「価値の体系」で述べる。以上のような「記号」の二重性と現実領域の分節の様子を示したものが、図 1 である。中央は言語事象で、複数の「記号」が立ち現れることによって、下側の伝達に利用される音や書記の側の実質と、上側の伝達される内容の側の実質が分節化される。

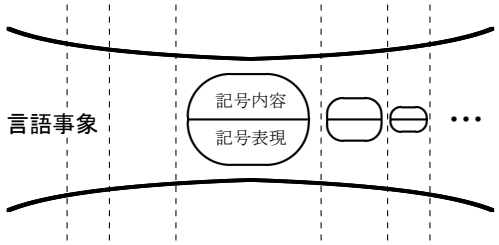

図 1 言語記号の二重性と分節化
Fig. 1 Sign and Articulation

2.1.2 価値の体系

それではそれ自身組織化の原理を持たない二つの実質から、どのように記号を含む言語的な存在が可能となるのであろうか? これを見ることによって、後程、テクストアーカイブズにおいて言語的な事象をどのように記述するか、という問題を議論できる。

ここでは「価値の体系」という概念が中心的な役割を果す。「体系」とは一般に、それに属する個々の要素が相互に関わり合うような総体という意味に使われるが、ここでの「体系」は要素単位が何かの本質によって、実体的に与えられるようなものではない。Saussure は価値が存在するための因子として、以下の二つが必要であると言う。

(1) 一つの異種の事物で、いま価値を決定しようとしている事物と交換できるもの
(2) 幾つかの同種の事物で、いま価値が問題になっている事物と比較できるもの

語のような言語的な存在も、何か異種のもの、即ち観念と交換できるし、同じ性質のもの、即ち他の語と比較することができる。言語的な存在は (1) の意味作用のみで決まるのではなく、(2) における他の存在の否定的な限定による関係性によってその存在の同一性を持つことができる。例えば隣接観念を表現する語は全て相互に制限しあっている。このような情況を「言語は一つの体系であり、その辞項はことごとく連帯的であり、そこでは一辞項の価値は他の辞項たちの同時的現前からしか生じない」と述べる。

⋆1 書物のアーカイブズ一つを取っても、それを芸術作品として、或いは文化人類学の資料として、など様々にアーカイブズの対象とすることができる。本稿では言語学的なテクストとして扱うことを主眼とする。
⋆2 表現すべき概念が言葉以前に予め構成されていて、それに名前を付与する「名称目録」なる言語観を Saussure は否定する。
⋆3 文献[1] では西洋言語学の意味分節理論はそれほど長い歴史を持たないと指摘するが、東洋思想の分節理論との違いは、内容とともに表現も同時に分節されるとすることであろう。
⋆4 本稿は Saussure 学の文献学的な研究ではないので、用語の成立過程やその内容については言及しない。文献[15)11)] 等を適宜参照してほしい。

Vol.2009-CH-82 No.10
2009/5/23

2.2 コミュニケーションの記号学

前節までに自然言語の記号学的な側面について見てきた。テクストアーカイブズを議論するには、記号体系の言語学的な性質を保ったまま、もう少し広い範囲の記号機能を扱えるような体系を議論する必要がある。本節では文献[14] に従って、そのような体系を議論しよう。

2.2.1 信号とメッセージ

メッセージを伝達する上において使用される道具が信号である[⋆1]。発信者がメッセージ伝達を成功するためには、以下の二つの条件が成り立つことが必要である。

(1) メッセージを伝達しようとする意図があることを、発信者が受信者にさとらせること
(2) 発信者の伝達しようとしている特定のメッセージを、受信者は受け取った信号から読みとること

このような伝達には、信号の指示のメカニズムと呼ばれる機能が使用される。これは「特定の信号は許容されるメッセージ群を限定する」というものである。信号によって許容されるメッセージ群は、一つのクラスを構成し、また、その補クラスはその信号によって排除されるメッセージ群を決定する。この相補的な二つのクラスは全体集合としての意味の場を形成する。ある信号によって許容されるメッセージ群からなるクラスを、⟨記号内容クラス⟩と呼ぶ。反対にある特定の⟨記号内容クラス⟩に対して、そのメッセージ群を許容する信号群はクラスを形成し、それを⟨記号表現クラス⟩と呼ぶ。このクラスは補クラスとともに全体集合としての指示の場を形成する。⟨記号表現クラス⟩とこれに対応する⟨記号内容クラス⟩を合せて、⟨統合記号⟩と呼ぶ。

指示の場と意味の場を関係づける⟨統合記号⟩を使ってコミュニケーションを実現できる[⋆2]。図 2 に見られるように、発信者及び受信者の記号行為は⟨統合記号⟩に関して反対方向の選択による。まず発信者は受信者に伝えたいメッセージを、特定の⟨統合記号⟩の⟨記号内容クラス⟩に含まれていることを確認し、対応する⟨記号表現クラス⟩から発信する信号を選択する。信号を受けた受信者は、その信号がある⟨統合記号⟩の⟨記号表現クラス⟩に含まれていることを確認し、発信者が伝えたいと思うメッセージを、対応する⟨記号内容クラス⟩から選択するのである[⋆3]。

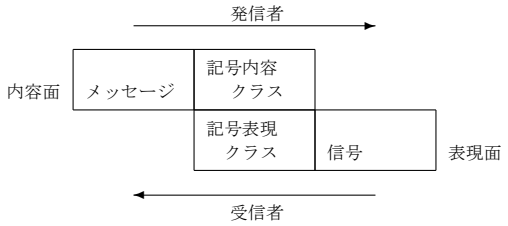

図 2 伝達の仕組み (文献[14] から一部修正)
Fig. 2 Communication

2.2.2 因子クラスと分節

あるクラスが幾つかのクラスの論理積で表現されるとき、その複数のクラスを元のクラスの「因子クラス」と言う。⟨統合記号⟩の⟨記号表現クラス⟩と⟨記号内容クラス⟩が、ともに因子クラスに分解され、因子クラス同士も⟨記号表現クラス⟩と⟨記号内容クラス⟩と同様の関係を持つとき、その分解を「第一次分節」と言い、その結果生ずるものを⟨記号⟩と言う。更に⟨記号表現クラス⟩が分節されるとき、これを「第二次分節」と呼ぶ。これらは自然言語が持つ重要な特性である「二重分節」にちょうど対応している。

本稿ではテクストアーカイブズを扱う体系として、「二重分節を持つような統合記号の体系」を前提とすることにしよう。また論理的には、同様に⟨記号内容クラス⟩が分節されるものを考えて、「第三次分節」と呼ぶことができるが、これは意味論の構築とも関係する。

3. テクストアーカイブズ構築の記号論的な解釈

文献[4] ではアーカイブズを記号体系として捉えたが、アーカイブズの構築過程をモデル化することはできなかった。ここでは前節で述べた「コミュニケーションの記号学」における記号体系を使って、テクストアーカイブズの構築について議論する。

3.1 仮想的な状況

まず最初に以下のような仮想的 (ある意味で理想的) な状況を想定してみよう。

(1) 原テクストにおける言語体系が完全に与えられている
(2) この体系を使用して原作者は原テクストを作成する

⋆1 文献[14] では信号と指標を厳密に区別する。指標と異なり、コミュニケーションの意図があるものが信号であり、この区別はコミュニケーションの記号学の基礎をなす。文献[14] では指示のメカニズムを指標について調べたのち、信号による記号行為について述べているが、本稿では言語を対象としているので、信号のみを扱う。なお言語分析の対象を明確化する上でも、この概念上の区別は重要になってくるであろう。

⋆2 このメカニズムでは伝達の成功失敗、誤解等を説明することができるが、これについては文献[14] を参照願いたい。

⋆3 信号による表意指示だけでは不確定性が残り、その足りない部分を補うのが「情況」である。これについても文献[14] を参照願いたい。

原テクストの作者は読者にメッセージを送るために、自分の使用する言語体系を使って、そのメッセージに最適だと思う原テクストを作る。読者が作者と共通の言語体系を持っている、或いは作者の使用している言語体系を再構築して、それを使用するならば、作者のメッセージを理解できる。これは原作者の使用した自然言語の体系が既知であるという想定である。また、文献[4] でも述べたように、テクストアーカイブズも記号体系として扱われる。従って、テクストアーカイブズ構築には以下の二つの記号体系が関与する (図 3 参照)。

(1) 原テクストの言語体系
(2) テクストアーカイブズの記号体系

テクストアーカイブズの記号体系は、何らかの意味で原テクストの言語体系の表現であるはずである。それでは原テクストをアーカイブズするとは、この二つの記号体系にどのような関係があれば良いのであろうか? 最初に考えられるのは、原テクストの言語体系を使ってアーカイブズを作成することである。即ち原テクストが許容するメッセージと同じメッセージを許容するように、アーカイブズとしての信号を作成すれば良い。しかし、ここには主に、密接に関係する二つの困難がある。

(1) 自然言語の記号体系はデジタルアーカイブズとして扱えない
(2) 記号の二重性が媒体変換を困難にする

一番目は文献[4] でも述べたテーゼを受け入れるならば必然的な事柄である。二番目は前節で述べたように、記号が二つの世界を同時に分節化する二重な存在であるため、名称目録のように記号を扱うことは原理的に不可能であることを示す。即ち、上の想定は内容面のみが決まっていて、表現面だけをテクストアーカイブズ用に作り変え、テクストアーカイブズ用の記号を作ろうとしている。

このように原テクストの言語体系が既知であるような仮想的な場合にも、テクストアーカイブズ用に新たに記号体系を作成しなければならない。どのような記号体系を作成すればアーカイブズを構築したことになるのか、これについては節 6.1 で示唆したいと思う。

3.2 現実的な情況

現実には原テクストを書くために使用した言語体系が、明確であるという状況は仮想的なものであろう。本来はこの体系は潜在的なものであり、また原テクストのみからこれを完全に構築することは難しい。更に文献学的な研究にテクストアーカイブズを使用する場合、原テクストを作成するために使用された体系が曖昧なだけではなく、原テクストそのものが曖昧性を持っている。即ち、テクストアーカイブズには原テクストの再構築という意味も含めなければならない。これには校勘学の営為をモデル化する必要がある。

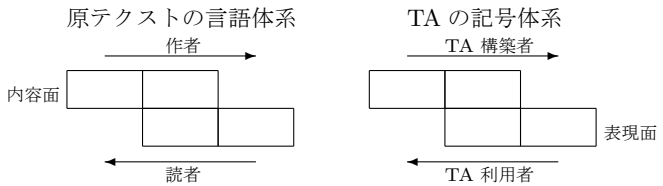

図 3 原テクストとテクストアーカイブズ (TA) の記号体系
Fig. 3 Symbol System for Text Archives

3.3 全文テクストデータベース

従来の全文テクストデータベースの枠組みをテクストアーカイブズとして使用する場合には、通常、原テクストを作成したときの言語体系を問題にすることは余りない。また、信号としての原テクストを符号化文字集合という記号表現に移すだけであり、その移行が何に基づいてなされるのかを考慮することもない[⋆1]。従って、検索一つを取っても、その対象が何であるのか明確ではない。本来、検索を含む問合せには意味論が必要であり、テクストアーカイブズの場合、それは記号体系によって与えられるべきである。

4. テクストアーカイブズの言語学的な検討要因

前節の議論を前提に、テクストアーカイブズを構築する上で、言語学的な考察が何処に出現するかについて纏めておく。

(1) 原テクストの言語体系 — 原テクストは自然言語で書かれている訳であるが、作者が使用した言語体系が存在した筈である。しかし、その体系が原テクストとともに与えられている訳ではないので、テクストアーカイブズの作者がその言語体系を部分的ながら再構築しなければならない[⋆2]。
(2) 媒体変換における言語単位の保存 — 原テクストの表現である記号表現には第二分節が仮定されている。その場合、テクストアーカイブズはこの分節を自然な形で変換することが、文献学的研究に使用する場合に前提とされる。特に漢字文献の場合は文字をどのようにアーカイブズとするかという問題が重要な役割を演ずる。

⋆1 アーカイブズの立場から文字を論じたものに文献[7] がある。また文字の解釈については、例えば文献[5] を参照。
⋆2 テクストの読みにおける解釈学的な問題については別に考察が必要であろう。

Vol.2009-CH-82 No.10
2009/5/23

(3) 校訂テクストの意味 — テクストの校訂という問題を考えるとき、テクストに使用した言語体系抜きには校訂の意味を扱うことはできない。本来、校訂の意味は言語体系を離れて考察することに意味がないが、このことはテクストアーカイブズを校訂の道具として使用する場合に、校訂と言語体系の問題が先鋭化することになる。即ち、校訂テクストというテクスト構築の営為を記号学的に理解せねばならない。

言語テクストを機械的に処理することによって実現される情報抽出や自動翻訳等、自然言語処理の目的には、特定の自然言語に対するある程度完成された文法、即ち、統辞論、形態論及び語彙論的な知識が必要とされる。しかし、テクストアーカイブズに必要なのは完成されてしまった文法知識ではなく、言語分析の方法とその分析過程で得られる知識を記述する枠組みである。テクストアーカイブズでは、完全なる言語記号を記述することができる訳ではなく、言語テクストを研究するための方法論を実現する必要がある。この方式として次節でテクスト分析の理論について検討する。

5. テクスト分析の理論

テクストアーカイブズを実現するために前提となるのは、〈言語学的なるもの〉[⋆1]を画定すること、即ち、テクストアーカイブズの対象となるものを明確化することである。今迄も見てきたようにこの対象は本質的な実体として定義することはできなかった。従って、対象を画定するための分析方法が必要である。本節では文献[12)⋆2]で述べられた言語素論 Glossemantics から、テクストアーカイブズに必要なテクスト分析の理論を議論しよう[⋆3]。

5.1 分析と機能

言語素論では対象をその性質によって記述するのではなく、他の対象との関係性に注目して、形式的に定義する。これは言語記号の特徴を非常に良く表現している。テクスト分析は与えられたデータとしての「テクスト」を出発点とするが、「分析」によって言語学的に関心のある対象を構成する。「分析」の形式的な定義は以下である。

> **分析** Analysis は一つのオブジェクトに対する、他のオブジェクト (複数) の互いの関わりと、そのオブジェクトへの関わりによる均質的な記述である[⋆4]。

⋆1 文献[11)] では、Saussure の業績を〈言語学的なるもの〉の構築とする。
⋆2 原典はデンマーク語である。参考文献の英訳は改訂版であり、文献[13)] によると、著者 Hjelmslev は「テクストを全部校訂し直し、訳者に多くの改良点を示した。」とある。
⋆3 本節で述べるテクスト分析は、Digital Humanities におけるテクスト分析とはテクストの概念が異なる。
⋆4 この形式的な定義に含まれる「記述」「オブジェクト」「均質性」「関わり」等は未定義語である。以下、あらかじめ定義されていない用語は断らない限り未定義語である。また定義される用語を太字で示す。

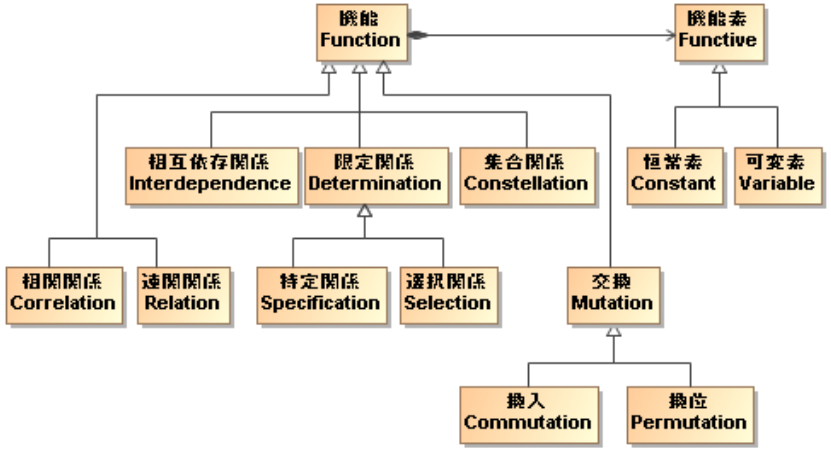

図 4 機能
Fig. 4 Classification of *Functions*

以後、その操作的な内容に鑑みて「分析」を「分割」とも言う。「分割」の対象となるオブジェクトを**類** Class と言う[⋆5]。一つの「分割」で導入されたオブジェクト (複数) を**区分単位** Component と言う。類 (複数) の類を**階層** Hierarchy と言う。階層についての用語を用意しておこう。**分割派生単位** Derivates とは、全く同一の演繹における、ある類の区分単位及び区分単位の区分単位である。また、分割派生単位がその最下位の共通する類に達するまでに経過する類の数を**次階** Degree と言う。

文献[12)] では目指す言語理論のために、「分析」に要求される原理が述べられているが、本稿ではテクストを記述する枠組みを提出することが目的であるので、これについては触れない。そのような「分析」の要求条件を充した関わりを**機能** Function と呼び、機能を持つオブジェクトを**機能素** Functive と言う。テクスト分析の理論では「類」「機能」「機能素」が基本的なオブジェクトであり、これらのオブジェクトの分類が述べられる (図 4, 5 を参照)。機能と機能の間にも機能の存在することがあるので、機能も機能素でありうる。機能でない機能素を**占在体** Entity と言う。

機能素が**恒常素** Constant であるのは、その現存がその機能が持つもう一方の機能素の現

⋆5 関係性に注目するため、「類」と言ってもそれ自身、物を集めたものという訳ではない。

Vol.2009-CH-82 No.10
2009/5/23

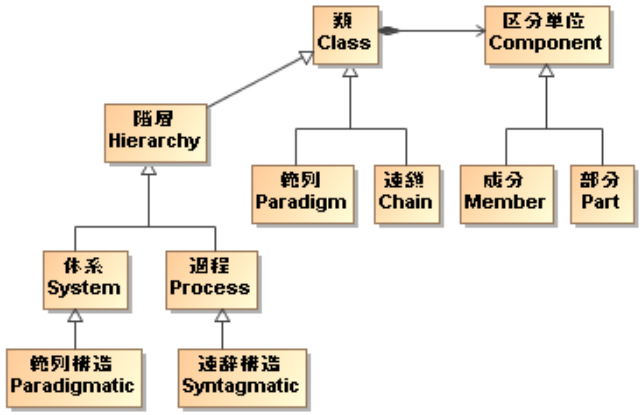

図 5 類, 階層
Fig. 5 Classfication of *Classes* and *Hierarchies*

存の必要条件であるときである。また、必要条件でないとき**可変素** Variable と言う。機能素のこのような分類によって、機能を分類すると以下のようになる。**相互依存関係** Interdependence は二つの恒常素の間の関係、**限定関係** Determination は一つの恒常素と一つの可変素の間の関係、**集合関係** Constellation は二つの可変素の間の関係である。なお限定関係にある恒常素 C と可変素 V に対して、$V \to C$ と書いて、C は V により限定される、或いは V は C を限定すると言う。

複合分割 Analysis complex とは、全く同一の類に対する分割の類のことである。分割が継続的に行われる場合、後の分割は前の分割が行われていることが前提である。この様子を定義したものが「演繹」である。**演繹** Deduction とは、継続的な分割あるいは複合分割で、その分割の間に限定関係があるものを言う。

言語理論にとって重要な機能として以下の区別がある。either-or 型機能と both-and 型機能であり、各々、**相関関係** Correlation と**連関関係** Relation と言う。この多重分類はこれ以降出てくる分類に非常に関係が深く、図 5 における全ての多重分類、また、図 4 においては、相関関係、連関関係より低い位置に図示してある多重分類は、なんらかの意味でこの区別に関係している。例えば、**体系** System は相関関係の階層であり、**過程** Process は連関関係の階層である。階層はこれらの機能分類に従って分類されている。

5.2 記号、表現と内容

ここでの記号機能は表現と内容という二つの占在体の間に想定される機能である。従って、表現と内容は機能素の操作上の単なる名称であり、それ以上の意味はないものとして扱う。**記号言語** Semiotic とは、その全ての区分単位が、互いに連関関係によって定義される類へと分割されるような階層のことである[⋆1]。但し、ここでの**定義** Definition とは、記号内容または記号表現の部分分割のことである。即ち、記号言語の最初の分割は表現と内容への分割を意味する。記号言語の体系を**範列構造** Paradigmatic 、記号言語の過程を**連辞構造** Syntagmatic と言う。また、範列構造内の類は**範列** Paradigm、連辞構造内の類は**連鎖** Chain と言う。

5.3 不変体と可変体

テクスト分析においては、分析の結果で多くの機能素が導入され、それらの目録ができる。従って、多くの機能素を同一視する方法が提供されなければならない。ここで「変換」の概念を導入しよう。**交換** Mutation とは、全く同一の類における一次階分割派生単位間に存在する機能で、その類と同じ序列[⋆2]に属す他の一次階分割派生単位間の機能に対して連関関係を持つものを言う (図 6 参照)。また、範列の成分間の交換を**換入** Commutation、連鎖の部分間の交換を**換位** Permutation と言う。相互に換入を持つ相関体を**不変体** Invariants、相互に代入を持つ相関体を**変異体** Variants と言う。機能素を不変体と変異体に分類することが、テクスト分析の一つの目標である。

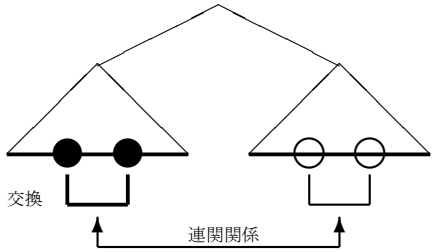

図 6 交換, 換入及び換位
Fig. 6 Mutation, Commutation and Permutation

⋆1 この情況を後で導入する用語「交換」を使って表現すると、これらの類の全てが、相互の交換によって定義される分割派生単位へと分割される場合を言っている。
⋆2 **序列** Rank とは、全く同一の過程、または全く同一の体系に属す同じ次階の分割派生単位 (複数) である。

Vol.2009-CH-82 No.10
2009/5/23

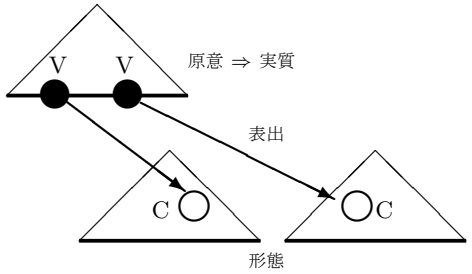

図 7 原意, 形態及び実質
Fig. 7 Purport, Form and Substance

5.4 言語図式と言語使用

Saussure は記号による分節以前のカオス、未分節状態の意味志向を原意 Purport と呼んでいた。この原意の言語による記述について議論しよう。原意は言語学以外の科学によって分析することができるが、それを言語によって表現する方式は以下である。原意に言語以外の方法による分析を施して得られた階層が、「演繹」によって得られた言語の階層に対して機能を持つ場合、言語の階層を「言語形式」、非言語の階層を「言語使用」と呼び、「言語使用」は「言語形式」を表出すると呼ぶ。このようなオブジェクトの形式的な定義を与えよう。

表出 Manifestation とは、階層間の、あるいは異なる階層の分割派生単位間の選択関係のことである。このとき表出の恒常素を**形式** Form、可変体を**使用** Usage と呼ぶ。**原意** Purport とは、二つ以上の連辞構造の中で二つ以上の連鎖を表出する、かつ/あるいは二つ以上の範列構造の中で二つ以上の範列を表出する可変素の類を言う。**言語** Language とは、その範列が全ての原意によって表出される範列構造である。また、**テクスト** Text とは、これに関連して全ての原意によって表出される連辞構造である。このとき、言語である形式を**言語図式** Linguistic schema、言語図式を表出する実質を**言語実用** Linguistic usage と呼ぶ (図 7 参照)。

原意の定義ではテクスト分析が連辞と範列に関して演繹されることを仮定している。これらはテクストと言語体系の分析であり、各々は次に表現と内容に分けられて、分析が継続される。この表現と内容は同じ手続によって並行的に分析される。

6. テクストアーカイブズ構築に向けて

6.1 テクスト分析の理論の意義

前節でテクスト分析の理論について述べたが、ここではテクストアーカイブズを構築する上で果すその意義について議論しよう。以下のように大きく三点に纏めることができる。

(1) 原テクストの言語体系の明確化とその記述
(2) 媒体変換としてのテクストアーカイブズの構築
(3) 非言語的な知識の記述の分離

(1) は原テクストに対するテクスト分析の結果、得られる不変体がアーカイブズの対象となる単位を明確にする。また、表現面と内容面からなる対象の記述が記号体系を構成している。従来、言語学的な対象の扱いについては文法学的なアプローチが主であったが、今まで見てきたように、原テクストの言語体系には不完全で曖昧な点が存在し、そのためテクストアーカイブズ構築においては文法学以前の言語学が必要である。テクスト分析の結果を出発点にテクストアーカイブズを構築すべきである。

文献[2] では歴史記述のテクストを表現するために、「抽象テクスト」という概念を提出した。これは「単位」から構成されるものであるが、この「単位」を定義する方法を提出できなかった。テクスト分析の理論はこの「単位」を定義するための原理を与えていると見ることができる。また、アーカイブズへの操作、検索などの対象をこの観点に立って設計すべきである。

(2) は言語図式と言語実用の枠組みを、テクストアーカイブズを構築するために利用できないかという点である。テクストアーカイブズの構築には原テクストとテクストアーカイブズに対する二つの記号体系の関係を扱う必要があるが、二つの記号体系においては内容原意は共通である。また表現原意については、その範列に関する部分は構造を保持しなければならないため、多くの部分で共通点がある可能性が高い。問題は表現原意の連辞に関する部分を考察しなければならないという事である。

(3) は (2) で述べた言語図式と言語実用の枠組みを使うことで、非言語的な知識を言語体系から比較的独立に分析、記述可能である。内容原意を如何に表現するかについては各分野の検討が必要である。

6.2 次世代 SAT における検討項目

大蔵経プロジェクトでは、これまで全文テクストデータベースである「大正新脩大藏經テキストデータベース」[6](以降、SATDB と呼ぶ) を公開している。次世代の大蔵経データ

ベース (以降、新世代 SAT) では、「はじめに」でも述べたような人文学研究の新しい基盤としてのテクストアーカイブズを志向している。しかし、その前段階として、SATDB に索引、シソーラス、用語辞書等を付加することで、意味論的な検索を実現することも次世代 SAT のためには必要なことであろう[⋆1]。SATDB では全文テクストデータベースという枠組みを使用したがための問題点がある。これについては節 3.3 で検討した。

大蔵経のような地域や時代を異にする様々な経典を収蔵したような文献においては、更に以下のような問題を検討していかなければならない。

(1) 複数の共時的な言語体系の扱い
(2) 言語体系を形成するには至らない翻訳語のような単位の混入
(3) 非言語的な (人文学的) 知識の表現

項目 (1)(2) については言語素論の共示的分析[12] が参考になるであろう。(3) の非言語的な知識の表現については節 6.1 で議論した。以上、今後の課題としたい。

7. おわりに

昨今、デジタルアーカイブズの様々な成果が発表されているが、多くは研究実績を急ぐあまり、理論的な短絡を無視するような傾向にあるように思う。構築されたデータやシステムが将来的に無駄になることのないように、我々に今できることは、自分達がどのようなデータやシステムを作っているのかを良く理解できるような、着実でしっかりした基礎的な研究を展開することであろう。

本稿で議論した言語理論が言語に対する一つの視点である以上、その視点に基づくテクスト理解も一つの可能性に過ぎない。従って、そこから導き出されたテクストアーカイブズ構築の方法も一つの試みである。この立場でのテクストアーカイブズへの取り組みは、一つ一つその正当性を積み上げていき、有用なものへと仕上げていくことは我々の当面の課題であろう。

文献[10)8)] は文字の表現のために、その知識表現である素性の集合による Chaon モデルを提出している。また、文献[9)] でも示されているように、素性の集合というこの枠組みは一般の知識表現にも拡張できる。しかし、素性の集合をどのような原理によって決めるかということは、知識表現という面からも重要な課題である。テクスト分析の理論はそのための指針を与えることができると期待できる。具体的な展開については別稿とする。

⋆1 文献[3)] では、人文系データベースに対してその分野固有の意味論が必要であることを述べた。

謝辞 次世代人文学開発センターでお世話になっている下田正弘先生に感謝いたします。大蔵経プロジェクトでは永崎研宣さん、清水元広さんにお世話になっております。御三方には人文学が必要とする研究インフラの理想について多くの示唆を賜っています。秋山陽一郎さん、守岡知彦さん、山田崇仁さんには人文情報学の方向性について多くのご教示を頂いております。また、多くの方の教えと助力に感謝いたします。最後に妻留美と家族の常日頃の励ましに感謝します。

参 考 文 献

1) 井筒俊彦: 意味分節理論と空海, 「意味の深みへ」所収, 岩波書店, 1985.
2) 白須裕之: 歴史記述に対する概念分析の試み, 情報処理学会研究報告 2007-CH-74, 2007.
3) 白須裕之: 人文系データベースを構築するとはどういうことか？ , 漢字文献情報処理研究 第 9 号, 2008.
4) 白須裕之: 記号機能としてのアーカイブズ, 情報処理学会 人文科学とコンピュータシンポジウム「じんもんこん:-)2008」, 2008.
5) 白須裕之: 文字の指示概念に関する試論, 情報処理学会 人文科学とコンピュータシンポジウム「じんもんこん:-)2008」, 2008.
6) 大蔵經テキストデータベース研究会: 大正新脩大藏經テキストデータベース, http://21dzk.l.u-tokyo.ac.jp/SAT/, 2008.
7) 當山日出夫: 文字とアーカイブ, 情報処理学会研究報告, 2008-CH-79, 2008.
8) 守岡知彦: 文字オントロジーに基づく文字処理について, 情報処理学会研究報告, 2006-CH-72, 2006.
9) 守岡知彦: Concord: プロトタイプ方式のオブジェクト指向データベースの試み, Linux Conference 2006, 2006.
10) 守岡知彦, 師茂樹: 文字素性に基づく文字処理, 情報処理学会研究報告, 2004-CH-62, 2004.
11) F. Gadet: Saussure, Une science de la langue, Presses Universitaires de France, 1987. (邦訳 立川健二訳: ソシュール言語学入門, 新曜社, 1995.)
12) L. Hjelmslev: Prolegomena to a Theory of Language, translated by F.J. Whitfield, The University of Wisconsin Press, 1961. (原典 Omkring Sprogteoriens Grundlæggelse, Ejnar Munksgaard, Copenhagen, 1943) (原典よりの邦訳 竹内孝次訳: 言語理論の確立をめぐって, 岩波書店, 1985.)
13) G. Mounin: Introduction à la sémiologie, Minuit, 1970. (邦訳 福井芳男他訳: 記号学入門, 大修館書店, 1973.)
14) L.J. Prieto: Messages et signaux, Presses Universitaires de France, 1972. (邦訳 丸山圭三郎訳: 記号学とは何か — メッセージと信号, 白水社, 1974.)
15) F. de Saussure: ソシュール 一般言語学講義 — コンスタンタンのノート, 訳者 影浦峡, 田中久美子, 東京大学出版会, 2007.