翻訳過程の代数構造について

新　田　義　彦

要旨

1　はじめに

翻訳および機械翻訳については，膨大な研究報告が発表されているが，そのほとんどが翻訳における意味の保存と伝達に主軸をおくものである．本研究では，翻訳過程，つまり言語系Aから他の言語系Bへの翻訳の処理過程を少し抽象化して代数構造の変換として解釈すると，見えてくるもの（新知見）について論じる．この研究の動機と目的は，新しい機械翻訳の方向の策定，そして言語教育に翻訳過程を積極的に導入する効用の発掘にある．

これまで翻訳そして初等的言語教育（特に外国語教育）においては，語と語の対応（いわゆる対訳語の列挙と記憶奨励），語句の並び順（いわゆる構文）の学習に，第一の重点が置かれていたように見える．（数多く有る英語学習書では，語彙と文法という大分類の下でこのような知識を学習させようとしている．）一方またコミュニケーション英語という標語の下では，情景を詳細に叙述規定したあと，必然的に発生するであろう会話や質問応答を，一種の言語利用の常識的パターンとして，初学者に修得させようとしているように見える．

どちらのアプローチも言語教育あるいは外国語教育として正鵠を射ていると筆者は思う．本論文で提案しようとする，翻訳過程の代数的構造による理解は，前記二者のような直接的・実用的な教育効果を持つものではない．翻訳過程をより抽象度と透明度の高い代数構造として捉えることにより，見えてくるものをまず探りたい，というのが本研究の素朴な動機である．そして副次的目標として，抽象性ゆえに難解性が高いとされる圏論（Category Theory）や表示的意味論（Denotational Semantics）の具体的応用例を平明に提示したいという動機も地下に存在する．

2　言語表現の担う意味と統語構造

自分の思考過程や思考結果を他人に伝える手段として最も重要なものが言語表現であり，言語表現は典型的には「文」の集合である．文は単語列であり，単語は文字列として表現されるから，文も結局文字列として表記される．しかし，本論文では考察対象の最小単位として，「文字」ではなく「単語」を設定する．

結局，文は，複数個の単語を線状に並べて思考過程や思考結果，そして他者に伝えたい情報（＝メッセージ）を表現したものである，と考える．これらの表現対象を「意味情報」と呼ぶことにしよう．文が担う意味情報は相当に複雑精緻である．しかし本論文では，考察を加速するために，まず計算論的に厳密に定義された“意味概念を表現する記号（符号）系”が存在すると仮定する．さらにその“記号系”

の上に「文の意味」を写像できると仮定する．すると，その像（＝写像した先の値）は，精緻かつ複雑な立体的なグラフ構造を持つことになるであろう．あるいは多次元的なグラフ構造というべきかもしれない．いずれにせよ，この像（イメージ）の全体を正確に把握することは実際上困難である．文が担う意味情報には，多義性，比喩性，象徴性，含意性，など未解明な要素が数多く存在するからである．（cf.［Grice 1975］,［Saraki & Nitta 2008］,［Moon 1987］,［Marcus 1980］,［Nitta 1986 a , 1986 b ］）

このような理由により計算機による意味の取り扱いにおいては，文が持つ意味内容（＝情報内容）のうちの必要部分だけを形式化あるいは抽象化して処理せざるを得ない．

文の意味を形式化あるいは抽象化する方法の前提，文の構造を記号により抽象化する必要がある．抽象化した文構造は，統語記号と機能語の列として表現する．機能語は抽象化する以前の文の単語（文字列）を使って表現される場合が多いことを注意しておく．具体例を提示して考察を進める．

下記（1）および（2）のような表層文に対応する抽象化の例は，下記の（3）のようになる．

（1）社長が工場に自動組み立てロボットを入れる．
（2）主人が客にコーヒーを入れる．
（3）Ｎ１がＮ２にＮ３を入れる．

ただし，Ｎ１，Ｎ２，Ｎ３などは名詞または名詞句の存在を示す統語記号である．それらが（社長，工場，自動溶接機）のような名詞〔句〕集合，あるいは（給仕，私，コーヒー）のような名詞〔句〕集合を抽象化している．

日本語の「入れる」は，様々な意味をもつ．具体的には，（1）および（2）の日本語文に対応する英語文を考えてみるとよくわかる．Ｎ１，Ｎ２，Ｎ３などの統語記号はそのまま継承するが，適当に英語の名詞〔句〕集合を抽象化していると見なすことにする．

（4）Ｎ１ installs Ｎ３ in Ｎ２．
（5）Ｎ１ makes Ｎ３ for Ｎ２．

（4）や（5）のような差異（多義性，訳し分け）は，名詞［句］Ｎ１, Ｎ２, Ｎ３などの意味属性の差異に依存していると考えるのが妥当である．意味属性を導入した抽象化は，たとえば下記のようになる．

（5）Ｎ１〈＃３ 主体〉がＮ２〈#389 施設〉にＮ３〈#962 機械 , #2007 設備〉を入れる．
（6）Ｎ１〈＃４ 人〉がＮ２〈＃４ 人〉にＮ３〈#857 飲み物〉を入れる．

上記の（5）と（6）に出現した意味属性記号，つまり“＃３，＃４，#389，#962，#2007”のようなコードは，日本語語彙大系（岩波書店刊行の全５巻の辞書［Ikehara et al. 1997］）において定義されている意味類型コード番号を引用した．

上記した文の意味情報の抽象化表現は，比較的単純な例である．意味情報の差異が，文を構成する体言（≒名詞［句］）の意味属性の差異だけから説明できるからである．文の統語構造，特に用言（≒動詞［句］）支配構造に起因する意味の差異あるいは多義性は，相当に複雑である．具体例に即して説明する．

日本語の文は用言が１つの単文の場合は，統語構造は比較的単純であり，意味情報を考慮した抽象化や代数化は比較的簡明である．用言が２つ以上あってそれらが，連用形で接続している場合には慎重な意味情報解析が必要となる．日本語文に固有の連用形の接続，「～して　～する」の形式を中核に置いて議論を進める．

［Saraki and Nitta 2005］では，日本語文で典型的に出現する「・・・して～する」という連用形の接続形を「シテ型接続」と呼び，その適正な統語構造解析の重要性を議論している．佐良木昌氏によるシテ型分類の統語論と意味論の概要を以下に示す：

■　３つのカテゴリーによる「シテ型接続」の統語論的分類

（１）

(1.1) Ｖ１してＶ２　　verb conjunction only
(1.2) Ｖ１ようにしてＶ２　　comparing situation
(1.3) Ｖ１ようとしてＶ２　　intention
(1.4) Ｖ１させてＶ２　　causative verb
(1.5) Ｖ１られてＶ２　　passive verb
(1.6) Ｖ１しないでＶ２　　negation

（２）

(2.1) Ｖ１してＶ２する
(2.2) Ｖ１して VP ２する
(2.3) VP １してＶ２する
(2.4) VP １して VP ２する

（３）

(3.1) Ｖ（Ｐ）１してＶ（Ｐ）２
(3.2) Ｖ（Ｐ）１してＶ（Ｐ）２してＶ（Ｐ）３

註：　Ｖは動詞あるいは用言，VP は動詞句または用言節を表す．１や２などのインデクスは，動詞や用言の異なりを示す．

■　４つのカテゴリーによる「シテ型接続」の意味論的分類

A. Collateral condition（平行・付随する状態）
（１）Agent condition（動作主の状態）
(1.1) Posture change（姿勢の変化）
太郎は小首をかしげてしきりに考えていた．

Taro was thinking hard with his head on one side.

（1.2）Put on-off（持ち上げと下げ）

父はグレーの背広を着て出かけた.

Father went out in his gray suit.」

（1.3）Carrying（運搬）

ずっしり重いかばんを下げて兄が出張から帰ってきた.

My elder brother came home from his business trip carrying a heavily packed bag.

（2）Mental condition（心的状態）

（2.1）Inner mental action（内的な心理による行為）

彼はあわてて彼女を押しのけて行った.

He hurriedly pushed past her.

（2.2）Exposed mental condition（心理状態の露呈）

花子は口許に微笑を浮かべてわたしたちを迎えてくれた.

Hanako welcomed us with a smile on her lips.

（3）Agent's action（動作主の行為）

その犬は鼻をくんくんさせて食べ物を探し回った.

The dog sniffed around for food.

（4）Collateral condition（平行条件）

（4.1）Agent's condition（動作主の状態）

彼は先に立って歩いた.

He walked in front.

（4.2）Condition when main event（main predicate）occurs（中心的事件による条件）

義雄はよくテレビをつけっぱなしにしてうたた寝している.

Yosio often dozes off with the TV on.

（4.3）Conditions expressed by similar events or metaphor（類似事件や象徴的事件が原因）

太郎と花子は人目に立たないようにして会っていたものだ.

Taro and Hanako used to meet secretly.

B. Temporal condition

（1）Temporal circumstances（時間的状況）

しばらくして手に痛みを覚えた.

After a while, I felt a pain in my hand.

（2）Temporal succession（時間の継続）

母はスープの味見をして塩を加えた.

My mother tasted the soup, and then added salt.

（3）Temporal processing flow（時間的処理の継続）

その実業家は十分に足固めをして新しい事業に取りかかった.

That businessman made all necessary preparations before embarking on a new enterprise.

（4）Moment of action（瞬間的行為）
多くの難民が脱走しようとして撃たれた.
Many refugees have been shot while making a bolt for freedom.

C. Originating condition（元来の状態）
（1）Cause（reason of involuntary event occurrences）（原因，無意思の事件の発生が理由）
その箱を持ち上げようとして梅子は肩の筋を違えた.
Umeko strained her shoulder lifting the box.

（2）Reason（理由）
(2.1) Reason of agent's voluntary action（動作主の有意思行動が関与する理由）
太郎は欲が出て，失敗した.
Taro failed because he was too eager.
(2.2) Agent's judgment（動作主の判断）
危険が起こるかもしれないと考えて彼はあとに残った.
He stayed behind in view of possible danger.
(2.3) Bases of judgment（of implicit agent）（潜在動作主の判断）
安全性，経済性の両面から考えて，このストーブを買った.
I have bought this stove after considering both economy and safety.
(2.4) Reason why agent is affected by other voluntary actions（他者の有意思行動が関与する理由）
彼は信号無視をして重い罰を受けた.
He was severely punished for running a red light.

（3）Objective-oriented cause（目的が関与する原因）
(3.1) Subjective intention
父はこちらを振り向かせようとして空咳をした.
My farther gave a dry cough to make them turn and face this way.
(3.2) Presentation of objective
多くの候補者が市長の椅子をめざして選挙運動中です.
Many candidates are campaigning for the mayor ship.

（4）Methodological cause（方法や手段が関与する原因）
その国会議員は父を買収して何も言わせないようにした.
The congressman bribed my farther to say nothing.

（5） Condition（状態）
水を無くしては（水無しでは）どんな生物も生きていけない.
Without water, no living thing could survive.

D. Parallel（並列）

二人は芝生の上に仰向けになって寝ころんだ.

The two lay on the lawn face up.（註：英語の対応語句なし）

上記のような用言の並びは抽象化と代数化が難しいのであるが，用言接続を与える函数のパラメータにカテゴリ分類を反映させればなんとか対応できる．もう少し優雅な対応法は今後の課題である.

3　代数的に表現した等価変換

等価変換とは, 文の持つ意味情報を保存（維持）したままで進行する言語変換である．言語変換とは，言語表現の変換，つまり文の表現形式の変換である．文や文章の中に取り込められた「意味情報」を不変のまま変形・変換するにはどうすればよいか．変換（つまり翻訳や言い換え，書き換え）により，元の意味が変更されてしまう現象（誤変換）を最大限回避しなければならない．最大限という修飾語をつけた理由は，文や文章の表現形式を変更すれば，なにがしかの微妙な意味は変化せざるを得ないからである.

本論文で検討している「等価変換」は，実は古典哲学でも重きをなす概念であり，理論物理学における基礎概念でもある．有名な相対性理論をアインシュタインが考察する際の思考的基礎になっていたことも知られている．本章においては，工学や技術分野にける新発見や発想の基盤を哲学的に研究した市川亀久彌［Ichikawa 1963］の定式化をまず見てみよう．市川亀久彌は，“独創性”の背景にある“等価的類推思考”を下記のような記号式により定式化した.

（1）　C（ A/α　=/ε　B/β ）

ただし，式（14）は“条件Cの下で，＝の左辺と右辺の間で意味的等価関係が成立している”というように解釈する．ここで“意味的等価関係”とは，条件Cの立て方により，“類推”であったり“翻訳”であったり“言い換え”であったりする．“＝”の右辺と左辺の間には，“等価性”という関係があるだけであり，“方向性”や“非対称性”はない.

もう少し詳細に記号の意味を検討する．A/αは言語体系αの上で記述された表現であり，B/βは言語体系βの上で記述された表現である．換言すれば，言語体系αやβの上で思考した際の,“対象となった事象･事物”や“結果（あるいは過程）としてのイメージや想念”, を表す記号がA/αやB/βである．ギリシャ文字のα，β，εは，思考や言語表現の基底にある系（システム）を表す．特に“ε”は“＝”の左辺と右辺に共通する思考系／言語表現系を表す．たとえば，αが感情や感覚系であり，βが音楽系であり，εが音楽家一般の思考系を表現するものとすれば，式（1）は音楽家が自身の感情や感覚を等価変換して楽譜に変換するという行為，つまり“作曲”という知的行為を定式化したこととなる．同様に，βを図形・色彩・構図などの画像系とすれば，画家の“描画”という芸術的創造行為を定式化したことになる.

このように市川亀久彌氏が提唱する“等価変換”は，創造，工夫，発明，発見，など，非常に広範囲の知的所為の基底にある哲学的構造を扱うものであるが，本論文では，定式化（1）の作用域を言語現象のみに限定して考察することとする．具体的には，下記のような枠組みａ），ｂ），ｃ）の下で考察を進める．

ａ）条件Ｃは，文の意味のうち代数的に抽象化できるものを扱う．代数的に抽象化できるという意味は，微妙な含意や示唆（ほのめかし）の類は扱わず，論理的に明解な意味だけを扱うという意味である．

ｂ）言語表現ＡおよびＢは，自然言語の文とする．文の構造は単文・複文・重文などであるが，特に本研究では複文に関心がある．単文の意味の抽象化にはある程度の見通しが，日本語語彙大系の開発により得られている［Ikehara et al. 1997］からである．

ｃ）表現の基底にある表現系 α や β は，日本語や英語，ドイツ語，フランス語，といった自然言語の区別を表すこととする．一般に $\alpha \neq \beta$ の場合には，等価変換は言語 α から β への翻訳を表すこととなる．また α と β とが同一の言語系に属していて，単に語彙集合や統語解析規則の相違などがある場合には，言い換え・要約を表すことになる．

ｄ）等価変換においては，双方向性や対称性を仮定せず，左辺から右辺への一方向への変換を考える．双方向性を仮定すると，翻訳や言い換えの場合に，双方向翻訳や相互（可逆）変換を実現しなければならなくなるが，これは現状の自然言語処理技術の水準に照らして困難である．ゆえに，式（1）における等号“＝”の代わりに，等価変形“⇒”を使う．

ｅ）等価変形“⇒”における基底表現系は，述語論理（predicate logic）とする．つまり等価変形“⇒”において，不変に保存される意味情報は，述語論理で表現可能なものに限定することにする．換言すれば，本論文で扱う“意味論”は“論理学的意味（logical semantics）”であり，語用論的意味（pragmatic meaning）や文脈的意味（contextual meaning）は扱わぬ（現状では断念する）こととする．語用論的意味とは，例えば，「君，財布持ってる？」という疑問文の〔実際的〕意味が「食事の代金は君が払ってくれたまえ」という要求であるというような状況解釈に立ち入る意味論である［Nitta, Yoshihiko 1988］．「財布は持ってないけど名刺入れならあるよ」と答えてすましていてはいけない，という解釈も語用論的解釈，談話理論的解釈から導き出されるが，本論文が取り上げる等価変換による意味の保存の範疇からは外しておくこととする．

結局，本研究における代数的抽象化を指向する，等価変換の定式化は（1）式の代わりに下記（2）式として定式化できる．

（2）　$C(A/\alpha) \Rightarrow/\varepsilon \; C(B/\beta)$

上式における記号の読み方を念のために再論する．“$C(A/\alpha)$”は言語 α で記述されている文（一般的には言語表現）Ａを，代数的抽象化Ｃによって記述（表現）したものである．記号“$C(B/$

β）”の解釈も同様である．記号“⇒/ε”の解釈は，述語論理（ε）で表現（抽出）される意味を保存しつつ，左辺の言語表現を右辺の言語表現に変換（変形）することである．

言語α（たとえば日本語）から言語β（たとえば英語）への機械翻訳の場合を具体例として，等価変形の手順を以下に示す．

第1段）言語αで記述されている文A/αを，“言語αの意味類型知識ベースKB α”と照合させつつ，代数的抽象化C（A/α）に変換する．

第2段）代数的抽象化C（A/α）を，述語論理系εの形式A/α/εに変換する．この変換処理においては述語論理規則系と共に，表層文A/αおよび意味辞書SDαなどを参照する．

第3段）述語論理形式A/α/εを参照しつつ，“言語パターン変換規則PTα β”にしたがって，代数的抽象化C（A/α）を代数的抽象化C（X/β）に変換する．ここで，X/βは，言語βによる記述文を変数化（汎化）した表現を表しており，これに言語αと言語βに共通する意味類型コード（註：厳密に言うと回りくどくなるが，代数的抽象化Cが定めるコード）を付与した表現が，C（X/β）である．

第4段）代数的抽象化C（X/β）を，“言語βの意味類型知識ベースKB β”と照合させつつ，代数的抽象化C（B/β）に変換する．多義性の処理は，前述のように補助的な意味処理や辞書参照により適宜行なうものとする（註：本論文では詳細な議論は略す）．

第5段）代数的抽象化C（B/β）から，言語βによる表層文B/βを生成する．生成においては，“言語αと言語βの語彙的対照辞書TDα β”などを援用する（註：対象言語生成技術の詳細は略す）．

元の入力文A/αを代数的抽象化する際に，多義性やあいまい性が発生した場合の処理は一般に面倒であるが，1つの簡便な方法は，適当な優先度（順位）を付与して複数個の代数的抽象化を対応させることである．絞り込みは，後段で補足情報と補助手段により行なう．

上述の変換（変形）過程を“⇒”により記号化すると下記（3）式のようになる：

（3）　A/α
　　⇒/KB α, SDα
　C（ A/α ）
　　⇒/ε, SDα
　A/α/ε
　　⇒/PTα β, C（ A/α ）
　C（ X/β ）
　　⇒/KBβ, SDβ
　C（ B/β ）

⇒ / SD β, TD α β
B / β

ただし，“⇒ / x ”は左辺から右辺への変換において，“知識 x の制御を受けること”，もしくは“知識 x を参照・照合しながら変換すること”を意味する記号である．また KB α，KB β はそれぞれ，“言語 α の意味類型知識ベース”および“言語 β の意味類型知識ベース”である．SD α, SD β はそれぞれ“言語 α の語彙的意味辞書”および“言語 β の語彙的意味辞書”である．また TD α β は，“言語 α と言語 β の語彙対照辞書”である． ε はある種の“述語論理系”であるが，その詳細は本報では述べない．PT α β は，“言語 α から言語 β への形式変換（トランスファー）知識”である．その他の記号，A / α，B / β, C （ A / α ）, C （ B / β ）などの意味は従前の通りである．

4　意味類型知識ベース

“意味類型”とは，人間の認識・思考・伝達・などの知的活動の結果を言語の形式で表現したものを，計算論的な観点から抽象化･汎化したものである．たとえば，「肯定」,「否定」,「比較」,「評価」,「因果」,「理由」，などは人間の“判断”という知的行為を“言語化”もしくは“概念化”したもののカテゴリー（＝下位範疇化，下位分類）である．このような“判断に関わる概念”の表現形式は，言語に依存して多様に変化する．本研究のアプローチではコーパス言語学の方法論に準拠して，各々の言語（前述の記号では α や β）ごとに，コーパスベースの学習・抽出を網羅的に行なう．コーパスから，人間の知的活動の結果としての概念（あるいは想念，イメージ）を表現するパターン（＝意味類型パターン）を収集し，これを整理統合して知識データベースを構築する．前章の記号を使うと，“言語 α の意味類型知識ベース KB α ”および“言語 β の意味類型知識ベース KB β ”などを，コーパスベースの学習抽出処理により構築することが本研究の基礎作業の一部である．以下に，意味類型知識ベースの具体例を示す．

（1）日本語の意味類型パターン：「２つの行為が“相反”していると判断するパターン」
・Ｎ１はＶ１したが，Ｖ２しなかった．
・Ｎ１はＶ１したけれども，Ｖ２しなかった．
・Ｎ１はＶ１したにもかかわらず，Ｖ２しなかった．

（2）英語の意味類型パターン：「２つの行為が“相反”していると断定するパターン」
・Ｎ１ Ｖ１ accidentally but it was NP ２．
・Ｎ１, which Ｖ １ by accident, was NP ２．
・It was by accident that Ｎ １ Ｖ１, which was though NP ２．

（3）日本語の意味類型パターン：「２つの行為が“同等”であると判断するパターン」
・Ｎ１はＮ２をＶ１するが，それは，Ｎ３がＮ４をＶ２するのと同じである．
・Ｎ１はＮ２をＶ１するのと同様に，Ｎ３はＮ４をＶ２する．
・Ｖ１を〔Adverbial － Modifier のように〕することは，Ｎ１と同じである．

意味類型パターンは，コーパスデータから抽出・収集あるいは学習・習得して決定する．意味類型パターンの決定における最大の課題は，“汎化の程度”の決定である．つまり，多様な表現自由度を持つ表層単語列の，どの部分をどのように変数化するのか，およびどの部分をカテゴリコードに変換するのか決定すること，あるいはその選定方法を決定することが重要課題である．

言うまでもなく，カテゴリコード系つまり意味類型コード系の仕様決定も課題であるが，本研究では当面，“日本語語彙大系”の意味分類コードを準用することとする．意味類型パターンを抽出・収集する際の“汎化”の問題は今後の課題とする．

5　変換知識ベース

前章で例示したように，日本語意味類型パターンを英語意味類型パターンに変換する〔あるいはその逆向きの変換をする〕際に，参照・照合（パターンマッチング）するための対応規則を集積したデータベースが，「パターン変換知識ベース」である．結局その本質は，日本語（α）と英語（β）における２つの意味類型パターンの対照表示データベースである．しかし，両者の対応は１対１にはならず，一般にm対nの関係となるから，αからβへの言語変換計算はそれほど単純ではない．ただし，本論文では対応変換計算の詳細には言及しない．

以下に，パターン変換知識ベースが与える“対応パターン”の例を示す．ただし，文法コード，統語カテゴリコード，意味類型コードの表示は省略している．これらをすべて表示すると，相当に重装備な素性束のシーケンスとなる．下記の（1），（2），（3）はそれぞれ，前章の（1），（2），（3）のパターンに対応する相手言語パターンである．

（1）“相反”概念を表現する英語意味類型パターン：
・N１ V１, but it did not V ２.
・N１ V１, however it did not V ２.
・Although N １ V１, it did not V ２.
・N１ did not V ２ despite V １ - ing.

（2）“相反”概念を表現する日本語意味類型パターン：
・N１は偶然V１したがNP １だ.
・偶々N１はV１したが，実はNP １であるのだ.
・偶然にV１したN１は，本当はNP １だ.

（3）“同等”概念を表現する英語意味類型パターン：
・N１ V１ N２, similarly N ３ V２ N４.
・N１ V１ N２, in the same way as N ３ V２ N４.
・V１ - ing〔Adverbial - Modifier〕is/was tantamount to N １.
（註：　末尾の英語意味類型パターンは，V１-ing と N１が論理的意味では（厳密には）同等ではないときに用いる修辞的な英語表現である.）

6　翻訳過程の統計的解釈

前節まででは，翻訳過程，特に機械翻訳過程を異言語間における意味的等価変換としてみて定式化を進めてきた．さらに抽象化を進めて，言語を単なる記号あるいは信号の情報列とみて，その間の統計的最適化処理とみて定式化してみよう．この考え方は，統計的機械翻訳と言われる最近の機械翻訳方式を直裁に理解するためにも有効である．

意味的等価変換と統計学的解釈（あるいは統計的変換）との関係について補足する．

以前にみた代数的な表現は，

（1）　C（ A/α　=/ε　B/β ）

あるいは

（2）　C（ A/α ）　⇒/ε　C（ B/β ）

であった．

ここで，代数数式の構成要素の意味は下記のようになっている．

C：　条件，あるいは　代数的抽象化

A / α：　言語αで表現されている文，αはたとえば日本語

A / β：　言語βで表現されている文，βはたとえば英語

C（ A / α ）：　言語αで表現されている文を　代数的に抽象化したもの

C（ B / β ）：　言語βで表現されている文を　代数的に抽象化したもの

=/ ε：　代数的言語系（たとえば述語論理）の範囲で等価性が成立するという意味

⇒/ ε：　代数的言語系（たとえば述語論理）の範囲で等価性を維持しながら，左辺を右辺に変換（写像）する

説明の簡略化のため，条件あるいは代数的抽象化を表す記号Cを省略する．

（1）’　A/α　=/ε　B/β

あるいは

（2）’　A/α　⇒/ε　B/β

そして，=/εや⇒/εを，統計的に成立する等式あるいは等価変換とみなす．つまりεは，統計的尤度を表すとみる．実際上計算処理としては，統計的尤度の計算だけでは，言語表現の変換はできないので，統計的探索・検索処理により代替する．

様々な言語（たとえばαやβなど）で記述された文が，世界中に（特にインターネット上に）存在する．そのような“文の巨大集合”の中を捜せば，必ずや任意に与えられた文A/αに対して，これと意味的等価な文B/βを検出できるはずである．

このように楽観的に考えて，現代の超高性能な計算機の力にものを言わせて，B/βを統計的探索する．統計的探索においては，論理的に厳密な等価は求めない，適当に定めた制約式の上で意味的等価の尤もらしさが最大値になればよしとする．

A/α，つまり言語α（日本語）で記述された文を，今一度簡略表現して　j　と記すことにする．

A/β，つまり言語β（英語）で記述された文を，今一度簡略表現して　e　と記すことにする．

このように記号を簡略化すると，統計的［近似］等価変換は，

$$\hat{e} = \mathrm{argmax}_e P(e|j)$$

という代数式で表現される．

以下では，この代数式を中核に据えて，機械翻訳のメカニズムについて議論を展開する．

■　日英機械翻訳過程は，下記の式により定義できる．ただし j は入力日本語文，$\hat{e}$ はこの入力文に対応する英語訳文である．e は翻訳の途中過程で考慮される，英語訳文の候補である．

P（e| j）はこの訳文候補 e の適訳性（ j の翻訳としての相応しさ）の程度を定量化したスコアである．確率値として定量する場合は，0から1の範囲の正の実数値となる．しかし値は確率値に固執する必要はない．種々の候補訳 e の間に序列が付けばよい．

argmax_e は，“最大スコアを取る e ” を選択する函数である．選択された e が，$\hat{e}$ として函数より出力される．

一般に，j を入力される起点言語の文，そして e あるいは $\hat{e}$ を出力される対象言語の文，と解釈すれば，下式は2言語間の機械翻訳過程を定式化したものと解釈できる．

$$\begin{aligned}\hat{e} &= \mathrm{argmax}_e P(e|j)\\ &= \mathrm{argmax}_e P(e)P(j|e)/P(j)\\ &= \mathrm{argmax}_e P(e)P(j|e)\end{aligned}$$

■　最新の統計的翻訳 SBMT も，古典的文法ベース（トランスファー型）翻訳 GBMT も，MT としての機能の本質は同じである．上の定式化により統一的に，機械翻訳メカニズムが説明できる．

■　上記の式を統計的信号処理の式として解釈してみよう．

本来の英語文 e が通信路の雑音や暗号化処理により，日本語文 j に化けて伝わった．この暗号 j に処理（翻訳）を施して，本来の英語文 e に復元する．この処理では文字列を連続信号列のように扱っている．通信理論・暗号理論のアナロジーにより，翻訳メカニズムを「雑音交じり信号から本来のオリジナル信号を復元する」メカニズムとして定式化している．一番尤もらしい（雑音混入の少ない）e が $\hat{e}$ である．　このメカニズムを，暗号解読や雑音信号解析と同様に，統計的に処理する方式が統計ベース機械翻訳（SBMT）［Brown, P. F. et al.（1993）］［Koehn, P（2010）］ である．

■　$\mathrm{argmax}_e P(e|j)$ により，直接的に入力日本語文 j から　出力英語文 e を変換出力しない理由は，次のように説明できる．この処理では，雑音の混入度が最小の e を選択出力するので，単語対応のみを考慮した“ぎごちない訳文 e ”のが出力される可能性が大きい．訳文 e のぎこちなさ（英語文としての不自然さ）を最小限度に抑制するために，英語文法モデル P（e）というフィルターを通すのである．この処理の付加により出力訳文 $\hat{e}$ の英語らしさが高まる．

■　もしも日本語文と英語文の自然的対応（翻訳）の対のすべてを，機械翻訳システムが，知識ベースとして所蔵していれば，勿論，英語文法フィルター（英語モデル）P(e) は，不要である．機械

翻訳システムは，言語変換をするのでなく，単なる対情報検索システムとなる．このときの機械翻訳システムの振る舞いは，「哲学者サールの中国語の部屋」のようになる．

つまり閉鎖した「英語の部屋」を作り，「日本語しか理解できない人間」と「“意味を共有する日本語文と英語文のペア”のすべてを収録した対訳リスト」を部屋に入れておく．この部屋の住人は，外から差し入れられた「日本語文」をみて対訳リストを検索して該当するペアの英語文を紙片に書き込み，ドアの外に返す．ドアの外の翻訳依頼人は，部屋の中に完璧な翻訳者（あるいは機械翻訳システム）がいて適切な日英翻訳をしてくれたと感心する．

■　しかしながら，３－グラムであっても可能な対訳アラインメントの数は　指数爆発に地近づくのであるから，翻訳を「サールの中国語の部屋方式」で実行することは不可能である．ここで「３－グラムの対訳アラインメント」とは，単語３つ分だけに注意を向けて対訳のペアを作った，翻訳用知識ベースのことである．サール流の英語の部屋を作るためには，任意の n（註：実際上は十分に大きな n でよい）に対して，すべての日英対訳 n グラム・アラインメントを作らねばならぬが，これは爆発的な数量のアラインメントになるから実現不可能である．

■　翻訳過程の数学的表現を，　統計的翻訳 SBMT として解釈するやり方を，もう少し厳密に再論すると下記のようになる．

$$\hat{e} = \mathrm{argmax}_e P(e|j)$$
$$= \mathrm{argmax}_e P(e)P(j|e)/P(j)$$

入力信号は　変動しないので，max をとる処理とは無縁．よって消去できるので，

$$= \mathrm{argmax}_e P(e)P(j|e)$$

ここで

$P(e)$：　言語モデル英語のモデル　英語コーパスから統計処理（sta と略）により自動生成

$P(j|e)$：　翻訳モデル．対訳コーパスから sta により自動生成

$\mathrm{argmax}_e P(e)P(j|e)$：　デコーダ．翻訳の実質を担うエンジン．高速アルゴリズムが盛んに研究開発されている．

■　文法的翻訳 GBMT，人間（≒言語の初等的学習者）の翻訳 HT としての解釈は下記のようになる．

$$\hat{e} = \mathrm{argmax}_e P(e|j)$$
$$= \mathrm{argmax}_e P(e)P(j|e)/P(j)$$

$P(e)$：英語の文法モデル　人手構築

$P(j)$：日本語の文法モデル　人手構築

$P(j|e)$：日本語と英語の間の変換規則（翻訳文法）人手構築

argmax_e　翻訳アルゴリズム，e を探索して $\hat{e}$ を選択するのではなく，$P(e)P(j|e)$，$P(j)$ を参照しつつ，$\hat{e}$ を生成する（訳文としての $\hat{e}$ を構成する）

$P(e)P(j|e)$，$P(j)$ は，すべて対訳コーパス，単一言語コーパス，言語学的知見などから，人間が

構築する.

■P(e)P(j| e)，P(j）の構築をコーパスから半自動的に行う方式が，EBMT（Example Base MT）である.

■言語の特異性（ideosyncracy）（Yoshihiko Nitta（1986a））に注目して，P(j| e）を　パターン対として構築する方式が　PBMT（Pattern Base MT）である．パターン対は，人手構築あるいは半自動構築する.

■　RBMT あるいは GBMT の弱点：　RBMT あるいは GBMT の弱点は高品質の変換文法を頭脳作業で作る難しさ，特に　複雑な特殊な文型と比較的単純な文型の両立のむずかしさ，を持つことである.

■　SBMT（統計ベース機械翻訳）の弱点：　SBMT の弱点は，高品質の対訳コーパス（バイリンガル，トリリンガル，マルチチリンガルーコーパス）の確保がむずかしいいことである．3-gram レベルの対訳学習や変換文法学習までは　なんとか実行できるが，一般の n-gram（n ＞= 4）については，手本にすべき学習用コーパスがスパース（過疎）となり学習困難に陥ることである.

7　翻訳過程の代数的解釈

これまで，翻訳過程を意味の保存変換としてみる定式化，および信号的な対応の最大尤度検出としてみる定式化についてそのメカニズムを概観してきた．これらの概観を踏まえ，さらに翻訳過程全体を，ある種の代数的写像として捉える観点を提案する.

翻訳過程で授受変換される種々の知識ベースを対象（Object）としてまず定式化，次に対象間の写像（射　Map）として翻訳の部分的処理を定式化する．この考え方は，カテゴリー理論［Mac Lane1997］［Eilenberg & Mac Lane 1945］［Freyd 1964］［Yoneda1954, 1960］およびその計算機科学への応用およびソフトウェア工学の精密化への応用を示唆する様々な提言〔Goguen1989］［Esterbrook 19999］の影響を受けている．ただしこれらの先人達の著述では，翻訳過程に特化した定式化には言及していない．一般に対象と射を，圏の間の写像つまり関手（Functor）として捉え，自然性や同型性を考察すべきことを示唆する段階で言述をやめている.

種々の代数的概念を翻訳過程に適用するアイデアは，未だ具体化検討の余地があるが，本研究の創案である.

言語 a を使う世界（圏）を α とする．世界 α で語られる文あるいはメッセージをＡとする．一方また別の言語 b を使う世界（圏）を β とする．世界 β の中に，メッセージをＡと情報的に等価（同内容）の「言語 b で語られる文あるいはメッセージをＢ」が存在すると仮定する.

（注意：　この仮定が妥当か否か，現実的に意味的等価な対訳表現がいつも存在するか否か，という議論はしばし棚上げしておく.）

そして，次のような代数的関係をトップダウンに導入する．

$$\alpha\ (\ T\ (A),\ B\ )\ =\sim\ \beta\ (A, T^{*}(B))$$

ここで　Tは世界αにおける　言語 a から言語 b への翻訳（Translation）とする．

また，T * 世界βにおける　言語 b から言語 a への翻訳（Translation）とする．

翻訳Tと翻訳 T*　は随伴関係にあると呼ぶことにしよう．

T*は　必ずしもTの逆写像［特にこの場合は，逆翻訳］T^{-1}ではないことを注意する．

容易に理解されるように，２つの翻訳過程，α(T(A), B)　とβ(A, T*（ B))との間には，強い関係がある．この関係をも随伴関係（Adjoin Relationship）と呼ぶことにする．

この随伴関係はある種の同型関係と見てもよいことがある．つまり同型関係にある翻訳は，ある意味同類項とみて，時に同一視することが許されることもあるという意味である．

ここで導入した随伴関係を，拡大敷衍して翻訳過程全般を見通しよく抽象化・般化することは，今後の課題である．

上記でみた　文A（これは言語 a で記述されている）と文B（これは言語 b で記述されている）とは，自然翻訳の関係にあると言える．つまりはじめから，２つの異なる言語圏αと言語圏βとに，パラレルに存在する〔自然な〕文であり，たまたま文Aの意味メッセージと文 b の意味メッセージが，等価あるいは同値になっているのである．しかし一般にこのような意味的等価文が，２つ（あるいはそれ以上の）異言語圏に存在すると期待することはできない．実際，共存しない事物を述べる文をとってくれば，自然翻訳は存在しないことが確認できる．

そこで上記の随伴関係あるいは同型関係を，少し緩くすることを考えよう．

$$\alpha(T\ (A), B'\ )\ =\sim\ \beta(A', T^{*}(B))$$

ここで，B’は，文Aの等価物（自然翻訳）Bのように，ピタリと対応する翻訳ではないが，原文Aにほぼ対応する翻訳文（受容できる訳文）というようなものである．ダッシュ（’）は，そのような受容可能近似というような意味合いである．　同様にA’は，Bの自然翻訳Aのようにピタリと意味的等価性が保障できる訳文ではないが，ほぼ等価な（Bの）翻訳文という意味である．

このように厳密な意味的等価性，自然翻訳性を断念して近似翻訳で我慢することにした場合，そのズレの調整・保障は，個々の語句や補助語が担う義務を負う．この機序を代数的に，写像として定式化すると種々の興味深い知見が得られる．得られた知見を，具体例と共に列挙することは次の報告に回す．

翻訳対象の文が，質問文，特に情報検索（探索要求）文である場合には，ラムダ式を経由する表示的意味論（Denotational Semantics）が有効である．すべての質問文あるいは検索要求文は，ある特定の対象空間への写像として表現できる．この写像はラムダ式により厳密に規定できる［Main & Benson 1983］．逆にラムダ式から　任意の言語表現に変換（翻訳）することも比較的容易である．

１つだけラムダ式への翻訳例をあげる．

What owns no black stone

black stone:

λ u. λ z(stone)uz ∧ (black)uz

no black stone:

λ x. λ u. ¬∃ z.(stone)uz ∧(black)uz ∧ xuz

own no black stone:

λ u. λ z'. ¬∃ z.(stone)uz ∧(black)uz ∧(own)uz'z

what own no black stone:

λ y. y(λ u. λ z'. ¬∃ z.(stone)uz ∧(black)uz ∧(own)uz'z

White に対応する意味対象が答えとなる.

つまり　世界‖ g ‖ ψにおける　‖ λ x. λ u. xu(White) ‖ ψが応答になる.

ただし　‖ ¬∃ z.(stone) gz ∧(black)gz ∧(own)g(White)z ‖ ψ

8　おわりに

翻訳過程を代数学的に解釈することにより，新規に得られたこと，今まで見えなかったことが顕在化したことは何か，要約列挙してみよう.

1）異言語間における言語表現が，翻訳写像の対象として掌握できるようになること.

2）個々の原文，翻訳文という具体的言語データ（コーパス）を，捨象して（一時的に忘却して），翻訳という言語行為を純粋な代数的写像として掌握できること.

3）上記のような超越的な翻訳観から，新しい異言語コミュニケーションの可能性，新しい言語教育・外国語教育へのヒントが得られること.

4）言語解析や文法適用などの人工知能的な思考を除去してしまう「統計的翻訳」の手法と古典的な「文法ベース翻訳」手法の間の連携・共存・協調のヒントが得られること.

カテゴリ論の種々の補助定理（End を扱う米田のレンマ，完全系列を扱う蛇のレンマ，など）を，より積極的に導入・援用した，代数的翻訳理論の開拓，そして新しい外国語教育の具体的提案の策定，表示的意味論（Denotational Semantics）を基礎におく，機械翻訳プログラムの半自動生成，などを今後の課題としたい.

9　参考文献

Bentivogli, L. and Pianta, E.（2005）Exploiting Parallel Texts in the Creation of Multilingual Semantically Annotated Resources: the MultiSemiCor Corpus. *Natural Language Engineering 11*(3):247-261

Brown, P. F. et al.（1993）The Mathematics of Statistical Machine Translation: Parameter Estimation, *Computational Linguistics* 19（2）:263-311

Eilenberg, S., Mac Lane, S. General Theory of Natural Equivalences, Trans. Am. Math. Soc. No. 58（1945）pp.231-294

Esterbrook, Steve, An Introduction to Catedory Theory for Software Engineers, Dept. of Computer Science, Univ. of

Toronto, http://www.cs.toronto.edu/~sme/presentations/cat101.pdf（1999）

Freyd, P., Abelian Categories: An Introduction to the Theory of Functors, New York（1964）Harper and Row

Goguen, Joseph A., A Categorical Manifesto, Technical Monograph PRG-72, Oxford University Computing Laboratory, Programming Research Group（March 1989）

Grice, H. P.（1975）Logic and Conversation, in: P. Cole and J. L. Morgan（eds.）*Syntax and Semantics,* vol.3, New York, Academic Press, pp.41-58

Hornby, A. S.（1978）*Guide to Patterns and Usage in English.* Oxford University Press（Japanese Translation: Kenzo Ito.1978. *An English Model, Usage.* Oxford University Publication Office）

Ikehara, S. et al.（1997）*Japanese Lexical Compendium,* Iwanami Shoten, Tokyo, Japan

Ikehara, S. et al.（2002） 池原悟，佐良木昌，宮崎正弘，池田尚志，新田良彦，白井諭，柴田勝正，類推思考の原理に基づく言語の意味的等価変換方式，人工知能学会論文誌，（2002）

Kinyon, A.（2001）A Language-Independent Shallow-Parser Compiler, *Proc. 39th ACL Ann. Meeting（European Chapter）*:322-329

Koehn, P（2010）*Statistical Machine Translation,* Cambridge University Press, 433p

Lehrberger, J. J.（1978）*Automatic Translation and the Concept of Sublanguage,* Groupe de Recherché en Tradution Automatique（TAUM）, Universite de Montreal, Canada

Lehmann, W. P.（1980）*The METAL System,* Linguistic Research Center, University of Texas, Texas, USA

Mac Lane, Saunders, Categories for Working Mathematician, Springer-Verlag, New York LLC.（March 27, 1997）
（訳本：S.マクレーン　著，　三好博之，高木　理　訳，圏論の基礎，丸善出版（2012-3）

Maimon, O and Rokach L（eds.）（2010）*Data Mining and Knowledge Discovery Handbook, Second Edition,* Springer Verlarg 1285p

Main, Michael G. and Benson, Dravid B.（1983）, Denotational Semantics for “Natural” Language Question-Answering Programs, American Journal of Computational Lingustics, vol.9 no.1（1983）pp.11-21

Marcus, M. P.（1980）*A Theory of Syntactic Recognition for Natural Language.* The MIT Press

Mihalcea, R. and Simard, M.（2005）Parallel Texts. *Natural Language Engineering* 11（3）:239-246

Moon, R.（1987）The Analysis of Meaning, in:（Sinclair（ed.）, 1987）Chapter 4:86-103

Munday, J.（2008）*Introducing Translation Studies,* Taylor & Francis Group : 訳本　ジェレミー・マンデイ著，鳥飼玖美子（監訳）翻訳学入門，　みすず書房（2009）

Nakamura, Y.（1983）*How far can we go in translation*?（翻訳はどこまで来たか？） Japan Times, Tokyo, Japan

Nagata, M.（2003） Natural Language Processing by Statistic Model, in: Amari, S. et al（eds.）, Statistics for Language and Psychology, Chapter 2 : 59-128（甘利俊一　他　編　統計科学のフロンティア　第10巻　言語と心理の統計　第2章　永田昌明，確率言語モデルによる自然言語処理），岩波書店（2003）

Nida, E. and C. Taber（1969）*The Theory and Practice of Translation,* Leiden: E. J. Brill（1969）

Nierenburg S. et al.（eds.）（2003）*Readings in Machine Translation,* The MIT Press

Nitta, Y. et al.（1982）A Heuristic Approach to English-into-Japanese Machine Translation. in: J. Horecky（ed.）*Proc. COLING 82 (at Prague) (=Proceedings of the 9th International Conference on Computational Linguistics)*, North Holland Publishing Company: 283-288

Nitta, Y. et al.（1984）A Proper Treatment of Syntax and Semantics in Machine Translation, *Proc. of COLING 84 (at Stanford)) (=Proceedings of the 10th International Conference on Computational Linguistics)*, Association for Computational Linguistics: 159-166

Nitta, Yoshihiko, et al.（1982）A Heuristic Approach to English-into-Japanese Machine Translation, in: Horecky, J.（ed.）: Proc. COLING 82（at Prague）（=Proceedings of the 9th International Conference on Computational Linguistics）, North Holland Publishing Company（1982）pp.283-288

Nitta, Yoshihiko, et al.（1984）: A Proper Treatment of Syntax and Semantics in Machine Translation, Proc. of COLING 84（at Stanford)）（=Proceedings of the 10th International Conference on Computational Linguistics）, Association for Computational Linguistics（1984）pp.159-166

Nitta, Yoshihiko（1986a）, Idiosyncratic Gap: A Tough Problem to Machine Translation, Proc. Comp. Linguistics, COLING'86 ACL（Assoc. Comp. Ling.）（1986）

Nitta, Yoshihiko（1986b）, Problem of Machine Translation Systems: Effect of Cultural Differences on Sentence Structure, Future Generation Computer System, Vol. 2, No. 2, North-Holland（1986）

Nitta, Yoshihiko（1986c）, Machine Translation: A Problem of Understanding, Japan Computer Quarterly, No. 64, JIPDEC （1986）

Nitta, Yoshihiko（1987）, Natural Language Understanding Viewed as Human Interface, Proc. Human Interface, SICE（Society for Information and Control Engineering）（1987）

Nitta, Yoshihiko（1988）, 新田義彦, 自然言語理解の基礎：意味論と語用論, 情報処理 , Vol. 30, No. 10, 情報処理学会 pp.1182-（1988, Oct.）

Nitta, Y.（1993）Referential Structure: A Mechanism for Giving Word-Definitions in Ordinary Lexicons. in: *Language, Information and Computation,* LSK（Linguistic Society of Korea）

Nitta, Y.（2002a）A Study of Semantic Typology Patterns and their Transformations, *Economic Review of Nihon University,* 71(4) Nihon University, Tokyo:131-155

Nitta, Y.（2002b）Problems of Machine Translation: From a Viewpoint of Logical Semantics, *Economic Review of Nihon University*. 72(2) Nihon University, Tokyo: 23-42

Nitta, Y.（2002c）A Study of Descriptive Language for Sentence Patterns, *Economic Review of Nihon University*. 72(3) Nihon University, Tokyo: 35-59

Pim, A.（2010）*Exploring Translation Theories,* Routledge, Taylor & Francis Group : 訳本　アンソニー・ピム著, 武田珂代子（訳）翻訳理論の探求,　みすず書房（2010）

Saraki, M. and Nitta, Y.（2005）,The Semantic Classification of Verb Conjunction in the "Shite" Form（日本語文のシテ型

用言接続が英語で連体修飾に転化する現象）, *Proceedings of Spring IECEI Conference,* IECEI Japan

Saraki, M. ed. and Nitta, Y.（2008）, *Regular Expression and Text Minig*（正規表現とテキスト・マイニング）, *Second Printing,* 明石書店, 312p

Saraki, M. ed. and Nitta, Y.（2011）, 英語複合前置詞の諸形態と日本語複合辞との対照 --- 明晰・精密な表現への志向, 電子情報通信学会（IEICE）思考と言語研究会（TL 研）研究会研究資料（2011- 2 - 5 ）「文型と意味」特集

清水義夫 , 記号論理学講義, 東京大学出版会（2013）

Slocum, J.（1985）Machine Translation---Its History, Current Status and Future Prospects, *Computational Linguistics,* 11(1)

Yoneda, Nobuo, On the Homology Theory of Modules, J. Fac. Sci. Tokyo, Sec. Ⅰ .7（1954）pp.193-227

Yoneda, Nobuo, On Ext and Exact Sequences, J. Fac. Sci. Tokyo, Sec. Ⅰ .8（1960）pp.507-526

以上