# 比例類推的記号列形式言語

イヴ・ルパージュ
yves.lepage@slt.atr.co.jp
エイ・ティ・アール音声言語通信研究所

## 1　はじめに

本論文は類推のための口頭弁論である。その動機は、もう衰退してしまったにもかかわらず、生成言語学派の影響が今でも感じられるからである。類推の話であれば、生成言語学派が類推に強く反対したことはよく知られている。

その反対は二つの仮説に基づいている。一つは生得仮説であって、もう一つは文脈自由仮説である。その二つの仮説は否定されたので、昔から（例えば、(Saussure　16)、第IV章や(Paul 20)、第V章）言語学で使われてきた「類推」という概念が表舞台に帰り咲く時期ではないかと考えられる。

「昔から使われてきた」からと言う理由だけではなく、新しい議論を導入して類推に基づく形式言語の特性を証明し、自然言語にその十全性や妥当性を示す。

## 2　生得仮説　対　類推

生得仮説というのは、心理学的な議論である。言語を習っている子供には、推論するために想像する用例は少なすぎると言う理由で、実習手順も推論手順も言語そのものではないと言う結果を主張する。

言語学者のItkonenは(Itkonen　&　Haukioja　97)、132ページで次のように述べている。

> Chomsky (1957) had claimed that there was no method for deriving grammars from linguistic data; and because any such method has to be analogical, it followed (or seemed to follow) that there was no use for analogy in linguistics.

しかし、生得仮説に反駁は2種類あった。一つは、知的反論である。データが少ないというのは絶対間違いだと多くの言語学者が考えている上、最近得られた心理学実験の結果に基づいて反論された。チョムスキーが立てた仮説は記述的妥当性にミスがあって、次の文脈自由仮説の準備のための論点を先取りしただけであったと考えられる。さらに、最近データが少ない環境でも、ロボットでのシミュレーション実験で、文法（または統辞）の誕生が見られる実験が行なわれた(Steels 97)。つまり、生得仮説は前提が正しくなくなったため、議論にならなくなった。

その結果、類推という現象そのものの研究をしようという動きが出てきた（Paul, Saussure, Kuryłowicz, Mańczak）。ところで、類推の現象では言語学で伝統的に優勢であったのに、生成言語学派は類推を言語学の歴史の中ではじめて反対した。

言語学の歴史をふり返ると、ギリシア語とラテン語の文法伝統でも（Varro, Aulus-Gellus）、アラビア語の言語学伝統でも（qiyāsと呼ばれる現象）類推を意味する。カザン派（Baudouin de Courtenay, Kruszewski）でも音韻学（Hjemslev, Jakobson）でも類推の手順が見える。結局、その音韻学の目標は、類推の関係を見付けることであると考えられる。たとえば、/p/は/t/となり、/b/は/d/に当たると言うときはそうではないかと考えられる。

次節に関連して言うと、生成言語学派では、文脈自由形式言語であくまでも何でもしようという考えがあったが、類推を使う言語学者は、類推で何でも説明しようとしているわけではない。

## 3　文脈自由仮説　対　弱い文脈依存性

チョムスキーの考えでは、自然言語は、文脈自由現象である。だが、Itkonenは次のように述べている(Itkonen 94)。

> ...Chomsky practised some sort of 'universal grammar of English', taking the syntax of his native language to be innate component of the human mind.

有名な話だが、二つの実際言語の用例で文脈自由性に反駁があった(Culy 85)、(Shieber 85)。

両方の反論では、$a^n b^m c^n d^m$ という形式言語を使用して証明された。

ここで、チョムスキーの形式言語の分類について指摘を行なう。その分類は、固有的や本質的でもない特徴付けだと言えるかもしれない。文法体系によって、手順の観点からの特徴付けであるからである。従って、それぞれの異なる分類の特徴的な形式言語用例はよく似ている。素人の目では、その直感的によく似ている物が、なぜ違う分類に属するか考えにくいと思われる。

この実際に対して、たとえば $a^{2^n}$ と言う形式言語は全く自然言語では見つからないと考えられるので、自然言語は少しだけ文脈依存的であると考える。従って、次に、その「少しだけ文脈依存的」の形式的な定義が必要になった。

その点に対して、Aravind Joshi は「弱い文脈依存性」という言葉を提案した(Joshi 85)。しかし、弱い文脈依存形式言語がＴＡＧで解析できる言語であるという定義をすると、これは手順によっての特徴付けになるから、批判されるかもしれない。それに対して、本質的な異なる定義の提案もあった(Marcus & al. 96)：即ちあらゆる弱い文脈依存形式言語の要素に対して、長さが予め与えられた値 $k$ 以下の差異で同じ言語の違う要素が見付けられるという特徴付けである。つまり、要素の間の穴が無限に大きくならない。形式的に表現すると、次のようになる。

**定義 1 (制限された成長)** $\mathcal{L}$ を形式言語とすると、制限された成長の特性を持つならば、

$$\exists k \in \mathbb{N} \,/\, \forall w \in \mathcal{L},\ \exists w' \in \mathcal{L},\ \Big|\,|w| - |w'|\,\Big| \leq k$$

## 4 類推関係の基礎公準

以上の背景では、類推には研究の余地があるかも知れないという議論であった。これだけではなく、自然言語と形式言語の関係に対して、類推は十全であることを数学的証明に基づいて示す。

基礎的な類推の特性から始めると、アリストテレスの著書で見つかる特性を公理として建てる。二つの特性のうちの一つは、内項の交替であって、その公理に従うと、中央にあるものは交替できる。

**公理 1 (内項の交替と均等の対称)**

$$A : B = C : D \quad \Leftrightarrow \quad A : C = B : D$$

もう一つは、均等の対称に従っている。

**公理 2 (内項の交替と均等の対称)**

$$A : B = C : D \quad \Leftrightarrow \quad C : D = A : B$$

これら二つの特性は直感的なことであると考えられるので、その特性を公理とするのは誰にでも認められると思われる。以上の公理に基づいて、次の定理が得られる。この特性は $A$、$B$、$C$、$D$ がどの集合の要素であっても有効である。

**定理 1 (等値形)** 以下の 8 つの類推関係式が等値形である。

$$\begin{array}{ll}
A : B = C : D & (i) \\
A : C = B : D & (ii) \\
B : A = D : C & (iii) \Leftarrow ii{+}vi{+}ii \\
B : D = A : C & (iv) \Leftarrow iii{+}ii \\
C : A = D : B & (v) \Leftarrow ii{+}iii \\
C : D = A : B & (vi) \\
D : B = C : A & (vii) \Leftarrow ii{+}vi{+}iii \\
D : C = B : A & (viii) \Leftarrow iii{+}vi
\end{array}$$

以下では、$A$、$B$、$C$、$D$ は記号列であるとする。ある形式言語が類推形式言語であることを証明するためには、次の二つの公理も必要である。

一つは、多数の用例の検査によって、$A$ にある記号はすべて $B$ や $C$ にある順番で見つけられることである。

**公理 3 (順番での包含)** $\mathcal{V}$ を記号の集合とすれば、$\forall (A,B,C,D) \in (\mathcal{V}^*)^4$,

$$|A| \leq \mathrm{sim}(A,B) + \mathrm{sim}(A,C)$$

$|A|$ と $\mathrm{sim}(A,B) + \mathrm{sim}(A,C)$ の差は $\mathrm{com}(A,B,C,D)$ と呼ばれ、類推関係の四つの記号列の同じ順番で共通する記号の数を表す。

もう一つの公理の表現は複雑だが、意味は記号の交わりのない二つの類推関係が連続できると言う命題だけである。連続のしかたは三通りあるが、一つは直感的な順番である（公理 4、次ページを参照、左上）。

## 5 記号列類推関係の特徴付け

数学的に、以上の四つの公理だけで色々な定理の証明ができる。例えば、

公理 4 (連続) $\mathcal{V}$ を記号の集合とすれば、
$\forall(A_1,B_1,C_1,D_1,A_2,B_2,C_2,D_2)\in(\mathcal{V}^*)^8$,

$$\left.\begin{array}{l} A_1:B_1=C_1:D_1 \\ A_2:B_2=C_2:D_2 \\ (\overline{A_1}\cup\overline{B_1}\cup\overline{C_1}\cup\overline{D_1})\cap \\ (\overline{A_2}\cup\overline{B_2}\cup\overline{C_2}\cup\overline{D_2}) = \emptyset \end{array}\right\} \Rightarrow$$

$$\left\{\begin{array}{l} A_1A_2:B_1B_2=C_1C_2:D_1D_2 \\ A_1A_2:B_1B_2=C_2C_1:D_2D_1 \\ A_1A_2:B_2B_1=C_2C_1:D_1D_2 \end{array}\right.$$

定理 2 $\mathcal{V}$ を記号の集合、dist を文字列距離（削除コスト 1、置換コスト 2、挿入コスト 1）とすると

$$\forall(A,B,C,D)\in(\mathcal{V}^*)^4,\quad A:B=C:D \quad\Rightarrow$$

$$\left\{\begin{array}{l} |A|+|D|=|B|+|C| \\ \mathrm{dist}(A,B)=\mathrm{dist}(C,D) \\ \mathrm{dist}(A,C)=\mathrm{dist}(B,D) \\ \mathrm{com}(A,B,C,D)=\frac{1}{4}\times(|A|+|B|+|C|+|D| \\ \qquad\qquad -\mathrm{dist}(A,B)-\mathrm{dist}(A,C) \\ \qquad\qquad -\mathrm{dist}(B,D)-\mathrm{dist}(C,D)) \end{array}\right.$$

この方程式の体系は、記号列類推関係に必要かつ十分の条件になるためにはまだ不十分であるが、この方程式の体系により、取り敢えず実際のアルゴリズムの構成ができる。

## 6 類推的記号列形式言語の定義

本節では、類推形式言語の本題に入る。まず、類推的導出を定義する。

定義 2 $\mathcal{V}$ を記号の集合、$\mathcal{M}\subset\mathcal{V}^*\times\mathcal{V}^*$ とすると $(v,v')\in\mathcal{M}$ の要素を $v\rightarrow v'$ で表せば、$\vdash_{\overline{\mathcal{M}}}$ を $\mathcal{M}$ を法としての類推的導出という。

$$\forall(w,w')\in\mathcal{V}^*\times\mathcal{V}^*,$$
$$w\vdash_{\overline{\mathcal{M}}}w' \Leftrightarrow \exists v\rightarrow v'\in\mathcal{M}/\ w:w'=v:v'$$

ここでは、普通の形式文法や導出システムと同じ記号法を流用して、$\mathcal{M}$ の要素を矢で表した。しかし、普通の標準導出ルールでは、$v$ に対してマッチしたあと、$w$ から $w'$ が導出されるが、ここは、$v$（$w$ ではなく）が $w$ と $v'$ に同時にマッチするかどうかを判断した後、$w'$ が生成（導出）される。即ち、ある $\mathcal{M}$ の要素に沿って類推関係で $w'$ を $w$ から導出することになる。

類推的導出から標準的な要領で類推的言語を次のように定義する。

定義 3 $\mathcal{V}$ を記号の集合、$\mathcal{A}\subset\mathcal{V}^*$ と $\mathcal{M}\subset\mathcal{V}^*\times\mathcal{V}^*$ とすると $\vdash^{+}_{\overline{\mathcal{M}}}$ が $\vdash_{\overline{\mathcal{M}}}$ の推移的な閉包とすれば、$\Lambda(\mathcal{A},\mathcal{M})$ を類推的記号列形式言語という。

$$\Lambda(\mathcal{A},\mathcal{M}) = \mathcal{A}\cup\{\,w'\in\mathcal{V}^*\ /\ \exists w\in\mathcal{A}\ /\ w\vdash^{+}_{\overline{\mathcal{M}}}w'\,\}$$

$\mathcal{A}$ は観察形の集合であり、$\mathcal{M}$ はモデル形または観察変形の集合であると解釈できる。ある記号列の文法性を判断するため、モデルに従って約分を行なった後、観察形集合 $\mathcal{A}$ の依属で判断する。

その定義に対しての考察として、一つはこの形式言語の定義では、非終端記号がない。Solomon Marcus の「contextual grammar」(Marcus & al. 96) においても非終端記号はない。

以上の解釈は解析側の説明で、生成をする時に、まず $\mathcal{A}$ の観察形の集合から始まって、運に任せてあらゆる $\mathcal{M}$ の変形を適用し、類推解決して新しい記号列が得られる。

## 7 自然言語にの十全性や妥当性

今まで提案された類推的言語の定義に基づいて、次にその自然言語の十全性や妥当性の話をしたい。

まず、それぞれのチョムスキー形式言語分類の例はどうなるかと言う問題の検討をすると、興味深い成果で、$a^n$ と言う正規言語も、$a^nb^n$ と言う文脈自由言語も、$a^nb^nc^n$ と言う文脈依存言語も類推的言語であると証明された (Lepage 00)。

定理 3

$$\begin{aligned} \Lambda(\{a\},\{a\rightarrow aa\}) &= \{a^n/n>0\} \\ \Lambda(\{ab\},\{ab\rightarrow aabb\}) &= \{a^nb^n/n>0\} \\ \Lambda(\{abc\},\{abc\rightarrow aabbcc\}) &= \{a^nb^nc^n/n>0\} \end{aligned}$$

望ましいことに、これは一般化できる。即ち $a_1^na_2^n\ldots a_m^n$ も類推的言語である。さらに、チョムスキー形式言語分類でその異なる言語は類推的言語として表すと、同じような記述となる。

定理 4 $\Lambda(\{a_1a_2\ldots a_m\},\{a_1a_2\ldots a_m\rightarrow a_1^2a_2^2\ldots a_m^2\}) = \{a_1^na_2^n\ldots a_m^n/n>0\}$

形式言語の例に戻ると、自然言語は文脈自由言語ではないという証明で使われた有名な形式言語も類推的言語であると簡単に証明されたことになる。

定理 5 $\Lambda(\{abcd\},\{abcd\rightarrow abbcdd,abcd\rightarrow aabccd\}) = \{a^mb^nc^md^n/n,m>0\}$

以下に、二つ目の重要な成果について述べる。類推的言語が、恐らく自然言語に対して十全であるとすると弱い文脈依存性に関するもう一つの議論が必要になるかもしれない。そこで、次の定理の証明ができる。

定理 **6** あらゆる類推的記号列形式言語は制限された成長の特性を持つ。

しかし、いずれの正規言語も文脈自由言語が類推的言語であるかどうかという問いにまだ正しく答えられない。恐らくそうであろうと推測するが、数学的な証明はまだできていない。同時に、中埋め込み (central embedding) の質問については、van Dyck 形式言語も類推的言語であろうと考えているが、その証明もまだできていない。

## 8 おわりに

計算に対して付加（和のこと）が基礎演算であるなら、自然言語の記述に対して類推も同様に基礎的な演算であると考えられる。

本論文では、類推の復権についての弁論を試みた。公理を立て、その公理を出発点として記号列の類推関係について色々な定理を示した。そこから、類推的記号列形式言語の定義を標準導出システムと全く同じく定義の提案をした。類推形式言語は直感的であると主張した。すなわち、観察形集合の依属で判断し、類推解決で約分し、非終端記号を使用しない。

また、数学的に三つの重要な命題の証明をした。一つは、普通にチョムスキー形式言語分類を例証する用例は、すべて類推的言語であり、類推的言語の記述で同様に単純な類推言語である。それにより、チョムスキー形式言語分類の非直感さが避けられることを指摘しておく。

もう一つは、自然言語は文脈自由言語ではないという証明で使われた有名な形式言語も類推的言語であることを示した。

最後に、あらゆる類推的言語が制限された成長の特性を持つことを証明した。その特性は弱い文脈依存性に絡んでいる為、類推的言語は弱く文脈依存すると言えるかもしれない。従って、類推的言語は自然言語に役立つと考えられる。

## 参考文献


Christopher CULY
The Complexity of the Vocabulary of Bambara
*Linguistics and Philosophy*, vol. 8, 1985, pp. 345-351.

Esa ITKONEN & Jussi HAUKIOJA
A rehabilitation of analogy in syntax (and elsewhere)
in András Kertész (ed.) *Metalinguistik im Wandel: die kognitive Wende in Wissenschaftstheorie und Linguistik* Frankfurt a/M, Peter Lang, 1997, pp. 131-177.

Esa ITKONEN
Iconicity, analogy, and universal grammar
*Journal of Pragmatics*, 1994, vol. 22, pp. 37-53.

Aravind K. JOSHI
Tree adjoining grammars: How much context-sensitivity is required to provide reasonable structural description?
in Dowty *et al.*, *Natural language processing*, Cambridge University Press, Cambridge, 1985, pp 206-250.

Yves LEPAGE & 飯田仁
言語に依存しない早期終了型類推解決手法
言語処理学会第 4 回年次大会, 九州大学, 1998 年 3 月, pp. 266-269.

Yves LEPAGE
Languages of Analogical Strings
*Proceedings of COLING 2000*, vol 1, Saarbrücken, July-August 2000, pp. 488–494.

Solomon MARCUS, Carlos MARTIN-VIDE, Gheorghe PĂUN
*Contextual Grammars versus Natural Languages*
Turku center for Computer Science, TUCS technical Report No 44, Sept. 1996.

Hermann PAUL
*Prinzipien der Sprachgeschichte*
Niemayer, Tübingen, 1920.

Ferdinand de SAUSSURE
*Cours de linguistique générale [1ère éd. 1916]* publié par Charles Bally et Albert Séchehaye, Payot, Lausanne et Paris, 1995.

Stuart M. SHIEBER
Evidence against the Context-Freeness of Natural Language
*Linguistics and Philosophy*, vol. 8, 1985, pp. 333-343.

Luc STEELS
Origin of Syntax in Visually Grounded Robotic Agents
*Proceedings of IJCAI-97*, vol. 2, pp. 1632–1641, Nagoya, August 1997.