

自由語順の数学

The Mathematics of Free Word Order

金沢 誠
Makoto Kanazawa

どんな研究？

人間の頭の中にある母国語の文法はどのような形式で書かれているのかという問題に対して、形式言語理論という数学を使ってアプローチしています。

次のようなパズルを考えてみて下さい。このパズルでは、最初に、aとbを同数含んだ文字列を与られます。例えば次のような文字列です。

a a a a b b b a b b b a

aの数とbの数が同じであること以外にこの文字列に規則性はありません。パズルの目的は、ある制約に従いながら、この文字列をどんどん分割して行き、最終的に1文字だけからなる文字列に分解することです。従わなければならない制約とは、どの局面でも、手持ちのすべての文字列において、aの数とbの数の差が1以内におさまることです。例えば、上の文字列を次のように分割することは許されませんが、

a a a a b b b a b b b a

次のように分割することは許されません。

a a a a b b b a b b b a

各局面で2文字以上を含んだ文字列を1つ選んで、2つに分割します。そのような文字列がなくなったときにパズルは完了します。

このパズルの解き方はいろいろありますが、例えば、次のような戦略で解くことができます。

- 選んだ文字列の中にaとbが同数含まれているときは、最初の1文字と残りの文字列に分割する。
- 文字列の中にaの方がbよりも1つ多く含まれているときは、最初の文字がaなら、最初の1文字と残りの文字列に分割する。さもなくば、文字列の2文字目以降を

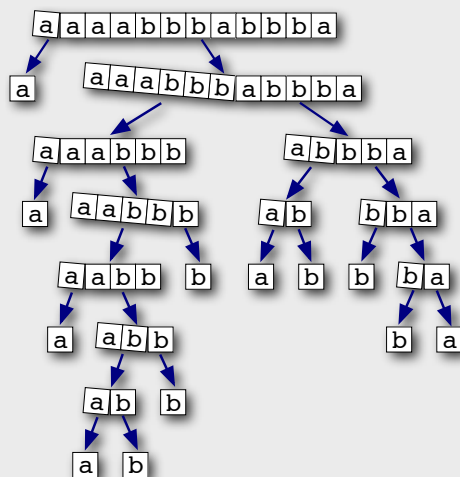
何がわかる？

「語順にまったく制約のない言語」を表す文字列集合を記述するのにどれくらい複雑な文法形式が必要になるかという問題を数学的に解明します。

1文字ずつ見て行き、初めてaの数がbの数に追いついた箇所で分割する。

- bの方がaよりも1つ多く含まれているときは上の規則でaとbを入れ換えた規則に従う。

この戦略に従って上の例を分解してみましょう。



形式言語理論では、文字列集合を**形式言語**または単に**言語**と呼びます。上のパズルに解法が存在することは、形式言語理論の言葉では、言語

$\{w \mid w \text{ は } a \text{ と } b \text{ を同数含む文字列}\}$

が**文脈自由文法**で記述できるという事実に対応しています。文脈自由文法は、プログラミング言語の文法の記述に標準的に使われている文法形式です。もともと言語学から生まれた概念ですが、自然言語の文法の記述という目的には必ずしも十分でなく、1980年代から**木接合文法**などのより複雑な文法形式が研究されて来ました。

さて、2つ目のパズルです。このパズルでは最初にaとbとcの3種類の文字を同数含んだ文字列が与えられます。パズルの目的は、与えられた文字列を1文字だけからなる文字列に分解することで、従わなければならない制約は、常にすべての文字列において「aの数 - cの数」と「bの数 -

cの数」のそれぞれが-1から1の範囲におさまることです。

最初のパズルと違って、2つ目のパズルには必ずうまく行く解法が存在しません。例えば、次の文字列をこのパズルの制約に従って分解することは不可能です。

a a b b b b c c c c a a

最初の局面で先頭のaを切り離すか末尾のaを切り離すことはできますが、その先で行き詰まってしまう。

さらに、パズルの制約で「-1から1の範囲」とあるところを、適当な自然数 n を取って、「 $-n$ から n の範囲」に弱めても状況は変わらないことがわかります。次の形の文字列を分解することができないからです。

$a^{2n}b^{4n}c^{4n}a^{2n}$

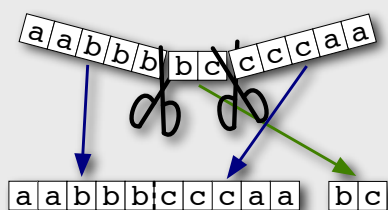
この事実、言語

MIX = { w | w は a と b と c を同数含む文字列 }
を記述する文脈自由文法が存在しないことを意味します。
前述の木接合文法は文脈自由文法よりも強力な文法形式で、文脈自由文法で記述できない言語も木接合文法で記述できる場合があります。では、MIXは木接合文法で記述できるでしょうか。この問題は、初めて提示されてから30年近く経ちますが、これまで誰も解決することができませんでした。

MIXは、人間の言語における自由語順の現象と関係があります。MIXは、まったく語順に制限のない言語、いわば究極の自由語順言語を表現した形式言語と見なすことができます。人間の言語には、語順の自由度が高い言語と低い言語がありますが、完全に語順が自由な言語は存在しません。もし木接合文法がMIXを記述できないなら、この事実

に純粋に形式的な説明を与えることが可能になります。
私は、Sylvain Salvati (INRIA Bordeaux) との共同研究で、この問題を解決しました。我々は、まず、木接合文法がMIXを記述できることと、次のパズルに解法が存在することが同値であることを証明しました。このパズルでは、上の第2のパズルのように a と b と c を同数含んだ文字列からスタートしますが、各局面で、選んだ文字列を2つに分割するという操作だけでなく、より複雑な次の操作が許されます。

文字列を3つに分割し、両端の文字列をつなげて1つの文字列にする。



ただし、継ぎ目にはそれとわかる跡がつくこととします。
また、この操作を適用する文字列にすでに継ぎ目がある場



Makoto Kanazawa
Sylvain Salvati

2ヶ月前

MIX is not a tree-adjoining language.
(MIXは木接合文法で記述できない。)



Sylvain Salvati

1年4ヶ月前

MIX is a 2-MCFL. (MIXは2-多重文脈自由文法で記述できる。)



Aravind K. Joshi 21年前
K. Vijay-Shanker
David Weir

It is not known whether

TAG can generate MIX. This has turned out to be a very difficult problem. (木接合文法でMIXを記述できるかどうかかわかっていない。これは非常に難しい問題だ。)



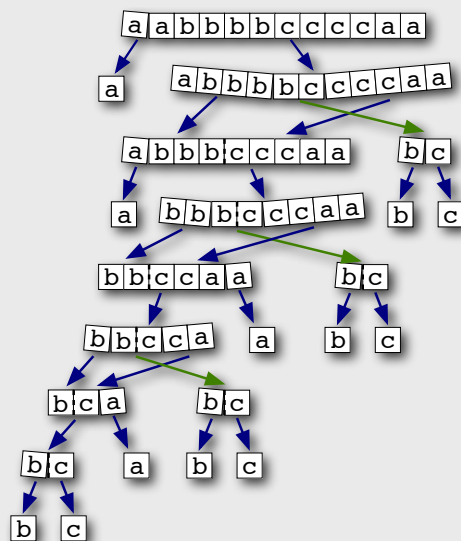
Aravind K. Joshi

27年前

TAGs cannot generate MIX, although the proof is not in hand yet. (木接合文法はMIXを記述できない。まだ証明はできていないが。)

合、継ぎ目は中央の文字列に含まなければならないとします。満たすべき制約は、途中に生じるすべての文字列において、「aの数 - cの数」と「bの数 - cの数」がそれぞれ、-2から2の範囲におさまることです。

第2のパズルで分解できなかった $a^{2n}b^{4n}c^{4n}a^{2n}$ の形の文字列も、このパズルのルールでは分解することができます。



しかし、一般にこのパズルに解法があるのかわからないのか、自明ではありませんでした。我々は、コンピュータを使っていろいろな文字列を調べ、ついに次の文字列が解を持たないことを突き止めました。

$a^5b^{14}a^{19}c^{29}b^{15}a^5$

これにより、MIXが木接合文法で記述できないことが証明でき、30年来の未解決問題が解決しました。