人間の文処理は左隅型である: 埋め込み構造と記憶負荷

橋田 浩一
電子技術総合研究所

1 はじめに

人間による言語処理においては、単なる左方枝分かれ (left branching) と右方枝かれ (right branching) では統語的な処理の負荷が生じず、**中央埋め込み** (center embedding) の深さに相関する負荷が生じているらしい (Miller & Chomsky, 1963)。英語では左方枝分かれの長さに制限があるとの説 (Yngve, 1960; Sampson, 1997) もあるが、左方枝かれ構造の処理が困難だという証拠はないので、この制限は記憶負荷以外の要因に帰せられるべきだろう。また、人間の言語処理は実時間で行なわれていると考えられる (Marslen-Wilson, 1975)。

下降型の統語解析では左方枝分かれの長さに応じた時間計算量と空間計算量が発生する。また、上昇型の統語解析では、右方枝分かれの長さに応じた時間計算量と空間計算量が発生する。これに対し、左隅型の統語解析は実時間の (入力単位ごとの時間計算量が定数である) 計算であり、また、左隅型の統語解析では左方枝かれと右方枝かれではなく中央埋め込みの深さが空間計算量と相関関係を持つ。

これらの事実から、下降型および上昇型の統語解析よりも左隅型統語解析の方が心理的実在性が高いと考えられる。以下では、人間の文処理過程が左隅型であるという仮説を検証する試みについて述べる。

2 左隅型処理と中央埋め込み

構文木を2分木とすると、左隅統語解析においては、構文木の根から最近の入力語に至る経路が屈曲するところにある節点を作業記憶に保持しておく必要がある。たとえば図1では、「本」が入力された直後に黒丸の節点が記憶されている (ハッチングは処理済みの構造を表わす)。その個数は中央埋め込みの深さと相関関係にあるが、中央埋め込みの深さに比例するわけではない。たとえば図2で黒丸を構文木の中の節点とすると、中

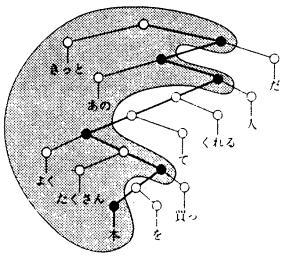

図1: 左隅統語解析における作業記憶

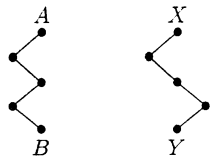

図2: 中央埋め込みと屈曲の関係

央埋め込みの深さは A と B の間でも X と Y の間でも2だが、屈曲の回数は A と B の間で3、X と Y の間では2である。

したがって、人間の文処理過程が左隅型なら次の予測が成立するはずである。

(*) 作業記憶の負荷は中央埋め込みの深さではなく屈曲の回数に比例する。

中央埋め込みの深さが等しく屈曲の回数が異なるような複数の文の組で屈曲回数の多い方が人間にとって処理が難しいような事例は知られている (Hasida, 1990) が、そのような組は最小対からはほど遠いため、少数の事例で (*) を立証することはできない。また、そのような事例を大量に作るのはあまりにも膨大なコストがかかる。

そこで、大量の言語データの統計的解析によって (*) の検証を進めている。これまでに、Penn Treebank の WSJ コーパスに関して構文木の根を始点とする経路の形の分布を調べた。まず、名詞または名詞句については3方以上の分岐を右方枝かれに、他の範疇については左方枝分かれにすることにより、各文の構造を2分木に変換した。たとえば図3の左側の局所木を右側の局所木に変換する。これにより、根を始点とする経路が L と R (それぞれ左下と右下への枝) の列によって記述できる。たとえば図2の A が根だとすると、B までの経路は LRLR である。

次に、このような経路の分布を調べた。L' と R' の対数成長率の分布をそれぞれ図4と図5に示す。ここで経路 PB の対数成長率とは $\log(PB$ の頻度$/P$ の頻

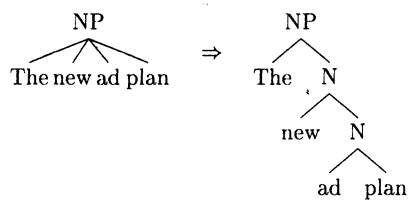

図 3: 2 分木への変換

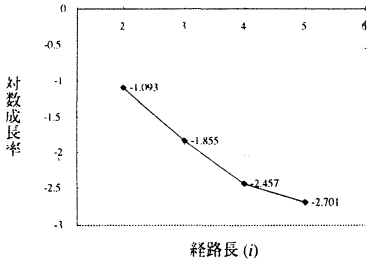

図 4: L^2 の対数成長率

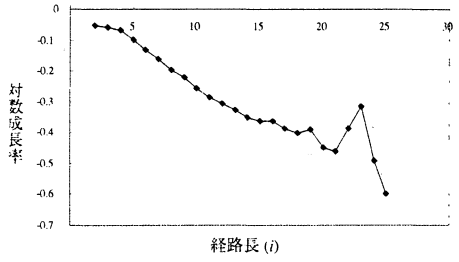

図 5: R^2 の対数成長率

度) である (P は任意の経路、B は L または R)。すなわち、経路の頻度は経路の長さに応じて超指数的に減衰する[1]ので、経路の頻度の対数ではなく対数成長率に関して線形回帰分析を行なう方が適切である。そこで、経路の長さ、R の個数、末尾の分岐 (L か R か)、中央埋め込み指数 (1 を加えて 2 で割った値が中央埋め込みの深さになる変数) および屈曲回数を説明変数として対数成長率の線形重回帰分析を行なった[2]。ここで、経路 PB に関する変数の値の組は P と同じ頻度を持つと見なした。以下同様である。

しかし、中央埋め込み指数と屈曲回数の寄与に関しては有意な結果が得られなかった。そこで、経路の長さと R の個数と末尾の分岐の値の各組に対して、中央埋め込み指数と屈曲回数と対数成長率を平均 0、標準偏差 1 に正規化し、それをすべて合わせたデータを用いて中央埋め込み指数と屈曲回数により対数成長率の線形回帰分析を行なった。その結果を表 1 に示す。これにより、中央埋め込みよりも屈曲の方が経路の頻度の減少の大きな原因であることが結論できるだろう。

表 1: 対数成長率の重回帰分析

	回帰係数	偏相関係数
中央埋め込み指数	−0.1112457	−0.0766958
屈曲回数	−0.3929171	−0.2820078

3 おわりに

人間の文処理過程が左隅型であるとの仮説を支持する新たな証拠を得た。しかし、この仮説を証明するには予測 (*) が成り立つことを示す必要があり、それには上記の統計解析では不十分である。そのためには、統計モデルを洗練して精度を高める必要があるだろう。

参考文献

Hasida, K. (1990). A Constraint-Based Approach to Linguistic Performance. *Proc. of the 13th International Conference on Computational Linguistics*, Vol. 3, pp. 149–154. Helsinki.

Marslen-Wilson, W. D. (1975). Sentence Perception as an Interactive Parallel Process. *Science*, **189**, 226–228.

Miller, G. A. & Chomsky, N. (1963). Finitary Models of Language Users. In R. Luce, R. Bush, & E. Galanter (Eds.), *Handbook of Mathematical Psychology, Vol. II*, pp. 419–491. John Wiley and Sons.

村田 真樹・内元 清貴・馬 青・井佐原 均 (1999). 日本語文における係り受けとマジカルナンバー 7±2.『言語処理学会第 5 回年次大会論文集』, pp. 48–51. 電気通信大学.

Sampson, G. (1997). Depth in English Grammar. *Journal of Linguistics*, **33**, 131–151.

Yngve, V. H. (1960). A Model and an Hypothesis for Language Structure. In C. A. Ferguson & D. I. Slobin (Eds.), *Proceedings of the American Phychological Society*, Vol. 104, pp. 444–466.

[1] Yngve らの主張に反し、左方枝分かれと右方枝分かれの間にこの点で定性的な差はない。

[2] 村田ら (1999) は経路の屈曲回数の最大値による文の頻度の分布を調べている。