日本語文における係り受けとマジカルナンバー7±2
— 英語文の場合も含めて —

村田 真樹　内元 清貴　馬 青　井佐原 均
郵政省 通信総合研究所 関西先端研究センター 知的機能研究室

1 はじめに

George A. Miller は1956年に人間の短期記憶の容量は7±2程度のチャンク[1] (スロット) しかないこと，つまり，人間は短期的には7±2程度のものしか覚えられないことを提唱した [1]. 本研究では，京大コーパス [2] を用いて日本語文の各部分において係り先が未決定な文節の個数を数えあげ，その個数がおおよそ7±2程度でおさえられていたことを報告する．この結果は，人間の文の理解過程において係り先が未決定な文節を短期記憶に格納するものであると仮定した場合，京大コーパスではその格納される量がちょうどMiller のいう7±2の上限の9程度でおさえられており，Miller の7±2の提唱と矛盾しないものとなっている．また Yngve によって提案されている方法 [3] により英語文でも同様な調査を行ない，NP程度のものをまとめて認識すると仮定した場合，必要となる短期記憶の容量が7±2の上限の9程度でおさえられていたことを確認した．

近年，タグつきコーパスの増加により，コーパスに基づく機械学習の研究が盛んになっているが [4]，タグつきコーパスというものは機械学習の研究のためだけにあるのではなく，本研究のような言語の数量的な調査にも役に立つものである．タグつきコーパスがあればこういったことができると思っていたのに，コーパスがないということだけでなしえずにきたことは数多く存在する．近年のタグつきコーパスの出現は，遠い昔に置き忘れてしまったものをもう一度思い起こす時代がきていることを意味している．

2 短期記憶と7±2

Miller は短期記憶の容量を計る，言葉，音感，味覚，視覚などを対象とした種々の実験におけるデータが，いずれも概ね7±2であったことから，人間の短期記憶の容量は7±2程度のチャンクであることを提唱した．7±2の「±2」は個人差を意味しており，一般の人は7個程度，人によっては二つ多いめに，もしくは二つ少なめに覚えることができることを意味している[2].

7±2の研究は心理学の分野に属するものではあるが，工学の分野にも応用することができる．例えば，文生成の研究では7±2の容量を越える文を作成するとわかりにくい文になるであろうから [7]，その条件を満足するように文生成を行なうということがある [3]. また，画像処理の分野では最近はやりのカーナビを構築する際に，一画面に多くの情報を与えすぎると人間の認識に支障をきたすので，7±2程度のものしか提示しないようにするなどの研究を行なっているものもある [8]. 7±2の研究は，単なる知的好奇心による人間の解明に役に立つだけでなく，実際の社会においても利用されうる有益な研究なのである．

3 日本語文での調査報告

本節では実際に京大コーパスを用いて行なった調査結果を報告する．(京大コーパスは1995年の毎日新聞のデータをもとに作成されたタグつきコーパスである．) まず調査方法を述べる．本研究では文の理解とは文の係り受け構造の解析であるとみなし，文を理解するときに短期記憶に格納することが必要とされるものは，係り先が未決定な文節であると考える．図1に例を示す．図は「その少年は小さい人形を持って

[1] チャンクとはある程度まとまった情報を計る，情報の認知単位のこと．

[2] Miller の7±2とは直接関係ないが，言語の特性を短期記憶と結びつけて議論しているものに Lewis のマジカルナンバ 2or3[5] という研究がある．これは中央埋め込みの数に関する研究で，英語では主節と一つの中央埋め込みの二つの文 (ここでいう文は，句点によって区切られる文ではなく，動詞によって構成される節のような部分的な文のことを意味する．) まで (日本語では主節と二つの中央埋め込みの三つまで) しか短期的に覚えることができないと主張するもので，これは英語については古くは Kimball の7つの原則 [6] のうちの四つ目の「文二つの原則」にあげられていることである．これらの研究は文理解において中央埋め込みの数に制限があるのは人間の短期記憶の容量に限界があるためであると考えているものである．

表 2: 係り先が未決定な文節の個数が 10 であった箇所を持つ文

調べでは、同町○○(地名)、建設業、○○○○(人名)容疑者は、九月六日告示の町議選を無投票にするため、逮捕された議員十五人が出し合った現金四百五十万円を告示日の六日、同町○○○(地名)、元町議で農業の○○○○(人名)容疑者に、また百万円を告示翌日の七日、新人で出馬予定だった同町○○○(地名)、会社員、○○○○(人名)容疑者と夫の会社代表、○○(人名)容疑者の二人に渡した疑い。

国はその後、このうち二十三点の公開は「やむを得ない」と認めたものの、主に電子機器などを置いてある地下部分の資料二十一点については「ＡＳＷＯＣはシーレーン防衛のための中枢基地であり、公開されると国防や警備上、重大な支障が生じる」などと主張、決定の取り消しなどを要求して争ってきた。

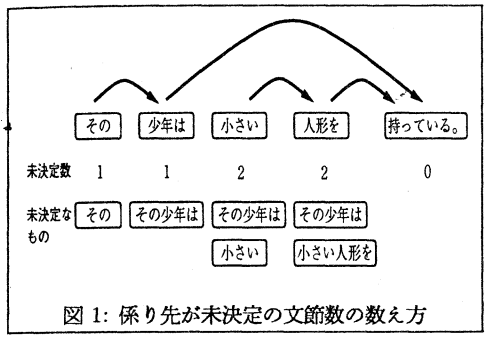

図 1: 係り先が未決定の文節数の数え方

表 1: 係り先が未決定な文節の個数の頻度統計

未決定文節数	頻度	
	文節数	文数
0	19954	90
1	52751	1352
2	59494	5022
3	38465	6823
4	15802	4468
5	4488	1593
6	1143	480
7	195	102
8	47	17
9	10	5
10	3	2

いる．」という文の係り受け構造を頭から解析するときに，各文節において係り先が未決定になっている文節の数を数えているものである．図の矢印は係り受け構造を示し，数字は係り先が未決定な文節の個数を示し，その下に係り先が未決定な文節として短期記憶に格納しなければならない要素を示している．この文を頭から見てみると，「その」が入ってきたときはそれの係り先はまだ決まっていないのでそれは覚えておかなければならず，係り先が未決定なものとして短期記憶に格納される．次に「少年は」が入ってきたときは，「その」は「少年は」に係るとわかり「その」単独ではもう今後係り受けの解析に利用する必要はないので単独で認識する必要はなく，「少年は」とくっつけて「その少年は」という形で認識され，結局係り先が未決定な「その少年は」が一つだけ短期記憶に格納されることにになる．その次に「小さい」が入ってくる．このときは新たに係り受け関係が定まるものはないので，「その少年は」と「小さい」が短期記憶に格納される．その次の「人形を」が入ってきたときは，「小さい」は「人形を」に係るので「小さい」はもう今後単独では解析に用いられることはないので，「人形を」とくっつけて「小さい人形を」とまとめて認識され，前から覚えていた「その少年は」とあわせて二つ覚えるだけでよい．最後に文末の「持っている。」が入ってくると，すべての係り受け関係が定まるので係り先の未決定数は 0 となり，短期記憶に覚えていたものはすべて忘れてもよいこととなる．

本稿では，人間が文を理解する際に以上のような過程をたどると想定し，実際に係り受け構造のタグがふってある京大コーパスにおいて，実際に上記の方法で係り先が未決定な文節の数を数えあげた．その結果を表 1に示す．この表の「文節数」の列の数字は，京大コーパス (未定義のタグがふってあった 2 文を除く 19,954 文，192,352 文節) の全文節において上記の方法で係り先が未決定な文節の数を調べ，その係り先未決定文節数ごとに，文節の頻度を調べたものである．また，表の「文数」の列は，一文中で最も大きかった係り先未決定文節数をその文の係り先未決定文節数と考えて，係り先未決定文節数ごとに文の頻度を調べたものである．この表では，未決定文節数が 10 つまり，Miller の 7 ± 2 の上限 9 を越える文が二つあったが，

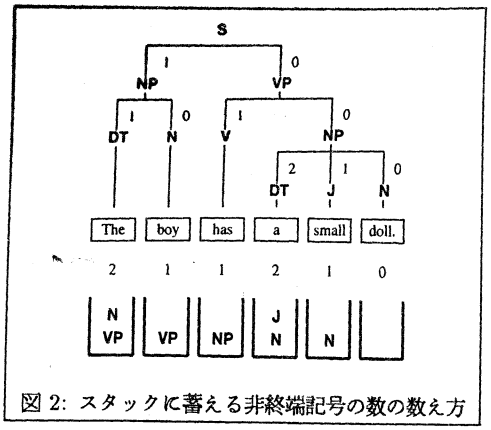

図 2: スタックに蓄える非終端記号の数の数え方

おおよそ 7 ± 2 の理論の範囲でおさまっていることを意味する．7 ± 2 を越えた二つの文を表 2に示す．これらの文は極めて読解が難解なもので真剣に読んでもなかなか理解ができない文である．7 ± 2 の上限の 9 を越えた文が少ないこと，また，7 ± 2 の上限の 9 を越えた二文も読解が難解な文であったことから，この調査結果は Miller の 7 ± 2 の理論と矛盾しないものとなっている．

本研究では京大コーパスの係り受けのタグにしたがって係り先が未決定な文節の個数を数えたが，京大コーパスは助詞「は」がつく文節が複数の係り先が想定される場合なるべく後ろの文節に係るようにタグづけされており，これを近くにかかるようにすれば統計結果は変化するだろう．また，接続詞など，記憶が必要でないかもしれない文節も数えてしまっている．これらに対して適切な処理を施せば，係り先の未決定な文節の個数はさらに少なくなると予想される．

4 英語文での調査報告

以上までは日本語コーパスを用いて文の理解に必要な短期記憶の上限を求めるものだった．本節では，英語コーパスで行なった同様な調査について報告する．

英語での文の構造解析に必要な短期記憶の容量を求める方法は Yngve[3] によってすでに提案されている．この方法は文をプッシュダウンオートマトンでトップダウンに解析する際にスタックに蓄えられる S や NP などの非終端記号を短期記憶するべきものと考え，このスタックにつまれる記号の個数を数えるものである．

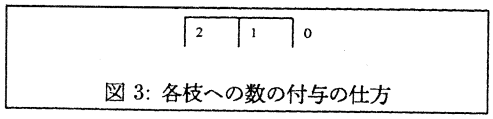

図 3: 各枝への数の付与の仕方

図 2は “The boy has a small doll.” をプッシュダウンオートマトンで解析したときにスタックに蓄える非終端記号の数の数え方を示したものである．図 2の下の方にある四角い箱は人間の短期記憶の格納庫に相当するスタックを表す．この文を頭から見て “The” が入ってきたときに，トップダウンで解析するのでまず最初に S から始まって S を (NP VP) に変形し VP を覚えておいて NP を (DT N) に変形して N を覚えておいて DT の部分を “The” と認識するので[3,4]，都合 “The” のところでは VP と N の二つをスタックに覚えておく必要がある．同様な考え方でスタックに積む必要のある非終端記号は図 2のようになり各単語でのスタックに積んでおく必要のある数は図 2のように “2,1,1,2,1,0” となる．Yngve はこのスタックに積んでおく必要のある数を簡単に数える方法も示している．それは，図 2の構文木の各枝に図 3に示した要領で数字をふり S から単語までの経路の数字を足したものがスタックに積む個数とする方法である．“The” を見ると S,NP,DT と見て 1 と 1 があるので足して 2 となりスタックの数 2 と一致する．

この方法で Penn Treebank[9] の Wall Street Journal のコーパス (49,208 文，1,122,857 単語) でスタックに積まれる数を数えてみた．その結果を表 3(a) に示す．表の「単語数」はスタックに積まれる非終端記号数ごとの単語の頻度を意味し，「文数」は一文で最も多く積んだときの非終端記号数をその文の非終端記号数としたときの非終端記号数ごとの文の頻度を意味する．ただし，ピリオドなどの記号は削除し，また “and” などによって構成される並列節の表現形式がこのコーパスではスタックの非終端記号の数を余分に数えるような構造になっていたので，その部分は図 4のよ

[3] このトップダウンの認識はわれわれ日本人には若干不自然に思えるものだが，英語文の場合はボトムアップよりも，文が成立すると仮定しまた主語と述部があると仮定して読み進めるトップダウンの方が人間の文理解のモデルとしてもよいとされている [6].

[4] 図 2では NP を展開する際に (DT N) と (DT J N) の二種類が想定でき，“The” が入ってきただけではそのどちらであるかを特定できないという問題がある．Yngve の方法はそういう問題を持っているが，タグつきコーパスからの計数方法が非常に容易であるため本稿はそれにしたがって計数している．

表 3: スタックに積まれる非終端記号の個数の頻度統計

(a) 単語部分で集計

積む記号数	頻度	
	単語数	文数
0	49208	132
1	377740	772
2	309255	3921
3	213294	9528
4	103864	13324
5	44274	11163
6	16478	6158
7	5750	2719
8	1939	981
9	661	338
10	243	111
11	92	29
12	43	17
13	15	14
14	1	1

(b)NP 部分で集計

積む記号数	頻度	
	NP 数	文数
0	69820	4546
1	102337	7634
2	74126	16847
3	30025	11489
4	11432	5780
5	3336	2020
6	963	633
7	273	187
8	76	51
9	29	13
10	13	8

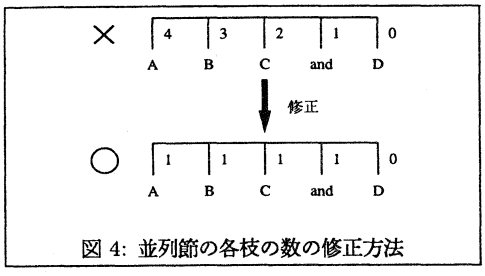

図 4: 並列節の各枝の数の修正方法

うに数値をふり直すことで余分に数えなくてすむようにして数えあげた. 表3(a) のように 7 ± 2 の上限 9 を越える文が多く存在することが気になる. そこで, 人間は日本語の文節のように個々の単語にまで分解せずに NP 程度のものはまとめて認識すると仮定して NP の部分におけるスタックの数を勘定した. つまり, S から NP にいたる経路に書いてある数字を足したあわせたものを用いて集計した. その結果を表3(b) に示す. この結果ならば Miller の 7 ± 2 の理論をほぼ満足する結果といってよいだろう.

5 おわりに

George A. Miller は人間の短期記憶の容量は 7 ± 2 程度のスロットしかないことを提唱している. 本研究では, 京大コーパスを用いて日本語文の各部分において係り先が未決定な文節の個数を数えあげ, その個数がおおよそ 7 ± 2 の上限 9 程度でおさえられていたことを報告した. また, 英語文でも同様な調査を行ない NP 程度のものをまとめて認識すると仮定した場合 7 ± 2 の上限 9 程度でおさえられていたことを確認した. これらのことは, 文理解における情報の認知単位 (チャンク) として日本語で文節, 英語では NP 程度のものを仮定すると, Miller の 7 ± 2 の理論と, 言語解析・生成において短期記憶するものは 7 ± 2 程度ですむという Yngve の主張を整合性よく説明できることを意味している.

謝辞

本研究の初期の段階において京大長尾真総長と議論した. また, 慶応大学三田メディアセンターの榎沢康子さんには文献検索において非常にお世話になった. また, 郵政省通信総研の藤原伸彦研究員には心理学の基礎的な事柄について教わった. ここに感謝する.

参考文献

[1] George A. Miller, The magical number seven, plus or minus two: Some limits on our capacity for processing information, *The Psychological Review*, (1956), pp. 81–97.

[2] 黒橋禎夫, 長尾真, 京都大学テキストコーパス・プロジェクト, 言語処理学会第 3 回年次大会, (1997), pp. 115–118.

[3] Victor H. Yngve, A Model and an Hypothesis for Language Structure, *the American Philosophical Society*, Vol. 104, (1960), pp. 444–466.

[4] 村田真樹, 内元清貴, 馬青, 井佐原均, 学習による文節まとめあげ — 決定木学習, 最大エントロピー法, 用例ベースによる手法と排反な規則を用いる新手法の比較 —, 情報処理学会 自然言語処理研究会 NL128-4, (1998).

[5] Richard L. Lewis, Interference in Short-Term Memory: The Magical Number Two (or Three) in Sentence Processing, *Psycholinguistic Research*, Vol. 25, (1996), pp. 93–115.

[6] John Kimball, Seven principles of surface structure parsing in natural language, *Cognition*, Vol. 2, (1973), pp. 15–47.

[7] 松岡正男, 文構造に着目した日本語文の理解しやすさ・しにくさの指標について, 京都大学工学部修士論文, (1996).

[8] 林武文, 乾敏郎, CG 空間におけるヒトの空間認知特性とそのモデル化に関する検討, 電子情報通信学会春季全国大会 D-183, (1991).

[9] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz, Building a Large Annotated Corpus of English: the Penn Treebank, *Computational Linguistics*, Vol. 19, No. 2, (1993), pp. 310–330.