文節係り受け木の根の構造について

高松　亮（埼玉大学経済学部）

Neighboring Structures at Root Node in Dependency Tree of Spoken Japanese

Ryo Takamatsu (Faculty of Economics, Saitama University)

１．はじめに

日本語の文節の係り受け関係を、文節をノードに，係り受け関係をエッジに対応付けたグラフとして表すと木構造になる．木構造の形態的特徴は文の構造を反映したものであり，例えば発話のジャンルが文の構造に与える影響を，木の構造を通じて観察することが可能である．前回の報告において、木の構造は「木の高さ」や「合計文節数」といった大域的特徴や，「各文節に係る文節数の平均」のような局所的特徴を用いて表すことができるため，観察を定量的に行なうことが可能であることを示した（高松(2013)）．その中で、木の大域的特徴の一つである木の高さの頻度について，ジャンルによる分布形の違いがあることが明らかになった．そこで本報告では，特にジャンル間の頻度差が大きかった高さの木について詳細な分析を行なう．その際，木の高さ以外の大域的特徴量として，木に含まれるノード数，すなわち文節数も考慮する．

具体的は，木の高さとノード数をパラメータとした場合の木の頻度分布を調べ，頻度がジャンルによって大きく異なるようなパラメータの値について，木の形態的特徴の傾向を明らかにし，そのような傾向がどのような言語現象に関連したものであるかを考察する．

木の構成要素の中で，根ノード（係り先を持たない文節，以下「根」という）は述語に相当する要素であり，根にかかる文節の数や種類が文の基本的な構造を決める特徴となるため，分析にあたってはこれらの特徴を重要な手がかりとして用いる．

なお，以下では文節の係り受けの情報を木構造として表現したものを係り受け木と呼ぶことがある．また，本報告では，係り元はあるが係り受けがない文節を根にもち，かつ高さ 1 以上(文節数 2 以上)の係り受け木を 1 本の木とする．

２．分析対象

国立国語研究所『日本語話し言葉コーパス(以下 CSJ)』(国立国語研究所(2006))のコア部分に含まれる学会講演と模擬講演を対象に，両者の比較を行なう．

学会講演は実際に行なわれた各種学会での講演を収録したもので，その中の比較的多数を占める理工系の学会では多くの話者が男性の大学院生である．発話のあらたまり度はやや高い．模擬講演は，年齢と性別のバランスをとった一般話者による、日常的話題(各話者に対して，3 種類のテーマから 1 つをあらかじめ指定)についての講演であり，話者の大部分が人材派遣会社からの派遣である．発話スタイルは学会講演よりもくだけたものである．

CSJ では係り受け構造の記述を行なう範囲として，文を認定するかわりに節単位という概念を導入し用いている．ほとんどの場合 1 本の係り受け木は 1 個の節単位に対応する．ただし，係り元があって，係り先のない文節が節単位中に複数存在する場合もあり，その場合にはそれぞれの文節を根に持つ複数の木を考えることにする．

なお，話者数および木の本数は，学会講演が 107 話者／8723 本，模擬講演が 70 話者／10046 本である．

３．木の大域的パラメータ：高さとノード数

木構造の大域的特徴を表わす特徴量の一つに木の高さがある．木の高さは，根から葉に至る経路をたどる際に通過するエッジの数（葉の高さ）のうち最大のものである．言い換えると，係り受け木の高さはある文において，文節が最大で何回の係り受け関係を経て述語に至るかに相当する．これは，国立国語研究所(1955)，小宮(1977)などにおいて「係り受けの次数」と
呼ばれている概念とほぼ同等である．既に高松(2013)において CSJ の学会講演と模擬講演，ならびに国立国語研究所(1955)における木の次数の比較を行なったが，これに小宮(1977)で報告された学校教科書(小中，高校，大学の 3 種類)での値を加えたものを図 1 に示す．なお，"対話"の値は国立国語研究所(1955)をもとに筆者が計算したものである．図より，話し言葉より書き言葉が，対話よりも独話が，あらたまり度の低い発話よりも高い発話が，より木が高い傾向があることが推測される．ただし，木の高さは文認定の基準の影響を受けるものであり，各数値を得た際の文認定の基準が必ずしも一致しているとはいえないため，これらの値を単純に比較することには注意を要する．

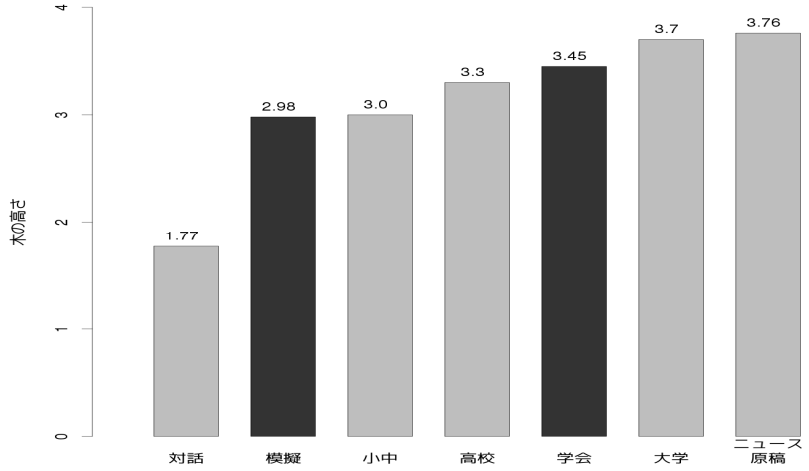

図１　発話ジャンルと木の高さ

大域的特徴として代表的な別の特徴量にノードの合計数(以下，単にノード数)がある．係り受け木の場合，ノード数はその文に含まれる文節数である．したがって，多くの文節を含む，長い文ほどその値は大きくなる．高松(2013)において，学会講演と模擬講演のノード数の相対頻度を比較し，文節数が 2 の木の頻度は模擬講演の頻度が高く，3 から 5 程度の範囲ではその差はわずかになり，それよりも文節数が多い領域においては，逆に学会講演の方がわずかに頻度が高いことから，木の高さ同様，ノード数にもジャンル依存性がみられることが明らかになっている．

以上のように，木の高さとノード数は木の大域的特徴を表わす代表的な量であり，いずれも発話ジャンルに依存して変化する性質を持つが，そもそも両者にはどのような関係があるだろうか．発話のジャンルを固定した場合には，ノード数が増加するにつれて木の高さも高くなることが予想されるが，その詳細は明らかではない．また，発話のジャンルが異なる木の集合同士を比較した場合に，どのような差異が見られるかも不明である．

そこで次節では，ある木の高さおよびノードの合計数を持つ木の相対頻度を求め，そのジャンル依存性を考察する．

４．ある高さとノード数を持つ木の頻度分布

学会講演および模擬講演について，ある高さとノード数を持つ係り受け木の頻度を図 2(学会講演)および図 3(模擬講演)に示す．なお図中では、木の高さに対応する平均ノード数を実線で、平均値に標準偏差を加算および減算した位置を点線でそれぞれ描いている．

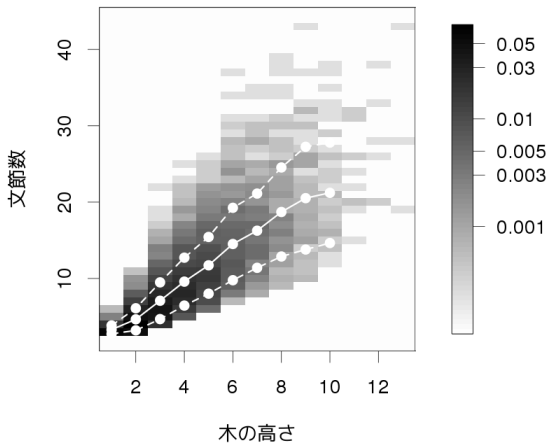

図２　木の相対頻度（学会講演）

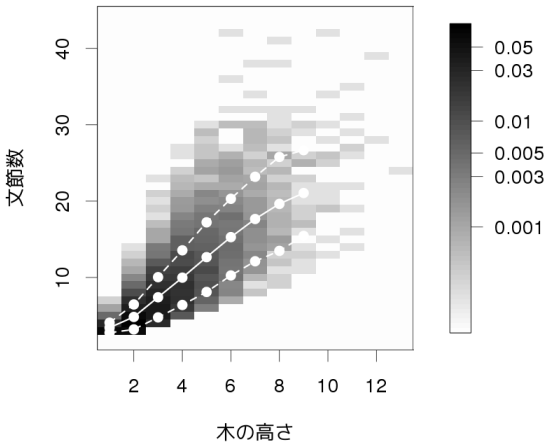

図３　木の相対頻度（模擬講演）

図 2 および図 3 から，木の高さが増加するとノード数も単調に増加すること，学会講演よりも模擬講演の方が増加率が若干大きく，したがって木の高さの割に相対的にノード数が多いこと，木の高さ，ノード数とも学会講演の方が模擬講演よりも広い領域に分布していることなどがわかる．

両者の頻度の差が最も明確に表われている領域を明らかにするために，学会講演の相対頻度から模擬講演の相対頻度を減算したものを図 4 に示す．

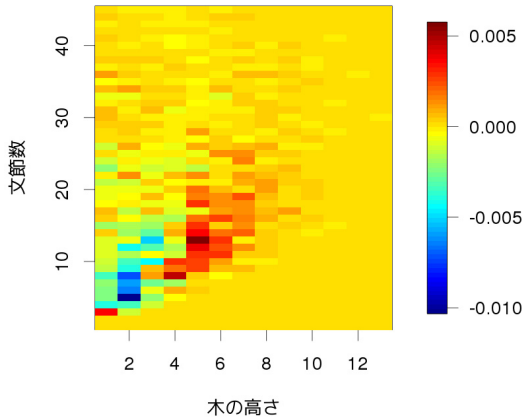

図４　木の相対頻度の差分（［学会講演］－［模擬講演］）

木の高さを h，ノード数を n とすると，図 4 において h=1 および h=2 の領域では，全体的に模擬講演の頻度が高い．特に h=2, n=5 の箇所では大きな頻度の差がある．h=3 および h=4 の領域では，ノード数が少ない部分では学会講演の方が，多い部分では模擬講演の方が，それぞれ頻度が高い．h が 5 以上の領域では全体的に学会講演が模擬講演より頻度が高い．

以上より，学会講演と模擬講演にはつぎのような相対的な差異があることがわかる．すなわち，学会講演は高さが高く，文節数が相対的に若干少ない文が多い．それに対して模擬講演は高さが低く文節数が若干多い文が多い．そこで以下では，図 4 においてそのような差異が最も明確な箇所の１つである h=2, n=5 に注目し，より詳細な調査を行なう．

５．根に係る文節数に基く分析

木の根は述語に相当する要素であり，根に何個のノード，すなわち文節が接続されているかは，文の構造と強い関連がある．そこで，以下では根にかかるノード数の相対頻度を h=2, n=5 において求め分析を試みる．ノード数毎の相対頻度を表 1 に示す。

表 1　根に係る文節数毎の頻度（h=2, n=5）

	学会講演(A)		模擬講演(S)	
根に係る文節数	頻度	相対頻度	頻度	相対頻度
1	17	0.0455	42	0.0766
2	182	0.487	210	0.383
3	175	0.468	296	0.540
合計	374		548	

表 1 より，h=2, n=5 であるような木の頻度は学会講演が 374，模擬講演が 548 である．両講演の全ての係り受け木に対する相対頻度を求めると，それぞれ 0.043(学会)および 0.055(模擬)となり，前節でも既に明らかなように模擬講演の方が頻度が高い．根に係るノード数毎の頻度に注目すると，いずれの講演でもノード数が 2 および 3 が頻度の大半を占めることがわかる．また，学会講演では 2 および 3 の頻度がほぼ同数であり，模擬講演では 2 よりも 3 の頻度がより大きい．さらに，模擬講演では 1 の頻度が学会講演よりも相対的に大きいこともわかる．このように，根に係るノード数の頻度分布が両講演で大きく異なることは，両者の係り受け木の形態の分布が大きく異なることを示唆している．

そこで，以下では根に係る文節数が 1 および 3 の場合について，それぞれどのように発話ジャンルに関係した性質を持っているかを観察する．

5.1　根に係る文節数が 1 の場合

5.1.1　模擬講演

根に係る文節がただ 1 つである場合，その文節の属性はどのようなものかを調査した．全 42 例のうち，言いよどみや述語の言い直しなどによる係り受けの乱れがある 2 例を除いた 40 例について，根に係る文節の最後を構成する短単位（国立国語研究所(2006)）の属性を調べると以下の通りであった．

- 節末（23 個，57.5%）
  - 引用節（格助詞「と(15 個)」，係助詞「は(1 個)」，副助詞「って(1 個)」）
  - タリ節，テ節，トイウ節，条件節タラ，理由節ノデ，並列節ガ（各 1 個ずつ合計 6 個）
- 節末でないもの（17 個，42.5%）
  - 助動詞（5 個），動詞（4 個），形容詞（1 個）の連体形（合計 10 個，25.0%）
  - 助詞（格助詞（3 個），副助詞（2 個））（合計 5 個，12.5%）
  - 助動詞の連用形（2 個，5.0%）

このように、模擬講演においては h=2, n=5 でかつ根にただ 1 つの文節が係るような場合には、その文節は節末、特に引用節であることが多い。引用節が根に係る 17 個の場合について、根の文節を調べるとその大半（15 例）が動詞「思う」によって構成されていた．

5.1.2　学会講演

根に 1 つの文節が係る例(18 例)のうち，言いよどみによる係り受けの乱れがある 1 例を除いた各文節の属性を以下に示す．

- 節末（14 個，82.4%）
  - 引用節（格助詞「と（10 個)」）
  - 「トイウ節」（動詞（連体形）「いう（3 個)」）
  - 「条件節レバ」（接続助詞「ば（1 個)」）
- 節ではないもの（3 個，17.6%）
  - 助詞（格助詞（1 個），副助詞（1 個））（合計 2 個，11.8%）
  - 助動詞の連体形（1 個，5.9%）

節が述語に係るケースが 18 例中 14 例（82.4%）と，模擬講演よりもさらに大きな割合を占めている．その中ではやはり引用節である例が 10 例と多く，これらが係る述語は動詞「思う」(6 例)，「言う」(2 例)，「考え（られ）る」(2 例）で構成される文節であった．

以上から，*h*=2, *n*=5 でかつ根に係る文節数が 1 の文は，述語に節、特に引用節が係る場合が多く，その場合に用いられる述語も限定されることがわかる．また、模擬講演の方が学会講演よりも頻度が高いが，述語に節が係る表現が占める割合は学会講演の方が多い．

5.2　根に係る文節数が 3 の場合

表 1 より，*h*=2，*n*=5 の文のうち模擬講演で最も頻度が高いのは，根に係る文節数が 3 のものであり，模擬講演に 296 本，学会講演に 175 本存在する．なお、*h*=2，*n*=5 の根付き木で根に接続するノード数が 3 であるものは 1 種類のみであり[1]，根に対して 2 つの葉が直接係ることが形態上の特徴である．

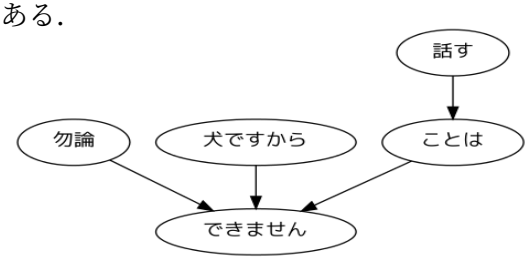

図５　*h*=2, *n*=5　かつ根に係る文節数が 3 の木の例

根に係る文節を構成する短単位のうち，最後のものの品詞の比率を延べ語数で求めたものを表 2，表 3 に示す．表 2 は，根に直接かかる全てのノードに関するもの，表 3 は根にかかる葉のみについてのものである．

表２　根に係る品詞の比率（全ノード）

品詞	高さ2　ノード数5(根に係る全ノード)				全体（根に係る全ノード）			
	模擬講演		学会講演		模擬講演		学会講演	
	個数	割合	個数	割合	個数	割合	個数	割合
助詞	599	0.675	362	0.693	16629	0.701	14602	0.730
格助詞	285	0.321	185	0.354	7607	0.321	8017	0.401
係助詞	132	0.149	90	0.172	2841	0.120	3183	0.159
その他	182	0.205	87	0.167	6181	0.261	3402	0.170
副詞	132	0.149	58	0.111	2059	0.087	1141	0.057
助動詞	58	0.065	29	0.056	2569	0.108	1943	0.097
名詞	52	0.059	39	0.075	992	0.042	865	0.043
形容詞	18	0.020	7	0.013	448	0.019	195	0.010
接尾辞	18	0.020	15	0.029	227	0.010	227	0.011
その他	11	0.012	12	0.023	798	0.034	1018	0.051
合計	888		522		23722		19991	

表３　根に係る品詞の比率（高さ 2，ノード数５の木）

品詞	高さ2　ノード数5(根に係る葉ノード)				全体（根に係る葉ノード）			
	模擬講演		学会講演		模擬講演		学会講演	
	個数	割合	個数	割合	個数	割合	個数	割合
助詞	358	0.605	220	0.632	6188	0.621	5082	0.670
格助詞	174	0.294	108	0.310	3048	0.306	2686	0.354
係助詞	92	0.155	66	0.190	1596	0.160	1611	0.213
その他	92	0.155	46	0.132	1544	0.155	785	0.104
副詞	129	0.218	57	0.164	2016	0.202	1126	0.149
助動詞	41	0.069	16	0.046	519	0.052	422	0.056
名詞	29	0.049	30	0.086	558	0.056	407	0.054
形容詞	16	0.027	6	0.017	292	0.029	99	0.013
接尾辞	12	0.020	11	0.032	133	0.013	116	0.015
その他	7	0.012	8	0.023	255	0.026	329	0.043
合計	592		348		9961		7581	

[1] ここでは同型な根つき順序なし木をもって「同じ種類の木」と考える．

表 3 より，h=2，n=5 における副詞の比率は，任意の（h, n）における副詞の比率にほぼ等しく，一般に模擬講演の方が学会講演よりも副詞の比率が高いことと，この型の係り受け木が副詞の比率に関しては平均的な性質を持っていることがわかる．

一方，表 2 における副詞の比率は学会講演，模擬講演のいずれにおいても h=2，n=5 の方が，任意の（h, n）よりも大きい．すなわち，葉以外のノードも含めた計数を行なうと，この型の木は副詞の割合が高くなる．これは，表 2 と表 3 の副詞の個数を比較すると明らかなように，ほとんどの場合副詞は葉として根に係ること，h=2，n=5 の木のうち本節で扱っている根に係る文節数が 3 のものは，根に係る文節 3 個のうち 2 個までが葉であることが要因となっている可能性がある．

これらの，根に係る葉に含まれる副詞について，どのような語彙が存在するかを調査したものを表 4 に示す．表は頻度の高い順に、第 10 位までの語が示されている。括弧内は頻度である。学会講演は少ない語彙が集中して出現し，模擬講演は様々な語彙が幅広く出現する傾向が見られる．それぞれの発話場面で好んで用いられる語彙の傾向が表れている．

表４　根に係る文節（副詞）

学会講演	まず(8)，ちょっと(5)，例えば(5)，一応(4)，色々(3)， こう(2)，ほぼ(2)，まだ(2)，やはり(2)，全て(2)
模擬講演	やっぱり(8)，あんまり(7)，よく(7)，ちょっと(6)，また(6)， もう(6)，やはり（6)，勿論(6)，色々(5)，大体(4)

６．まとめ

CSJ に含まれる学会講演と模擬講演の 2 つのジャンルを対象に，同じ高さ，文節数をもつ係り受け木の出現頻度がジャンルによって大きく異なる領域を明らかにした．特に大きな差異がある高さと文節数の組合せを対象に，木の根にかかる文節数ごとに，どのような差異があるかを調査した．一般に，2 つの木の高さと合計ノード数が等しいことは，両者が同型であるための必要条件である．今回はその条件に加えて木の根にかかる文節数が 1 または 3 という制約を課した分析を行なった．分析対象とした高さ 2，ノード数 5 の木においてはこの制約は同型であるための十分条件にもなっているため，結果として同型な木について，その出現頻度ならびに種々の言語学的特徴、さらにはジャンル依存性を調査したことになる．今後の課題としては，より広い範囲の高さ文節数を持つ係り受け木を分析することと，また木の同型性に基いた係り受け木の分類・分析手法の検討があげられる．

文　献

高松　亮(2013)「文節係り受け構造のジャンル依存性」第 3 回コーパス日本語学ワークショップ予稿集、pp.17-22
(http://www.ninjal.ac.jp/event/specialists/project-meeting/m-2012/jclws03/ よりダウンロード可能)

国立国語研究所(1955)『談話語の実態』，国立国語研究所研究報告 8
(http://db3.ninjal.ac.jp/publication_db/item.php?id=100170008 よりダウンロード可能)

国立国語研究所(2006)『日本語話し言葉コーパスの構築法』，国立国語研究所研究報告 124
(http://www.ninjal.ac.jp/csj/k-report-f/CSJ_rep.pdf よりダウンロード可能)

小宮千鶴子(1997)「読者層を異にする文章間に見られる文構造の相違(4)-係りの次数による日本史教科書の段階比較-」中央学院大学 人間・自然論叢，第 6 号

金 明哲(1993)「文節の係り受け距離の統計分析」社会情報：札幌学院大学社会情報学部紀要，5:2, pp.1-11 (http://hdl.handle.net/10742/754 よりダウンロード可能)