文節係り受け構造のジャンル依存性

高松　亮（埼玉大学経済学部）†

Genre Dependencies on Phrase to Phrase Modifications of Spoken Japanese

Ryo Takamatsu (Faculty of Economics, Saitama University)

１．はじめに

本報告は，文節間の係り受け関係の構造を木構造（以下，係り受け木と呼ぶ）としてとらえた場合，その形態的特徴が発話のジャンルによってどのように変化するかを定量的に記述・分析することを試みるものである．分析の対象としては，「日本語話し言葉コーパス」（以下 CSJ）の学会講演と模擬講演の発話を用いる．

２．発話のジャンルと係り受けの構造

発話のジャンルによって，そこで用いられる文体，スタイル，レジスタといった属性は異なる．これまでさまざまな視点からこの違いを定量的に捉える試みがなされてきた（樺島(1981)，Biber and Vasquez(2008)，小磯，小木曽，小椋他(2009)）．

本報告では，意味論的な構造を反映している最もプリミティブな要素であると考えられる文節間の係り受けの構造と，発話のジャンルとの関係に注目する．

文節間の係り受けの統計的性質については，係り受け関係を有する文節間の距離が拡張された Zipf の法則に従うことを指摘した例（丸山，荻野(1992)）がある．また，金(1996)は小説に関して，書き手が変わっても係り受けの距離の分布はほとんど変化がないことを示した．しかし，係り受け木の構造とジャンルの関係は調査されていない．

いま，文節をノード，係り元の文節と係り先の文節の関係をエッジと考えたグラフ構造で表わすと，一般的には係り受けの構造を木構造，すなわち係り受け木として表現できる[1]．

個々の係り受け関係は，文節間の修飾-被修飾や原因-結果のような，意味の呼応関係であるから，係り受け木は呼応関係の構造を表現したものである．学会における講演のように，複雑な論理的構造を持つ意味内容を，正確かつ明確に伝達することが必要な場面と，日常会話のような場面とでは，発話者が意味の呼応関係の構造を場面に応じて適応的に変化させている可能性がある．係り受け木の形態を定量的に表す特徴量を観測することができれば，そのようなジャンル毎の傾向の違いが観測値に表れることが期待できる．

３．係り受け木の定義

文節をノード，係り元の文節と係り先の文節の関係をエッジと考え，係り受け関係を木構造で表したものを係り受け木と呼ぶ．以下では，係り受け木の要素をグラフ理論の用語を用いて呼ぶことがある．各文節を「ノード」，係り元がなく，係り先がある文節を「葉」，係り元はあるが係り先のない文節を「根」と呼ぶ．係り受け関係のあるノード間を結ぶ線を「エッジ」，あるノード P から根 R に向かってエッジをたどる最短経路を考えるとき，経路上のノード Q に到達するまでに経たエッジの数を「P と Q の距離」，P から根 R までの距離を「ノード P の高さ」，ある木の葉から根までの高さの最大値を「木の高さ」という．

---

† rtakamat@mail.saitama-u.ac.jp

1 本報告ではグラフが木構造にならないような場合は扱わない．また，ノードの属性に文節の出現順序を加えた順序木も考えられるが，本報告では文節の出現順序の情報を捨象した係り受け木を用いた．

４．分析対象

本報告では，文節係り受けについて手作業によるアノテーションが施されている，CSJ のコア部分について「学会講演」と「模擬講演」の 2 つの発話場面の比較を行なった（表 1）．これは，合計 6 名の話者(以下，共通話者と呼ぶ)が両方の発話場面に収録されており，それらのデータを用いれば，同一発話者の発話場面による差異をも比較可能なためである．

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用いている(国立国語研究所 2006)．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．

表 1 ：分析対象の種類と規模（括弧内は共通話者 6 名についての値）

	話者数	節単位総数	木の総本数
学会講演	70(6)	8516(790)	8723(794)
模擬講演	107(6)	9675(613)	10046(640)

５．特徴量とその傾向

５.１　はじめに

学会講演と模擬講演のデータは，年齢や性別といった話者の属性の分布が同一ではない(国立国語研究所(2006))ため，両者の統計的な性質を単純に比較するべきではないが，両者に共通の話者(共通話者)が 6 名おり，共通話者の場合と全話者の場合それぞれについて比較することで，特徴量の異同の原因がジャンルなのか母集団の違いなのかをある程度判断できる．以下では，係り受け木の形態的特徴を表現する特徴量として，木の高さのような大域的な特徴と，ある文節に対して係る文節の個数やその平均値のような局所的な特徴について分析する．なお，係り受け木のうち，係り元，係り先の両方が存在しない 1 個の文節のみからなる木は，その多くがフィラーなどであるため，分析対象から除外している．

５.２　大域的特徴

５.２.１　木の高さの頻度

係り受け木の高さの相対頻度の分布を図 1 および図 2 に示す．

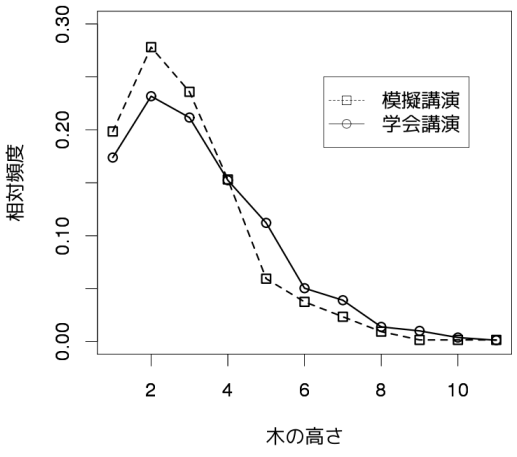

図 1 木の高さの頻度(全話者)

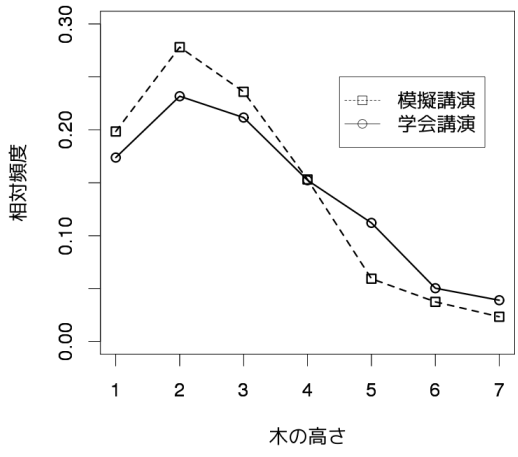

図 2 木の高さの頻度(共通話者)

図 1, 図 2 のいずれも学会講演の方が模擬講演よりも分布の幅が狭く，相対的に高い山の度数が多い．両者に共通する特徴としては，学会講演および模擬講演とも木の高さが 2 で最大値の頻度となり，それよりも木の高さが高くなるにしたがって頻度が単調に減少することがあげられる．学会講演の木の高さの平均値は 3.45(全話者)および 3.27(共通話者)，模擬講演の木の高さの平均値は 2.98(全話者)および 2.88(共通話者)であり，学会講演が模擬講演よりも木の高さの平均値が大きい．

また，全話者と共通話者の双方で同様の傾向を示すことから，学会講演と模擬講演による分布形の差異は，母集団の属性の偏りというよりは，ジャンルに起因する違いであることが推察される．

国立国語研究所(1955)においては「係り受けの次数」という，係り受け木の高さと同等のパラメータを用いて文の構造を分析しており，ニュース音声と日常的な場面における対話音声それぞれに表れる文の次数を比較した結果，ニュース音声(平均値 3.76)が対話音声(平均値 1.77)よりも次数の高い文が頻出すると指摘している(平均値は筆者による再計算)．参考のために本報告における値も含め，木の高さの値の順に並べると，

ニュース音声 > 学会講演 > 模擬講演 > 日常対話

となる．

ニュース音声は独話で，かつ改まり度が高く，本報告における学会講演に近い性質を持っている．また，本報告における模擬講演は比較的くだけた状況における独話であり，日常の対話とニュースや学会講演の中間的な性質を有すると考えられ，このことが木の高さの平均値の大小にも表れているものと考えられる．

５.２.２　文節数の頻度

1 本の係り受け木に含まれる文節の数は，木の規模の大小を表現するパラメターの一つである．図 3 および図 4 に文節数の相対頻度の分布を示す．図より，木の高さの頻度の場合と同様に，共通話者の場合も，話者全体の場合もかなり類似した傾向があることがわかる．すなわち，いずれの場合も文節数 2(図の最も左側のプロット)の頻度が例外的に高く以降単調に減少すること，文節数が 2 において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がわかる．

これらの特徴が図 3 と図 4 に共通して見られることから，文節数の頻度分布の傾向も，学会講演と模擬講演というジャンルの違いから生じていることが推察される．

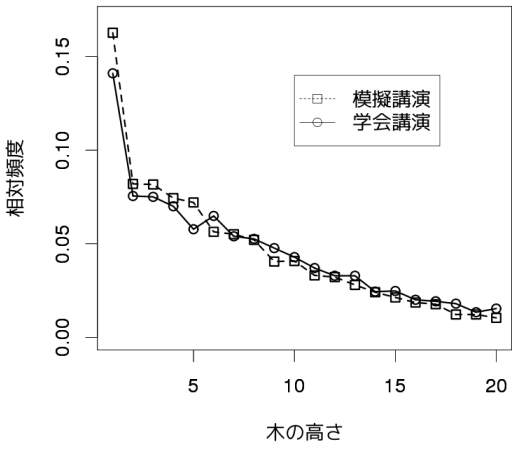

図 3 木に含まれる文節数の頻度(全話者)

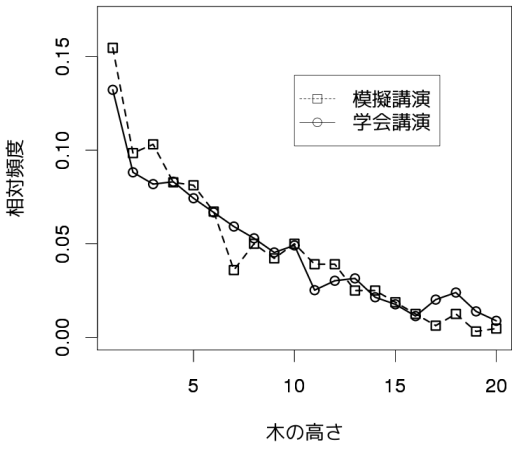

図 4 木に含まれる文節数の頻度(共通話者)

５.３　局所的特徴

５.３.１　係り元の文節数

係り受け木の局所的な特徴のうちもっとも基本的なものとして，ある文節に注目した場合に，その文節に係る文節(係り元)の数が n 個である場合の頻度を考える．図 5 および図 6 に係り元の数の相対頻度の分布を示す．なお，縦軸は相対頻度の常用対数である．

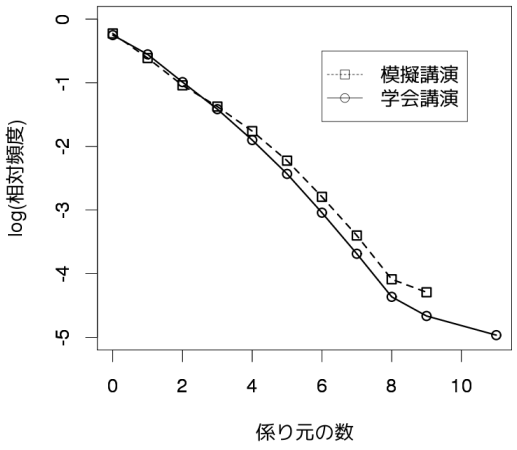

図 5 係り元の文節数の頻度(全話者)

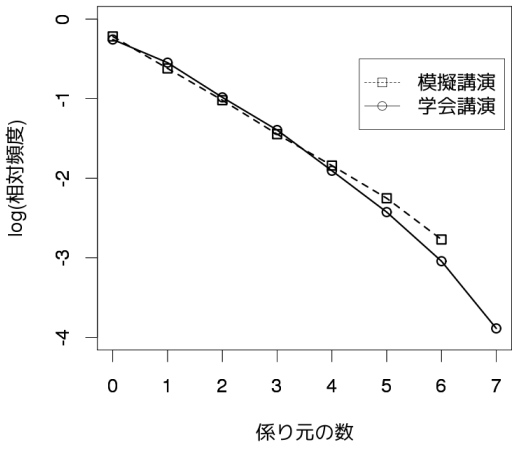

図 6 係り元の文節数の頻度(共通話者)

いずれのグラフもプロットが傾きがほぼ負の直線上にのっていること，係り元の数が 0，すなわち文節が葉である場合の相対頻度が学会講演と模擬講演とで一致すること，係り元の数が 0〜3 ないし 4 個の領域では学会講演が，それ以上の領域では模擬講演が，それぞれわずかずつ頻度が高い．共通話者と全体話者で傾向が一致することから，学会講演と模擬講演の間に見られたわずかな差異が，スタイルの差異から生じたものである可能性がある．

５.３.２　根の文節に係る文節数

根に相当する文節に，n 個の文節が係る場合の相対頻度を図 7 および図 8 に示す．

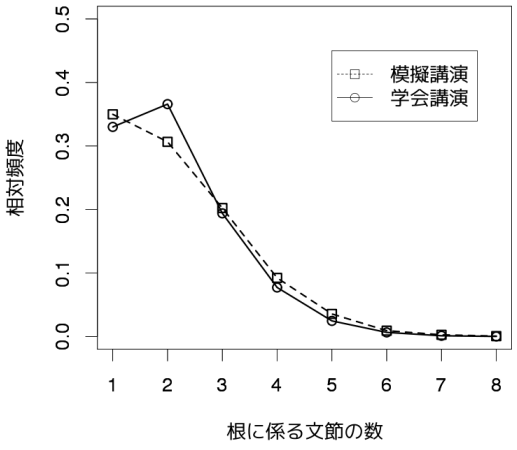

図 7 根の文節に係る文節数の頻度(全話者)

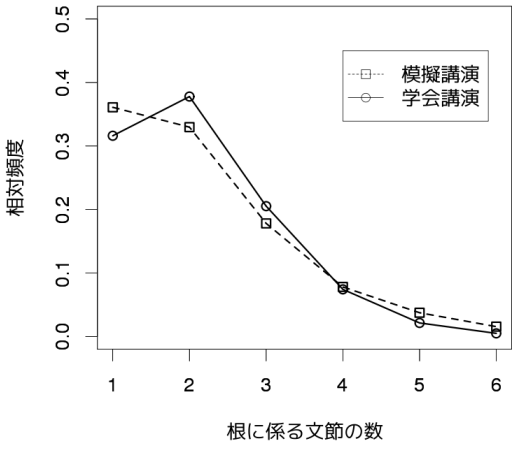

図 8 根の文節に係る文節数の頻度(共通話者)

学会講演は文節数 2 において最大値を，模擬講演は文節数 1 において最大値を取る．また，学会講演の方が分布の幅が相対的に狭い．これらの傾向が両方の図において見られることから，以上の差異が学会講演と模擬講演のスタイルの違いから生じている可能性がある．

５.３.３　葉の高さと葉の累計係り元数

ある葉の高さが n であるとき，葉から根まで辿って行く際に通過する各文節 N_i (i=1,2,..,n) が係り元を d_i 個ずつ持っているなら，d_i の合計数をその葉の累計係り元数と呼ぶことにする．葉の高さと累計係り元数の平均値の関係を図 9 および図 10 に示す．

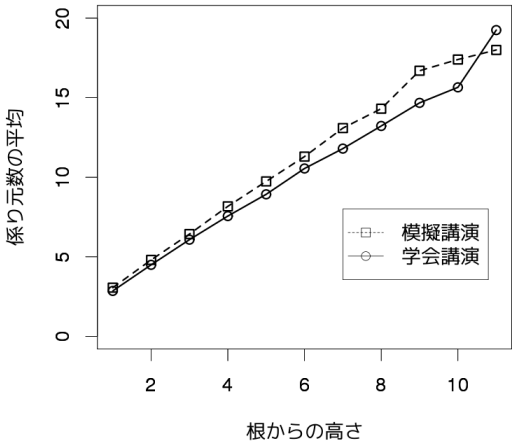

図 9 累計係り元数の平均値(全話者)

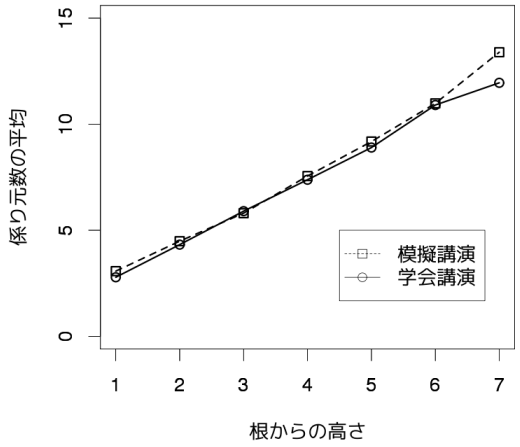

図 10 累計係り元数の平均値(共通話者)

全てに共通する特徴として，葉の高さが 1 から 6 ないし 7 程度までの範囲においては，プロットが傾きが正の直線上に良くのっていることが挙げられる．全話者においてはこの直線の傾きが模擬講演と学会講演とで異なり，学会講演の方が傾きが若干小さく，葉の高さが高くなった場合の係り元数の増加が少ない．一方，共通話者においては，学会講演の方が傾きが小さい点は全話者と同じではあるが，その差はごくわずかである．したがって，傾きの差異が発話ジャンルに起因している可能性はあるが，話者によってはそれほど明確な差が生じないことがあることがわかる．

６．まとめと今後の課題

今回得られた知見のうち，ジャンルによる量的な差異に関するものをまとめると次のようになる(A は学会講演，S は模擬講演を指す)．

- 大域的特徴
  - 木の高さの分布の平均：$A > S$
  - 木の高さの分布の幅： $A < S$
  - 文節数 2 の木の相対頻度： $S > A$
- 局所的特徴
  - 葉の相対頻度：$A=S$
  - 根に係る文節数の分布の最頻値：$A=2, S=1$
  - 根に係る文節数の分布の幅：$A < S$
  - 高さ n の葉から根までの累積係り元数：n に比例して増加(比例定数は $A < S$)

学会講演は木の高さが高く，高さの分布の散らばりも小さいこと，模擬講演は文節数が 2(すなわち高さで言えば 1)の木の相対頻度が相対的に多いことがわかる．また，高さ n の

葉から根までの累積係り元数は n にほぼ比例するが，学会講演の方が比例定数が小さいことから，葉が高い位置にあっても，根からその葉までの経路での枝分れがより少ない．以上より，学会講演は木の高さが高いが，枝分れの少ない構造を持つ傾向があると言える．

係り元を多く持つ文節ほど頻度が急速に減るが，模擬講演の方がよりロングテールな傾向を持つことから，模擬講演には 1 つの文節に多数の文節が係る表現が相対的に多いことがわかる．そのような構造の例を図 11 に示す．

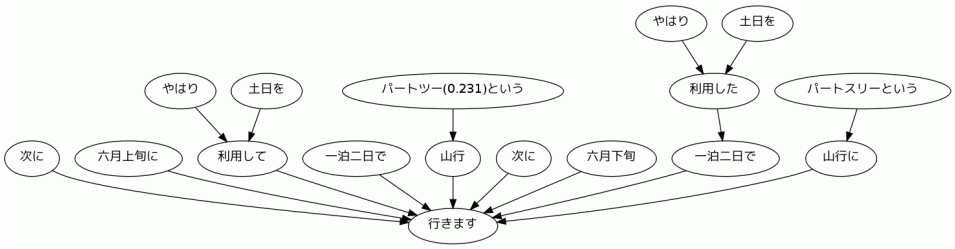

図 11 1 つの文節に多数の文節が係る構造の例

本報告では係り受け木の形態を表す特徴量として，大域的なものと局所的なものとを定義し，それらが学会講演ならびに模擬講演というスタイルの違いによってどのような傾向を持つのかについて調査し，いくつかの知見を得た．

今後の課題としてはまず，より多くの発話ジャンルについての調査を行なう必要がある．また，文節の順序関係についての情報を考慮に入れた場合，どのような傾向が見出されるかについても検討の必要がある．さらに，このような差異の傾向が見られる原因について，発話の生成過程についての知見との接続を行なう必要がある．

謝　辞

本報告でなされた研究は著者が国立国語研究所に外来研究員として滞在している際になされたものです．前川喜久雄先生，小磯花絵先生をはじめ多くの方々の御助力を頂きました．記して感謝を表します．

文　献

樺島忠夫(1981)『日本語はどう変わるか』岩波新書，岩波書店

Biber, Douglas and Cammila Vasquez (2008) "Writing and Speaking", in Handbook of research on writing, ed. C. Bazerman , pp.535--548, Routledge, Oxford, 2007

小磯花絵，小木曽智信，小椋秀樹，他(2009)「コーパスに基づく多様なジャンルの文体比較 -短単位情報に着目して-」言語処理学会 第 15 回年次大会発表論文集，pp.594-597

丸山 宏，荻野 紫穂 (1992)「日本語における文節間係り受け関係の統計的性質」情報処理学会 全国大会講演論文集，45:3, pp173-174. (http://ci.nii.ac.jp/naid/110002889591 よりダウンロード可能)

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)

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)