日本語話し言葉コーパスを用いた被験者実験による日本語予測文法の研究

高梨 克也　　河原 達也

京都大学 学術情報メディアセンター
{takanasi,kawahara}@ar.media.kyoto-u.ac.jp

1．背景と目的

話し言葉の実時間処理のためには文構造を前から解析できることが望ましいが, SVO 言語である日本語では通常述語は文末に来るため，文末を待たなければ文構造が決定できないという問題点がある．しかし，実際には，聞き手は進行中の文の途中の時点において「すでに言われたこと」に基づいて「これから言われるであろうこと」をある程度予測できているのではないかと考えられる．寺村(1987)は，文を途中の時点で切断し，後続部分を予測させ，その結果を分析するという実験を行い，「考えようによっては驚くほどの正確さで先を予測」ができると主張している．寺村の提案を受け，その後，陳述副詞や主題，格要素などについての「予測文法」的研究が行われているが（平田 1997 など），それぞれの研究は予め特定の言語特徴に焦点を限定していると共に実験の規模も小さく，こうした知見が体系的に整理されてきたとはいいがたい．そこで，本稿ではこの方向性を踏襲し,『日本語話し言葉コーパス』を素材とした被験者実験を行い，主に係り受け構造の観点から分析することによって, 予測において用いられる要因を解明することを目標とする．

2．係り受け関係

ある切断時点の以前の部分から以降の部分への予測に用いられるのは先行部中の要素から後続部中の要素への何らかの「関係」であると考えることができる．こうした関係の中には，述部の中での述語末要素の承接関係，呼応関係，節間関係などのさまざまなレベルのものがあるが，本稿では節内での係り受け関係に焦点を限定する．節内の主要な係り受け関係である格関係は，述語がその項としてどのような格要素を要求するかという観点から認定されるものであるため，一般的には「後ろから前へ」の関係であると考えられるが，実際には，格要素の生起から特定の（種類の）述語の生起が予測可能になっていることは多いと考えられる[1]．この点を本稿での分析の焦点とする．

3．実験

3.1 実験素材

『日本語話し言葉コーパス』（CSJ）[2]のコアのうち，談話境界情報が付与されている模擬講演 25 講演の中から「節単位」（高梨他 2004）を単位として選択した．25 講演中の全 2257 節単位のうち，以下の基準に合致した 532 単位からさらに無作為に 100 単位を選択して実験刺激とした．

・節単位末尾が[絶対境界][3] ．
・非流暢性に関する人手修正箇所 を含まない．
・節単位途中の節ラベルが<弱境界>のみで, その数が 0～3 個.
・フィラーF，語断片 D，雑音などのみからなる文節と B+の文節[4]，接続詞文節を除外した「有効文節」数が 5～20 文節.

3.2 手順

被験者は大学生・大学院生 5 名である．有効文節の直後の境界を「切断時点」とし，切断時点を一つずつ後方に移動させながら，すべての切断時点において，以下の教示に従って後続部分についての予測を記入させた.

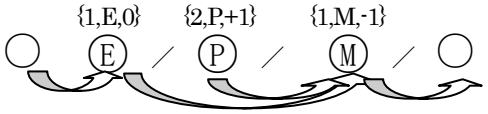
・表示された「現時点」の「先行文脈」の内容に基づき，その時点から文末までの内容を予測し，記入する.
・文末形式は丁寧体とし, 句読点やその他の記号は使用しない.
・答えは短めに書けばよい．無理に長くする必要はない.
・文末までではなく，予測できるのが文の途中の要素だけの場合にはその要素だけを記入してもよい.
・正解かどうかということにこだわる必要はなく，最初に思いついた内容を書けばよい.
・思いつくことがない場合には「null」と入力する.

4．係り先未定文節数の変化

ある文節の直後の予測時点においていくつの文節の係り先がまだ生起していないかを「係り先未定文節数」と呼ぶ．当該文節の係り先となる文節はまだ生起していないため，係り先未定文節数は必ず 1 以上である．また，すべての文節が直後に係る場合には，係り先未定文節数はどの予測時点でも 1 となる．従って，係り先未定文節数が 2 以上の箇所では，自分自身だけでなくそれ以前の一つ以上の文節が当該文節を越えて後続文節に係っていることになる.

次に，係り先未定文節数が直前の予測時点からどのように変化したかを「係り受け未定文節数の増減」（MEP）と呼ぶ．直前の時点と比較し，当該の時点での係り先未定文節数が増加する（P）のは当該文節が先行する文節からの係り先となっていない場合で，P の値は必ず 1 である．変化しない（E）のは当該文節が先行する一つの文節からの係り先となっている場合である．減少する（M）のは当該文節が先行する二つ以上の文節からの係り先となっている場合であり，P とは異なり，その値は 1 以上となることもある.

{係り受け未定文節数,増減,その値} =



MEP 及び節単位末毎に文節種[5]をカウントしたところ，節単位末については，助動詞の終止形が 65%，終助詞が 27%である

---

[1] 寺村(1987)も「日本語は『述語が最後に来るから，最後まで聴かなければ分からない』といった『説』がいかに事実から離れたものか」と指摘している.

[2] 本稿で参照した各種マニュアルについては本稿の末尾にリストを付す. CSJ の概要と入手方法については以下の URL を参照．各マニュアルもダウンロードできる．http://www2.kokken.go.jp/%7Ecsj/public/members_only/releaseinfo/index.htm

[3] CSJ の各節単位の末尾は必ずしも常にいわゆる文末形式とはなっていないが，[絶対境界]とは文末形式のことである.

[4] 当該の 2 文節は形態論的には 1 文節となるべきだったが，間に休止が入ったため 2 つの単位に分けられたものであり，係り受け関係付与の際にはこれらの箇所をつなげた上で関係を付与した（内元他 2004）.

[5] 文節末の形態素（長単位）の品詞情報に基づいて認定した.

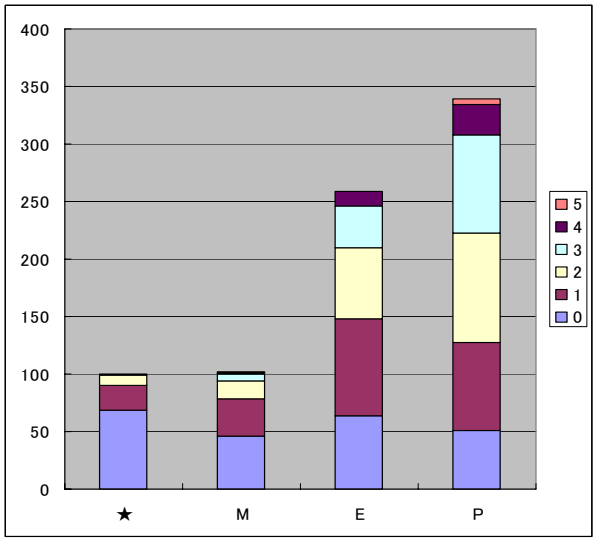
[表1] 正解得点及び述語文節の種類と被係り受け数

正解得点 述語分類	100≦X＜150 計	 平均	50≦X＜100 計	 平均	0≦X＜50 計	 平均	-50≦X＜0 計	 平均	合計 計	 平均
思考	6	1.33	8	1.63	2	1.5	0	-	16	1.50
存在	4	2.25	6	2.33	2	3	0	-	12	2.42
抽象	2	1	2	4	1	3	0	-	5	2.60
変化	2	2.5	1	1	0	-	0	-	3	2.00
具体	7	3.86	12	2.75	8	2.88	4	2.75	31	3.03
状態	0	-	2	3	3	2.33	2	2.5	7	2.57
名詞	0	-	0	-	5	2.2	2	2.5	7	2.29
抽象名詞	1	1	0	-	1	1	2	1.5	4	1.25
計	22	2.36	31	2.42	22	2.45	10	2.4	85	2.41

のに対して，動詞や形容詞の終止形は一例しかなかったため，節単位末の文節は通常述語を含むものの，述語単独で構成されることは極めてまれであるといえる．次に，M 文節に特徴的なのは助動詞の連体形（17.7%）と接続助詞（16.7%）であり，これらは述語を含む文節であると考えられ，従って他の文節からの係り先になる可能性の高い文節であるといえる．

5．予測結果の全般的傾向

使用した 100 節単位について，有効文節数の合計は 1037 個だが，最終文節直後は実験対象箇所ではないため，提示箇所の合計は 937 箇所である．被験者は 5 名であるため，被験者による予測が記入されうるスロット数は 4685 であったが，うち，被験者が何らかの予測内容を記入した箇所が 3466 箇所，null だった箇所が 1219 箇所で，null 率は 26.02%であった．



[図1] 係り先未定文節数毎の null の人数

MEP が「M」「E」「P」となる文節と最終文節（★）の直前の予測時点における null 人数を観察した結果を[図1]に示す．まず，節単位末の文節の直前においては null 人数が顕著に低く，被験者は最終文節の直前の直前の時点ではかなりの割合で最終文節についての何らかの予測を行うことが可能になっている．また，節単位末の文節の直前の予測時点と同様の傾向が M 文節の直前でも見られ，この傾向は E 文節や P 文節の直前とは明らかに異なる．第4節で見たように，最終文節や M 文節は述語を含む文節である可能性が高いため，これらの時点では，既に生起した文成分から後方（直後）に来る述語要素についての何らかの予測が形成されやすいという大まかな傾向を確認できる．

6．最終文節の直前での予測

6.1. 予測結果の採点

一般に，被験者の予測内容を話し手が実際に発話した内容（以下「正解」）や他の被験者の予測内容と比較するのは容易でない．特に，表層形式の一致だけでなく意味内容上の類似性も考慮しなければならないという点と，被験者の予測内容を正解と対応づける際に，意味内容を考慮しながら形態素，文節，連語などの単位を適宜使い分けなければならないという点が重要である．

そこで，今回はまず最終文節の直前のみを対象として被験者の予測結果を採点する方法を考える．こうした箇所では，「正解」の範囲は 1 文節のみであり，また被験者回答の多くも 1 文節であることが多いため，判定は比較的容易である．対象としたのは「最終文節の直前」and「null の人数が 1 以下」and「当該文節と直後の文節との関係が B+でない」の 85 箇所である．

本稿の分析の焦点は述語を含む文節以前の時点までに生起した内容からこの述語文節が予測できるか否かという点にあるため，予測結果の採点においては，被験者回答中の要素を可能な限り「正解」中の述語に対応させるようにした．また，日本語の述部文節は述語のみから構成されることはまれで，述語にさまざまな助動詞や助詞が後接することによって形成されているため（第4節），これらの述語後置要素の一致も適宜考慮した．基準の詳細は省略するが，要点は以下の通りである．

1．述語の一致を重視：述語部分についての被験者の予測が「正解」と一致した場合に各 20 点とし，全員が一致した場合に 100 点となる．
2．述語の形式的な一致だけでなく意味的な同義性や類義性を考慮：意味的な同義性や類義性の度合いを判断し，15 点，10 点，5 点の部分点を与えた．
3．予測の不一致が重要な問題となる場合を考慮：直後が最終文節であることが予測できていない場合とみとめ方が逆の場合には減点した．
4．述部内の述語以外の要素の予測も考慮：述語末のモダリティ要素（益岡 1991）などについては，無標のゼロ形態である場合と有標の形態である場合の区別があるが，被験者が有標の形態を予測できていた場合には 5 点加点した．そのため，正解得点が 100 点を越える箇所も生じる．
5．何も予測できなかったという事実を考慮：回答が null の場合，「述語」の項目において-10 点とする．

正解得点の範囲は 125 点～-40 点の間で，平均点は 58.1 だった．50 点間隔で対象箇所をグループ化し，これらのグループ毎に述語文節の種類をカウントし，さらに当該文節がいくつの文節からの係り先となっているかの平均を各分類項目毎に求めた結果を[表1]に示す．

[例1] [具体, 120点, 被係り受け数=4]

文節	文節	係り先	A	B	C	D	E
んで	0						
私は	1	12					
就職を	3	7	あきらめて	やめることにしました	あきらめて	断念して好きなことをやろうと決めました	Null
その	4	5					
時	5	6					
三年生だったので	6	7					
控えて	7	12	いまして	いました	いました	いる時期だったわけですけど	Null
五月の	8	9					
頃から	9	12	準備を	就職活動に関する資料を集めはじめていました	就職活動を	準備してました	Null
公務員の	10	11	試験を	試験勉強をしていました	試験を	試験勉強を始めたわけです	試験の準備を始めました
→ダブルスクールに	11	12	通って	通っていました	通っていた	通うことにしました	通い始めました
通っていました	12		▲	▲	▲	▲	▲

[例2] [具体, 95点, 被係り受け数=3]

文節	文節	係り先	A	B	C	D	E
僕はですね	2	6					
這いつくばるように	3	4					
⇒して	4	6	Null	そこにいました	何とか	地面に落ちたコンタクトを探しました	Null
→風呂から	5	6	あがって	出てきました	出ました	上がりました	出ました
上がりました	6		▲	▲	▲	▲	▲

[例3] [具体, 95点, 被係り受け数=2]

文節	文節	係り先	A	B	C	D	E
廊下を	1	2					
挟んだ	2	3					
向かいの	3	6					
三つ	4	5					
先ぐらいの	5	6					
部屋の	6	7	灯りが	中で	ドアに	ドアが薄く開いていたんです	Null
⇒目覚まし時計で	7	9	目を覚まして	大きい音のものがあります	起された	朝早くに起こされてしまいました	目が覚めました
→目が	8	9	覚めて	覚めました	覚めた	覚めてしまったんです	覚めました
覚めてしまうという(D よ)	9		▲	▲	▲	▲	▲

○述語文節の分類

A) 思考 :「思う」「感じる」「考える」など (話し手が主語)
B) 存在 :「ある」「いる」「ない」
C) 抽象 :「する」
D) 変化 :「なる」
E) 具体 : 具体的行為を表す述語
F) 状態 : 状態・性質を表す述語
G) 名詞 : 名詞述語文
H) 抽象名詞 :「ことだ」「感じだ」「次第だ」「状態だ」

まず, 正解得点と述語文節の種類の関係について, 述語文節の種類が「思考」「存在」「抽象」の場合に正解率が高く,「状態」「名詞」「抽象名詞」で低いことが分かる.「存在」では, 経験を表す「ことが//ある」や,「記憶が//ある」「思い出が//ある」などの抽象的な存在を表すものの方が多かった(//は予測時点).「抽象」の5例の内訳は「気が//する」「生活を//する」「〜たり//する」であり,「変化」の3例も「ことに//なる」「感じに//なる」「勉強に//なる」である. これらの機能動詞を含む箇所は内容的には1文節と見なした方がよいかもしれない[6].

次に, 被係り受け数について, 予測の正解率と被係り受け数の間に単純な相関はなさそうであるが, 述語文節の種類と被係り受け数の間には,「具体」で被係り受け数が多く,「思考」で少ないという傾向が観察される. 特に,「思考」では正解率が高いにもかかわらず被係り受け数が少ないという点が興味深い (6.3節). また,「具体」には被係り受け数が多いほど正解率が高いという傾向があるのかもしれない (6.2節).

6.2 具体述語の予測

最終文節が具体述語のもののうち, まず, [例1][7]は正解率が高い理由が被係り受け数の多さにあると思われるものである.

[6] 現行の文節認定基準では,「勉強する」は1文節なのに対し,「勉強をする」は「勉強を」と「する」の2文節に分けられる. こうした機能動詞の扱いについては村木(1991)が詳細な分析を行っている. こうした表現の位置づけを考える際には, 継起的 syntagmatic な関係だけでなく, 共起的 paradigmatic な関係を同時に考慮しなければならないという指摘は本実験結果の採点において被験者予測のある部分を正解中のどの部分と対応づけるべきかを判断する際にも踏まえるべき視点である.

[7] 例の見方 :「文節」の冒頭に「→」がある箇所が分析対象箇所である. A〜E は被験者の予測内容で, その中の「▲」は予測対象外の最終文節直後であることを示す. 係り受け関係について, 例えば,「係り先」が「12」ならば, その文節の係り先は「文節」が「12」の文節である.

［表2］ M 文節＋最終文節の例

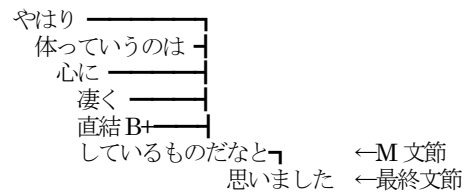
M 文節	被係り受け数	最終文節	被係り受け数	述語分類	得点
話だと	2	思っておりました	2	思考	110
話したいと	3	思います	1	思考	100
あるのではないかと	2	思ってます	1	思考	100
しようと	3	思いました	3	思考	80
ないと	2	思っています	1	思考	50
上手だったっていう	2	ことですよね	1	抽象名詞	110
置いてあるような	3	感じでした	1	抽象名詞	20
帰れる	2	状態ですね	2	抽象名詞	-25
残ってたような	4	時代でした	2	名詞	15
感じの	3	学校でしたね	1	名詞	10
できた	2	学校だったんですね	2	名詞	35
遊んだりとか	3	盛りだくさんだったんですね	2	状態	0

最終文節「通っていた」の内容は当該最終文節の直前ではじめて予測可能になっている[8]. 複数の項要素の組合せからこれらの共通の係り先となる述語が絞り込まれているのだと考えられる.

逆に，次の２例は必ずしも被係り受け数が多いとはいえないが正解率が高い例である．[例２]の場合には「風呂から」から「上がる」が，[例３]では「目が」から「覚める」が予測可能になっており，当該の２文節間の局所的な意味的結びつきの強さが予測の要因であるといえそうである．このように，ある名詞ないし格要素が慣習的に特定の述語と結びついて形成される連語は予測のための強力なリソースになっていると考えられる. しかし，これらの２例にも相違がある．[例２]では，「風呂から」の直前の時点(⇒)での予測は全く当たっていないのに対して，[例３]では，すでに「目が」の直前の時点（⇒）においてかなり正確な予測が形成されているためである. 後者の場合には「目覚まし時計（で）」という特定の表現が予測を限定する要因として重要であることが分かる.

6.3. M 文節＋最終文節

最終文節の直前がＭ文節という箇所が 14 箇所あった．最終文節直前の null 人数が２以上の２例を除く 12 例のうち，５例が「引用節＋思考動詞文節」，３例が「連体節＋抽象名詞文節」（いわゆる「外の関係」の連体修飾）である[表２]．これらの８例においては, 最後から二番目のＭ文節が埋め込み節の最終文節で，節単位の最終文節はこの埋め込み節を補部として持つ特定の種類の述語という構造になっている. 6.1 節で見たように，思考動詞述語は被係り受け数が小さいにもかかわらず予測の正解率が高いという傾向があるが，これらの箇所ではその直前のＭ 文節が多くの先行文節からの係り先となる被係り受け数の多いものであり，このＭ文節（だけ）が直後の最終文節に係るという構造になっているため，これらの２文節をまとめて考えれば被係り受け数が多い箇所であるといえる.



[8] それ以前の時点では被験者の予測は二転三転しているが，被験者間では一致しているという点も興味深い.

7．今後の課題

被験者の予測結果の収集自体は比較的容易である．問題は，収拾された予測結果の正解や一致をどのような方法で確実かつ効率的に判定し，意味ある発見を導くか，という点にある．こうした判定はある程度自動化できるかもしれないが，被験者の回答の特徴は「どの回答にも一理ある」という点であるため，徒に正解や一致のみを重視することには問題がある.

確率モデルを構築する際，例えば形態素ではなく文節を単位とする場合のように，より大きな言語単位を要素とする場合ほど異なり語彙数が多くなるため，モデルの学習は困難になる. コーパスに基づく確率的な予測とは異なる傾向が被験者の予測に見られるならば興味深い.

最後に，CSJ を刺激として使用したにもかかわらず，今回は刺激提示時に音声情報は用いておらず，また分析でも休止や話速，基本周波数（F0）といった韻律情報は利用していないが，この点も今後の課題としたい.

謝辞

木田敦子氏，大塚裕子氏（計量計画研究所），丸山岳彦氏（国立国語研究所），竹内和広氏，黒田航氏（情報通信研究機構）と被験者の皆様に感謝いたします．本研究の一部は，文部科学省リーディングプロジェクト e-society 基盤ソフトウエア「ユーザ負担のない話者・環境適応性を実現する自然な音声対話処理技術」の支援により行われた.

参考文献

平田悦朗(1997)『日本語学習者の文の予測能力に関する研究及び読解力・聴解力向上のための教材開発』（平成８年度文部省科学研究費補助金基盤研究(B)(2)研究成果報告書）

益岡隆志(1991)『モダリティの文法』（くろしお出版）

村木新次郎(1991)『日本語動詞の諸相』（ひつじ書房）

寺村秀夫(1987)「聴き取りにおける予測能力と文法的知識」『日本語学』6(3)（再掲『寺村秀夫論文集 II：言語学・日本語教育編』くろしお出版，1992．97-114）

『日本語話し言葉コーパス』マニュアル

西川賢哉・小椋秀樹・相馬さつき・小磯花絵・間淵洋子・土屋菜穂子・斉藤美紀 (2004)文節の仕様について.

小椋秀樹・山口昌也・西川賢哉・石塚京子・木村睦子(2004)『日本語話し言葉コーパス』の形態論情報の概要.

高梨克也・内元清貴・丸山岳彦 (2004) 『日本語話し言葉コーパス』における節単位認定.

内元清貴・丸山岳彦・高梨克也・井佐原均 (2004) 『日本語話し言葉コーパス』における係り受け構造付与.