因子分析による日本語文末表現の機能・叙法分析

玉城 伸仁　　黒橋 禎夫
京都大学大学院情報学研究科
{tamaki, kuro}@nlp.kuee.kyoto-u.ac.jp

1　はじめに

アスペクトやモダリティ、口調など様々な機能範疇が文末に畳み込まれて表出されることは日本語の大きな特徴のひとつである。文末表現は助動詞や接尾辞など複数の構成要素の複合体として形成される。本稿では文末表現の共起統計量の相関を因子分析によって分析し、抽出因子を文末表現の機能として解釈することを試みる。

これは、「〜だろうな」は「推量 + 会話調 + ···」、「〜であろう」は「推量 + 論説調 + ···」というように、文末表現に対する共起統計量を線形に分解して説明可能であると仮定することを意味する。一見、極めて乱暴な仮定である。しかし我々は玉城&黒橋 (2009) [7] において、副詞や機能語との共起統計量から構成した文末表現の特徴ベクトルを線形に分解することで、口調成分と非口調成分をある程度分離して評価可能であることを報告している。先稿は系外で作製した外部基準によって口調成分を分離することを狙ったものであったが、本稿では外部基準なしに変量間の相関関係のみを手がかりとした分析を試みる。

因子分析に類似の手法として主成分分析やその亜種である潜在意味インデキシング (LSI) [4] による次元圧縮手法がある。しかしこれらの手法では、実質科学的な仮説に基づいて直交軸を構成するであろう変数を選択しない限り、得られる構成軸に実体的な意味は期待できない。因子分析では各変数の独自分散の推定と、因子間の相関を考慮した斜交軸の構成が可能であるので [8]、仮説を持たない探索的な分析、結果の実質科学的解釈という矛盾した要求がある場合、適当なバランスをとりながら柔軟な運用が可能である。

重要な先行研究として仁科らの一連の研究 [2] [3] が挙げられる。一貫して推量副詞と文末モダリティ表現の共起関係を分析した研究である。Srdanović 他 [2] では推量副詞のクラスタ分析による分類と、ジャンルによるクラスタ構造の遷移について論じており、Hodoscek 他 [3] では逆にジャンルごとの文末モダリティの使用傾向について論じている。仁科らは非母語話者による日本語作文支援を念頭においているので、どちらかというと精度を重視した研究である。ジャンル別に収集した複数のコーパスを扱っていることが特徴的であり、文数は数十万、対象としている副詞は 20、文末モダリティは数十個である。対して本稿はむしろ網羅性を重視しており、WWW 由来のものに限るが 10 億文以上のコーパスを扱い、600 の副詞、1 万パタンの文末表現を対象とした。

本稿ではWWWコーパスの機械解析結果から計算される、文末表現と副詞や機能語の共起尺度を観測変数として用い、変数間の相関を因子分析によって分析する。抽出因子を文末に畳み込まれた抽象的機能範疇として解釈する。

2　資料と方法

WWW に由来する日本語文書 1 億ページから重複のない日本語文を抽出した [5]。得られた 16 億文を形態素解析器 JUMAN、構文解析器 KNP で処理したのち、2 文節以上である文のみを取り出した。得られたコーパスは 14.5 億文であった。

文末文節を構成する形態素のうち名詞、動詞、形容詞を汎化したものを文末パタンとよぶこととする。文節単位の認定には構文解析器 KNP の文節まとめ上げ結果をそのまま用いた。名詞は品詞細分類へ汎化する。複合名詞はひとつにまとめ、末尾名詞の品詞情報を残す。「お、御」のような接頭辞は複合名詞へ含めずに残す。動詞/形容詞は活用語尾へ汎化する。ただし「思う/おもう」のみは汎化せずに扱う。

- 参加 団体 むけ パンフレットです
  →「【複合名詞 普通】です」
- 違ってたんだ
  →「【動詞 って】たんだ」

頻出 10005 パタンを観測個体として設定した。以下では表記の煩雑さを避けるために、必要に応じて実際の形態素を補った汎化していない表現で代記する。

コーパス中の頻出副詞 608 個と文末パタンの文内共起数を計数した。共起統計量として出現確率比 p(副詞 | 文末表現)/p(副詞) を算出した。

解析環境には R 言語を採用し、psych ライブラリ [1] を一部書き改めたものを用いて因子分析を実行した。観測標本数は 10005、観測変数は 608 であった。因子数は分析の前にあらかじめ推定しておくものとし、堀 [9] のガイドラインを参考に、Velicer の MAP 尺度による最適数と Horn の平行分析法による推定数を基準とした。MAP 尺度による最適数が 120 因子、平行分析法による推定数が 128 因子であった。特に仮説は無いので中間の 124 因子を仮定した。因子の分布に正規分布を仮定した最尤推定によって因子構造を推定した。推定された因子パタン行列に対して斜交回転のプロマックス法を適用した。

因子群の階層性を検討するために階層的クラスタ分析を行った。斜交回転を行った際に使用した回転行列から因子間相関を計算した。“1 – 相関計数” を非類似度としてウォード法による凝集樹形図を作製した。

3 結果と考察

3.1 抽出された因子

表 1 は抽出された因子の例である。因子得点上位の文末表現と、因子負荷量上位の副詞を併記した。因子得点は Bartlett 法によって推定した値である。これを見ると副詞の「おそらく」などと「だろう/のであろう」の推量表現が呼応する現象を捉えた因子であると解釈することができる。他の因子では「さっそく-してみます」「ちょうど-ところだった」「なんと-ではありませんか」などの対応関係を反映した因子が観察された。100 以上のほとんどの因子について、言語的な特徴を人間が解釈できる余地があり、比較的良好な単純構造を得ることができた。

ただし一部には解釈に困難を生じる因子も観察された。実験に用いたのは機械解析コーパスであるから、一定の誤りを含むことは前提である。しかし解析を誤った表現は本来の特徴が捉えられなくなるため、因子負荷量、因子得点の上位に挙がってくることは結果的にあまり無い。表 2 の例は、WWW 特有の語彙と解析器の系統的な誤りの相互作用の結果生じた珍しい例外である。WWW 上で良くみられる「どっと」「どっとねっと」という表現が「どっと (副詞)」と誤解析され、これもおそらく WWW 特有の固有表現「もっとねっと」の解析誤り「もっと (副詞) ね (終助詞) っと (格助詞)」と共起する例を特徴的に捉えた結果、解釈不能な因子が生じたものである。これが最も問題のあった因子であり、他に 2 例明らかな解析誤りが上位を占める因子があった。

表 1: 抽出された因子の例

第 103 因子	寄与率: 0.436%		
因子得点	文末表現	因子負荷	副詞
15.104	【動詞 た】のであろう	0.848	おそらく
13.942	ためであろう	0.835	恐らく
12.997	【動詞 して】いるのだろう	0.180	よほど
12.815	【動詞 した】のだろう	0.109	少なからず
12.633	【サ変名詞】したのだろう	0.079	どれほど
12.541	ためだろう	0.073	最も
12.024	【サ変名詞】していたのだろう	0.065	きっと
11.410	【サ変名詞】しているのだろう	0.057	なんらか
10.671	ものであろう	0.054	思えば
9.581	【普通名詞】なのであろう	0.052	もっとも
9.332	【動詞 って】いたのだろう	0.052	とっくに
9.279	【サ変名詞】されたのでしょう	0.051	今に
9.257	ためでしょう	0.049	何らか
8.875	【複合名詞 普通】であろう	0.048	たぶん
8.764	【サ変名詞】したのでしょう	0.039	せいぜい
8.359	【動詞 んで】いるのだろう	0.037	内心

因子得点は Bartlett 法による。斜交回転を行った結果であるため、単独の参考値としての寄与率の重要性は相対的に小さい。

表 2: 問題のある因子

第 40 因子	寄与率: 0.605%		
因子得点	文末表現	因子負荷	副詞
94.825	もっとねっと	0.888	どっと
16.561	激安通販価格最新情報 *	0.885	ようこそ
16.218	【動詞 い】ました	0.524	いきいき
9.533	【動詞 む】の	0.352	楽々
8.289	お【動詞 し】いただきました	0.189	なんでも
6.241	あい【普通名詞】	0.185	ひとつひとつ
6.217	ＹＳＴｇｏｏｇｌｅサイトマップ完全対応済みです *	0.162	ほのぼの

* 汎化パタンが冗長なので出現回数最多の実表現で代替して示した。

表 3 は独自性が大きかった変数と小さかった変数である。独自性最大は「夜な夜な」である。独自性が大、すなわち共通性が低い変数であるから、モダリティ尺度、口調尺度などを構成していく際には単に省くことになる。しかし、そもそも我々の感覚からすると「夜な夜な」のような若干特殊な副詞が最頻出 600 個に含まれていること自体が奇妙である。そこで p(夜な夜な | 文末パタン)/p(夜な夜な) が大きかった文末パタンを調べたところ「【ナノ形容詞】【サ変名詞】してくれる」の値が突出しており、汎化前の表現は「フリーダウンロードしてくれる」というものがほとんどであった。WWW 特有のコロケーションにより極端な統計量を計上したものと結論づけられる。「一目」「多少」以下のものについては、観測変数を増せば共通性の高い因子がでてくるのか、あるいは本当に日本語の言語空間の中で特殊な位置を占めるものであるか判らない。本稿の範囲では判断を保留しておく。独自性最小は「はたして/ついに/例えば」の 0.005 である。前提

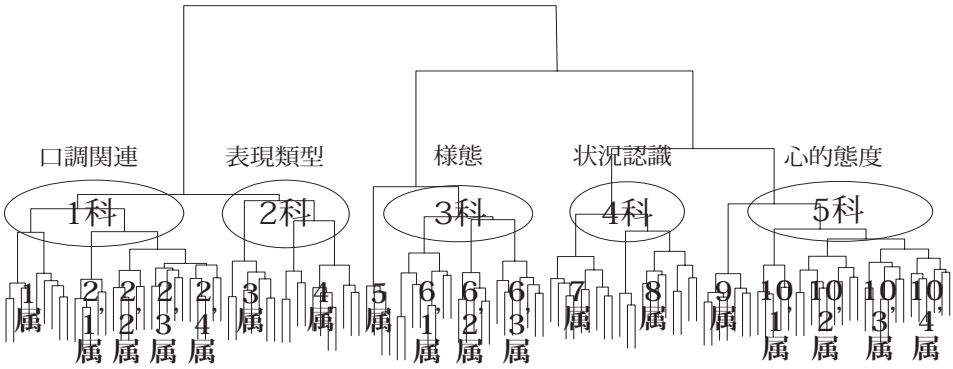

図 1: 共通因子の階層クラスタ構造

として実装上の特性を説明しておくと、独自因子の分散を推定する際に 0.005 ≤ 値 ≤ 1 の矩形制約を課した BFGS 法を用いている。これはその下限値ということになる。本当に共通性が大きいために独自性がほぼゼロになっている場合と、下限制約がなければ計算過程で不適解を生じていた場合が考えられるが、3 副詞ともいかにもモダリティ成分として文末表現と呼応しそうな頻出語であるから、実際に独自性が低い変数であったと考えられる。

表 3: 独自性最大と最小の副詞群

独自性	副詞	独自性	副詞
0.952	夜な夜な	0.005	はたして
0.941	一目	0.005	ついに
0.932	多少	0.005	例えば
0.922	ふっくら	0.010	一体
0.916	適宜	0.020	なんだか
0.915	即	0.022	もしかしたら
0.913	もれなく	0.024	いったい
0.906	たかが	0.027	確か
0.902	平常	0.028	たしか
0.899	どこまでも	0.032	ちょうど

各変数を説明する独自因子の分散を「独自性」として示した。

変数の多様性が増加した場合に安定した分析が可能であるかを確認するために、副詞に加えて接続詞、感動詞、助動詞など他の叙法・機能要素との共起尺度も合わせて観測変数を 2000 へ拡張した条件で同様の分析を実行した。このとき因子数は 314 を仮定した。表 4 は抽出された因子の例である。変数の多様性とスケール増加の影響は特に無く、安定した分析が可能であった。ただし、本稿では全体の見通しを重視して副詞のみの結果から考察を進める。

表 4: 副詞以外への拡張

因子得点	文末表現	因子負荷	副詞
20.694	【サ変名詞】でしたっけ	1.152	感動詞_えー
17.656	【複合名詞 普通】でしたっけ	1.134	感動詞_えーっ
15.425	【数詞】だったかな	0.775	終助詞_っけ
13.713	どこかな	0.623	感動詞_えっ
13.214	【複合名詞 普通】だったっけ	0.379	感動詞_ええ
10.907	いつ【複合名詞 普通】	0.305	接続詞_ところで
10.696	【動詞 んだ】っけ	0.254	指示詞_あれ
8.098	【複合名詞】なんですか	0.193	副助詞_って
7.240	【動詞 いた】っけ	0.173	格助詞_っと

3.2 因子の階層分類

図 1 に因子間相関をもとにした階層クラスタ分析の結果を模式的に示した。樹形図のリーフが各因子を表す。結果の階層的な関係を検討するために、2 階層の分類粒度を仮設する。相対的に大きい分割群を「科」、小さい分割群を「属」とする。3 つの属は比較的大集団となったので配下にさらに 11 亜属を設ける。都合、5 科 10 属 (11 亜属)124 因子として全体を整理することにする。2 属配下の亜属を 2-1' 属、2-2' 属などと記す。記号的に記述すると {1 科; 1 属, 2 属 }, {2 科; 3 属, 4 属 }, {3 科; 5 属, 6 属 }, {4 科; 7 属, 8 属 }, {5 科; 9 属, 10 属 } である。

トップレベルの分岐は 1, 2 科と 3-5 科の対立である。1 科は伝達態度の違いに関わる因子群であると考えられる。1 属にカジュアルな普通体が多く、2 属配下の各亜属はカジュアルなデスマス、尊敬・謙譲を伴うデスマス、脱場面的という特徴的な文末表現の対立を示している。

2 科は仁田ら [6] でいう表現類型のモダリティと、認識や説明のモダリティの一部が含まれる。3 属に含まれる各因子の代表的な表現は「しみじみ-思うのだった」「おそらく-のであろう」「試しに-してみました」

などで、特に相手目当てでないものが出現している。4 属ではどちらかというと相手目当ての表現が現れ、「おおむね-のようになっています」「もっとも-どれでしょうか」「もともと-ものだそうです」などである。1, 2 科は広い意味で叙述様式に関わる因子群であると解釈できる。

3 科は 5 属と 6 属である。5 属は他に比べて小集団であるが、確かに分岐するだけの特徴があって、いわゆる擬態語が負荷量の上位を占める因子が集中している。副詞のみを列挙しておくと「スッキリ、はっきり、うろうろ、ゴロゴロ、キラキラ、生き生き、ワクワク、ドキドキ、ほのぼの、ほんのり」といった様子である。6 属は 3 亜属に分ける。6-2' 属と 6-3' 属には陳述の副詞を伴った表現が並ぶが、文末は比較的中立であり、全体として心的態度の表明はさほど強くない。6-2' 属が「なんだか-くなっちゃいました」「あいかわらず-ですなぁ」「なんとも-ではありませんか」6-3' 属が「かろうじて-れるようになりました」「さっそく-しなければならない」「心から-御 . . . りしております」となっていて、両亜属間に鋭い対立は観察されないが、6-3' 属の方が若干フォーマルといえる。6-1' 属は他の 6 属と 5 属の中間的な性格で「ぼちぼち-していきますかね」「ぜんぜん-くなかったですよ」「さっぱり-らんな」などとなっていて、陳述の副詞だけでなく、擬態語を含む程度、情態の副詞と、その誘導が感じられる文末という組み合わせが多い。3 科は全般に文末よりも副詞を軸にまとまりがある因子群である。

4 科は 7 属と 8 属である。7 属は「たぶん-じゃないかな」「なにせ-だからね」「いまさら-いわれてもなぁ」と、話者の事態に対する確信度が低く、相手の意向を期待するととれる表現が多い。8 属は逆に「ひとつひとつ-していきましょう」「改めて-させられました」「原則として-しかねます」のように話者の事態に対する確信度が高く、自らの意向を相手へ示す表現が前面に出る。4 科は事態に対する確信度の高低に対立軸のある因子群であると考えられる。

5 科は品詞の掃き溜めの本領発揮といえるかもしれない。9 属と 10 属の間の対立はさほど判然としない。文末表現は多様で主観的な色付きの表現が散見される。副詞は主に陳述の副詞であるが、6 属ほどまとまりはなく比較的雑多な印象である。「なんだかんだ-一番ですよ」「決して-ばかりではありません」「正直-ではありません」などが代表例である。5 科全体として言えることは、比較的強い心的態度の表明に関わる因子群ということである。

3-5 科を通して観察すると 3 科と 5 科が相対的に近く、5, 6, 9, 10, 7, 8 属と並べて、様態の副詞 ⇒ 弱い心的態度 ⇒ 強い心的態度 ⇒ 確信度の高い状況認識 ⇒ 非断定的な状況認識、という緩やかなまとまりを持った主観的モダリティ因子群が、叙述様式・伝達様式を担う 1, 2 科と対立していると整理できる。

今回の分析手続きは決して客観的なものではない。特に共起尺度、因子回転法、クラスタリング法の選択に大きな恣意性がある。しかし、これは分析者に与えられた大切な自由度である。内省的な研究をおこなう際に、様々に切り口を変えながら文法現象の整理・分類をおこなっていくことに相当する。

4 おわりに

本稿では、文末表現と副詞などの叙法・機能要素との共起統計量の相関を因子分析によって分析した。多くの抽出因子について言語現象として解釈が可能であることが確認できた。文法の研究は最終的に内省的な方法による一種の職人芸に頼らざるを得ないことが多いが、本稿のような数理的手法がその補助となり得ることを示せたと考えている。

参考文献

[1] http://cran.r-project.org/web/packages/psych/.

[2] I.S. Erjavec, A. Bekes, and K. Nishina. Distant Collocations between Suppositional Adverbs and Clause-Final Modality Forms in Japanese Language Corpora. *Lecture Notes in Computer Science*, Vol. 4938, p. 252, 2008.

[3] Bor Hodoscek, Andrej Bekes, 仁科喜久子. 推量的副詞の共起情報に基づいた genre 別の文末表現の分析. 言語処理学会第 15 回年次大会, pp. 598–601, 2009.

[4] T. Hofmann. Probabilistic latent semantic indexing. pp. 50–57. ACM New York, NY, USA, 1999.

[5] D. Kawahara and S. Kurohashi. Case Frame Compilation from the Web using High-Performance Computing. In *Proceedings of The 5th International Conference on Language Resources and Evaluation (LREC-06)*, pp. 1344–1347, 2006.

[6] 日本語記述文法研究会. 現代日本語文法 4 モダリティ. くろしお出版, 2003.

[7] 玉城伸仁, 黒橋禎夫. 文体横断的な文末機能の類似度測定. 言語処理学会第 15 回年次大会, pp. 434–437, 2009.

[8] 豊田秀樹. 共分散構造分析 [理論編] 構造方程式モデリング. 朝倉書店, 2007.

[9] 堀啓造. 因子分析における因子数決定法. 香川大学経済論議, Vol. 77, No. 4, pp. 35–70, 2005.