統計数理 第64巻 第2号

（通巻124号）

PROCEEDINGS OF THE INSTITUTE OF STATISTICAL MATHEMATICS

目　次

特集「統計的言語研究の現在」

2016年12月

大学共同利用機関法人 情報・システム研究機構 統計数理研究所

〒190-8562 東京都立川市緑町10-3　　電話 050-5533-8500（代）

本号の内容はすべて http://www.ism.ac.jp/editsec/toukei/ からダウンロードできます

ISSN 0912-6112

統計数理　PROCEEDINGS OF THE INSTITUTE OF STATISTICAL MATHEMATICS

第64巻 第2号 2016

統計数理 Vol. 64, No.2

PROCEEDINGS OF THE INSTITUTE OF STATISTICAL MATHEMATICS

統計数理研究所

統計数理

（年 2 回発行）

「統計数理」は，統計数理研究所における研究成果を掲載する統計数理研究所「彙報」として 1953 年に歴史を始め，1985 年に誌名を変更し今の形となりました．現在は，統計数理研究所の研究活動に限らず，広く統計科学に関する投稿論文を掲載し，統計科学の深化と発展，そして統計科学を通じた社会への貢献を目指しています．

投稿を受け付けるのは，次の 6 種です．

a.　原著論文　　b.　総合報告　　c.　研究ノート
d.　研究詳解　　e.　統計ソフトウェア　　f.　研究資料

投稿された原稿は，編集委員会が選定・依頼した査読者の審査を経て，掲載の可否を決定します．投稿規程，執筆要項は，本誌最終頁をご参照ください．

また，上記以外にも統計科学に関して編集委員会が重要と認める内容について，編集委員会が原稿作成を依頼することがあります．

その他，「統計数理」に関するお問い合わせは，各編集委員にお願いします．

All communications relating to this publication should be addressed to associate editors of the Proceedings.

大学共同利用機関法人　情報・システム研究機構
統計数理研究所
〒 190-8562 東京都立川市緑町 10-3　　電話 050-5533-8500（代）
http://www.ism.ac.jp/

印刷：笹氣出版印刷株式会社

表紙の図は本誌 151 ページを参照

PROCEEDINGS OF THE INSTITUTE OF STATISTICAL MATHEMATICS

Vol. 64,　No. 2

Contents

December, 2016

Research Organization of Information and Systems
The Institute of Statistical Mathematics
10–3 Midori-cho, Tachikawa, Tokyo 190–8562, JAPAN

統計数理(2016)
第 64 巻 第 2 号 141–143

「特集 統計的言語研究の現在」編集にあたって

持橋 大地[1]・前川 喜久雄[2]・浅原 正幸[2]

統計学と言語学の間には過去にも若干の交流があった．本誌のバックナンバーを繰ってみると，26 巻 1–2 号(1979 年刊)に村上征勝氏が「著者推定問題における統計的手法」と題した研究ノートを寄稿しておられるのが見つかる．その後 20 年ほどの間隔をあけて，48 巻 2 号(2000 年刊)に同じ村上氏がエディターを務めた「“ことば” 新研究」という特集が掲載されている．そのまえがきには，シェークスピア＝ベーコン説などの著者推定問題に始まり，その後細々と続いてきた言語に関する統計的研究が，「ここに来て，大きなうねりとなる兆候が見えてきている．というのも近年のコンピュータの著しい進歩により，大量の “ことば” のデータベースが構築できるようになり，加えて，安価で高性能のパソコンの普及により，比較的簡単に複雑な統計分析ができるようになった為である．“ことば” の分析は新たな段階に入ったといえる.」との記述がある．

村上氏が 20 世紀末に指摘したこの「うねり」は，その後，インターネットがもたらすビッグデータの登場とも呼応して，現在では関連学会のみならず産業界をも飲みこむ大波に成長している．現在，音声認識や自然言語処理の主流が機械学習と連携した統計的なアプローチにあることは，広く認識されているとおりである．

20 世紀末と現在を比較してもうひとつ気づくのは，統計手法の利用目的が仮説検定や主成分分析といった事後の分析から，データのモデリングに移行していることである．一般化線形混合モデル，MCMC による階層ベイズモデルなどの普及により，統計ユーザーがみずからの発意で自由に統計モデルを構築できるようになったことは，言語分析にかぎらず，人文学領域の問題に対する統計学の応用可能性を飛躍的に高めたように思われる(その背後には R や Python に代表される，無償かつ高機能な計算環境の普及があることはいうまでもない)．

本特集の出発点となったのは，「統計的言語研究の現在」と題された国立国語研究所と統計数理研究所の合同シンポジウム(2015 年 9 月 4 日開催)であった．国立国語研究所講堂で開催されたこのシンポジウムは，われわれの予想をこえる参加者 116 名の盛会となり，言語研究における統計的手法の注目度の高さを感じさせるものであった．シンポジウムの内容，および講演スライドはすべて，ホームページ[1)]から今でもご覧いただくことができる．本特集所載の論文のうち，新井論文，村脇論文，荒牧論文はこのシンポジウムでの講演者の手になるものであり，他の論文も，シンポジウムの企画段階で講師の候補に名前があがった方々に執筆していただいている．他に言語心理学や社会言語学の領域からも寄稿していただく予定があったが，諸般の事情で実現に到らなかったのは残念であった．

従前の統計的言語研究は，主に単語などの頻度情報に基づくものであった．一方，本特集で扱う言語のデータは多様な形式からなる．単語列や品詞列などを抽象化した系列，系統樹や構文構造の根幹をなす木構造，被験者の反応・経年変化の手がかりとなる時間，さらには Twitter 発言の GPS 情報や言語接触がもたらす空間情報からなる．このような複雑な構造を持った情報に基づいた統計処理を進めることが，言語研究にも必要になりつつある．

[1] 統計数理研究所：〒 190–8562 東京都立川市緑町 10–3
[2] 国立国語研究所：〒 190–8561 東京都立川市緑町 10–2

能地論文は構文解析のうち，ここ 20 年間の教師なし構文解析や文法推定（Grammar Induction）の動向について紹介している．構文木を扱う枠組として，句構造文法・依存文法・組合せ範疇文法についての研究について，どのように文法規則を推定するかについて概説している．この試みは「文法がどのくらい生得的なのか」について，テキストデータから統計的に示唆を与えるものであり，今後の進展が期待される．

村脇論文は言語類型論（Typology）に対する統計的なアプローチについて概説している．近年，言語の類型論的特徴データが整備され共有されるようになった．これらのデータの特性を解説しながら，系統樹を復元する統計モデルについて紹介している．5 節では日本語の起源についても言及しており，クレオール形成に着目した言語接触に関する議論は特に興味深い．

岡崎論文は系列ラベリング技術に関するものである．岡崎氏は CRFSuite[2)] と呼ばれる条件付確率場（Conditional Random Fields）に基づく系列ラベリングツールを公開しており，系列に対するロジスティック回帰についての丁寧な解説を行っているほか，系列ラベリング研究の最新の動向についても言及している．

新井論文は心理言語学で用いられる眼球運動測定器から得られるセンシングデータをどのように統計処理するべきかを紹介している．眼球運動測定器から得られる情報がどのような性質をもち，どのようにモデリングするかを解説している．実験のデザインから解説しており，心理言語学の分野に興味がある研究者への良きチュートリアル資料になっている．

最後に荒牧論文は，ソーシャルメディアサービスを用いた言語研究について紹介している．ソーシャルメディアは非文法的でノイズの多い言語データであり，多数の発言者のデータを発言時刻と発言地点などの GPS 情報とともに得ることができ，これらを用いた新たなアプリケーションについて言及している．

こうした研究は，統計的機械学習の一部である統計的自然言語処理，および言語学自体への統計的手法の導入から生まれたものである．本特集は，その中でも特に工学より言語科学に近い，今後の発展が見込まれる内容を議論して論文の執筆を依頼した．一方で本特集は CRF や教師なし構文解析といった，ある程度確立された分野の，これまでになかった詳しい解説ともなっており，自然言語処理の技術に興味がある方にも有益な内容となっていると考えている．

前回の 2000 年の特集は，今回扱っている統計的自然言語処理とデータ科学の時代を予感させるものであった．今回の論文の内容をよく読むと，次の時代の研究の種子があちこちに埋まっていることがわかる．たとえば能地および村脇論文は本質的に，言語に存在する木構造や分岐構造を生成する確率過程をどう考えるか，という問題であるし，荒牧論文では暗黙に，感染症のモデルである SIR モデルが言語の場合にどう現れるかという問題を扱っている．どちらも統計学の分野でこれまで研究はあるものの，統計学の場合は問題を数理的に解きやすくするために大きく単純化されており，今回のような言語の場合にどう適用していくかはまだほとんど未開拓の荒野であるといってよい．

また新井論文では，眼球運動データと文の読み時間が分析されており，逆ガウス分布の当てはまりがよいことが示されている．逆ガウス分布を使う理由については引用されている Lo and Andrews (2015) の中でも弱い形で説明されているが，もし眼球運動を（バイアスのかかった）ブラウン運動とみなせば，その一定距離，すなわちここでは文末までの到達時間分布が逆ガウス分布になることは統計学においてはよく知られた事実であり，これは新井論文の観察とも符合する．したがって，視線の運動を直接，特殊なランダムウォークとしてモデル化することもできそうである．

岡崎論文は CRF（Conditional Random Fields，条件付確率場）の入門と詳細な解説であるが，論文の中で述べられているように，CRF はロジスティック回帰を時系列化したものといえ，現代の統計的自然言語処理全般においてきわめて重要な方法となっている．本特集ではふれなかった

が，現在ニューラルネットワークに基づく手法が自然言語処理を席巻しており，本特集編集時にも，Google 翻訳がニューラルネット化して大幅に精度が向上したことが話題となった．ニューラルネットがどのように動作しているのかはまだ解明されていないものの，このような中で，ニューラルネットで CRF の素性関数を学習する CNF（Conditional Neural Fields）(Peng et al., 2009)，画像の場合であるが CRF の学習をレイヤー化して再帰的ニューラルネットとして捉え，ニューラルネットとの統合を図る研究（Zheng et al., 2015）などの研究が発表されており，統計的に堅実な手法である CRF がどのように利用されていくのか，今後の動向が期待される．

こうした現在の統計的言語研究の先端を詳しく紹介する中で，その内容とともに，次の時代の言語研究の息吹を読みとっていただければと考えている．最後に，新しい試みが多い本特集の論文をお願いできる方を言語学または自然言語処理の中で探すのには労を要したが，貴重な査読者の方々にはお忙しい中，大変有益なコメントをいただくことができた．この場をお借りして感謝を申し上げたい．また，再度にわたる原稿の修正を行っていただいた執筆者の皆様，およびお世話になった編集室の方々にお礼を申し上げたい．

注．

1) http://www.ism.ac.jp/~daichi/workshop/2015-statling/
2) http://www.chokkan.org/software/crfsuite/

参 考 文 献

Lo, Steson and Andrews, Sally (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data, *Frontiers in Psychology*, **6**, 1171.

Peng, Jian, Bo, Liefeng and Xu, Jinbo (2009). Conditional neural fields, *NIPS 2009*, 1419–1427.

Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C. and Torr, P. H. (2015). Conditional random fields as recurrent neural networks, *ICCV 2015*, 1529–1537.

統計数理(2016)
第 64 巻 第 2 号 145–160

特集「統計的言語研究の現在」
[研究詳解]

文に隠れた構文構造を発見する統計モデル

能地 宏†

(受付 2016 年 5 月 4 日；改訂 9 月 8 日；採択 10 月 7 日)

要　　旨

本稿は，自然言語の文法を単語列から自動で抽出する教師なし構文解析について，過去 20 年間に渡る研究の進展について紹介を行う．この研究で本質的に重要となるのは，言語の文法に関するバイアス，もしくは知識をどのようにモデルに組み込むか，という点である．本稿ではこの観点から様々な既存のモデルを比較し，どのような知識を仮定することでどの程度の文法が獲得できるようになったのかについてまとめることで，教師なし構文解析が今後向かうべき方向性についての議論の指針としたい．

キーワード：計算言語学，教師なし構文解析．

1. はじめに

自然言語処理において文の統語構造(木構造)を明らかにする構文解析は最も基礎的かつ重要な技術である．例えば図 1(b)のような依存構造木が得られれば，ここから単語間の意味関係，例えば read の目的語が the book であることが読み取れ，これらが機械翻訳や質問応答で利用される．

本稿では，構文解析に関する研究のうち，特に教師なし構文解析，もしくは文法推定(grammar induction)の問題に関する最近の進展についてまとめる．この問題は自然言語処理が統計的アプローチにシフトし始めた 90 年代初期の頃から存在し，また当時から非常に困難な問題として知られていた(Lari and Young, 1990)．木構造に対するモデルとして文脈自由文法などの簡単なモデルを仮定すれば，そのパラメータ(文法の各書き換えルールに対する重み)は Expectation-Maximization(EM)アルゴリズムを用いて文のみの集合から機械的に計算することができる．しかしながらそのようにして得られた文法は言語学的に正しいとされるものとは大きくかけ離れており，長い間文法の教師なし獲得は不可能であると信じられていた(Manning and Schütze, 1999)．このように一旦停滞しかけていた研究であるが，2004 年の Klein らの研究(Klein and Manning, 2004)によるモデル及び学習法によって再び注目を集め，その後約 10 年間で様々な改良が行われ，現在に至っている．本稿では，特にこの過去およそ 10 年間の進展をまとめることにより，今後の教師なし構文解析の方向性に関する議論の指針を与えたい．

近年行われている研究の多くは，モデルに文法に関する事前知識，もしくは常識をどのように取り入れるか，という点に焦点を当てたものが多い．これは言い換えれば，最小限の労力で新しい言語に対するツリーバンクを構築することを目標とした際に必要な事前知識，もしくは外部知識を明らかにする立場であるといえる．例えば Klein らの研究では本質的には EM アルゴリズムの初期値が最も精度の向上に寄与しており，この初期値は言語学的直感に基づいて設

†奈良先端科学技術大学院大学 情報科学研究科：〒 630–0192 奈良県生駒市高山町 8916–5

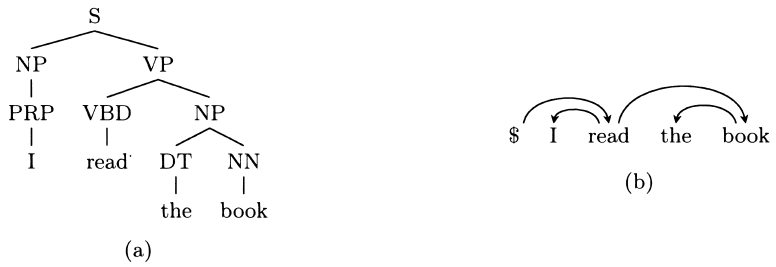


図 1. 構文解析が扱う木構造の例．(a)は句構造木，(b)は依存構造木を表す．$ は常に文頭に置かれる仮想的なノードで文のルート(read)を子に持つ．

計されている(3.1 節)．より最近の研究，例えば Bisk and Hockenmaier (2013)は，名詞と動詞間の言語普遍的な振る舞いを語彙化文法(組合せ範疇文法)の枠組みでモデルに組み込んでいる．この観点から見ると，過去の研究は，どのような知識の入れ方が最も効率的であるかについての試行錯誤であったと捉えることができ，また研究が進展するにつれ，どこまで知識を仮定すればどの程度文法が得られるのかについての知見が蓄積されてきたように思われる．機械学習の立場から別の興味深い問題は，モデルが正しいと仮定した際に最適解をどのように得るか，という問いであるが，この点に関する研究はまだ少ない．本稿では大きく扱わないが，分枝限定法(Gormley and Eisner, 2013)やモーメント法(Hsu et al., 2012)などの適用が検討されてきた．

上記の立場は基本的には工学的な有用性を探求する立場と言えるが，教師なし構文解析は理学的，あるいは哲学的にも興味深い問題であり，特に初期の研究にはそういった立場に立脚したものも見受けられる．例えば Clark (2001)は計算機が文の集合から文法が獲得できることを示すことにより，理論言語学における刺激の貧困(poverty of stimulus) (Chomsky, 1986)に対する経験的な反証が行えると主張している．本稿ではこの点についてはあまり触れないが，言語獲得のモデル化という観点からは両者は切り離せるものではないだろう．例えば上で述べた文法に関する事前知識を利用した学習は，言語の文法がどの程度生得的なのか，という問いに対する示唆を計算言語学の立場から与えるものであると考えられる．

本稿で扱うほぼ全てのモデルは，確率文脈自由文法に対する EM アルゴリズムもしくはその拡張によって説明が行える．その上でほとんどの議論は，どのような文法の枠組みが文の集合のみからの学習に適するか，そしてどのような推論法がより学習を促進させるのに有効か，という点に集約される．本稿ではまず 2 節でこの大きな枠組みを説明した後，その単純な応用である句構造文法の EM アルゴリズムがうまくいかなかったことを述べる．その後 3 節で大きな転換点となった依存構造の学習に焦点を当て，主要な研究をかいつまんで解説する．最後に 4 節で近年注目を集めている言語理論に基づく組合せ範疇文法の教師なし学習について紹介し，5 節でまとめを行う．

2. 確率文脈自由文法と EM アルゴリズム

2.1 確率文脈自由文法

確率文脈自由文法(PCFG)は $G = (N, \Sigma, P, S, \theta)$ の 5 つ組で表現される．ここで (N, Σ, P, S) は一つの文脈自由文法(CFG)であり，N が非終端記号，Σ が終端記号，P が書き換えルールの集合を表す．終端記号は構文木の葉ノードに出現し，非終端記号はそれ以外の内側に出現する記号として区別される．各書き換えルール $r \in P$ は $A \to \beta$ の形をもつ．ここで $A \in N$,

$\beta \in (N \cup \Sigma)^*$(空または記号の列)である．$S \in N$ は特別な文法の開始記号である．図 1(a)はある CFG が入力文 I read the book に対して与える解析例を示しており，NP, VP などが非終端記号，I, read などが終端記号，S $\rightarrow$ NP VP などがルールの例となっている．θ はパラメータであり，各 $r \in P$ に対して確率値を割り当てる．ここで θ_r で r に対する確率を表すと，

$$\forall A \in N, \sum_{A \rightarrow \beta \in P} \theta_{A \rightarrow \beta} = 1$$

が成り立つ．すなわち各非終端記号 A は子の書き換えに関する多項分布をもつ．

PCFG は構文木に対する分布を与える．ある PCFG G のもとで，一つの構文木 z は開始記号 S からの再帰的な書き換えルールの集合とみなすことができるので，その確率は，

$$p(z|\theta) = \prod_{r \in z} \theta_r$$

である．また，ある PCFG G が与えられたとき，入力文 $x = x_1, x_2, \ldots, x_n$(各 x_i は単語)に対する最適な構文木は，動的計画法である CKY アルゴリズム(Kasami, 1965; Younger, 1967)を用いて効率的に計算できる．

$$\hat{z} = \arg \max_{z \in \mathcal{Z}(x)} p(z|\theta)$$

ここで $\mathcal{Z}(x)$ は x に対してあり得る構文木の集合である．

本稿で扱う CFG は全て，ルールの右辺 β の大きさが 1 または 2 のものに限られる．上で述べた CKY アルゴリズム，もしくは以下の内側外側アルゴリズムは，この仮定により大きく簡略化される．

2.2 EM アルゴリズムによる学習

PCFG は隠れマルコフモデル(HMM)の木構造への一般化とみなすことができる．そしてこの観察から，HMM に対する EM アルゴリズムと同じように PCFG に対する EM アルゴリズムを導出することができる．文の集合 $\boldsymbol{x} = x^{(1)}, x^{(2)}, \ldots, x^{(m)}$ が与えられたとき，EM アルゴリズムは次の対数尤度を上昇させるように θ を更新する．

$$L(\theta) = \sum_{x \in \boldsymbol{x}} \log p(x|\theta) = \sum_{x \in \boldsymbol{x}} \log \sum_{z \in \mathcal{Z}(x)} p(x, z|\theta). \tag{2.1}$$

各更新は E ステップと M ステップの二段階からなる．E ステップでは，現在の θ のもとでのルール r の期待値 $e(r)$ を計算する．

$$e(r) \leftarrow \sum_{x \in \boldsymbol{x}} e_x(r),$$

$$e_x(r) \leftarrow \sum_{z \in \mathcal{Z}(x)} p(z|x) f(r, z). \tag{2.2}$$

ここで $e_x(r)$ は文 x における r の期待値，$f(r, z)$ は構文木 z 中でルール r が使われた回数である．M ステップでは，この期待値を正規化することで θ を更新する．

$$\theta_{A \rightarrow \beta} \leftarrow \frac{e(A \rightarrow \beta)}{\sum_{\alpha : A \rightarrow \alpha \in R} e(A \rightarrow \alpha)}$$

ここで最も重要なのは，ルールの期待値 $e_x(r)$ を効率よく計算することである．これには HMM での前向き後ろ向きアルゴリズムとよく似た内側外側アルゴリズム(inside-outside algorithm)が利用できる．以下 $r = A \rightarrow B\ C$，つまり右辺の大きさが 2 のルールを仮定する．またスパン

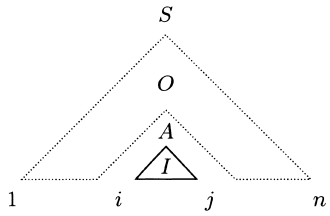


図 2. 内側確率(I)と外側確率(O)の概念図．n を文長，i,j を文中の単語のインデックスとして，各スパン (i,j) 及び非終端記号 A 毎にこれらが計算される．$x_{i:j} = x_i, x_{i+1}, \ldots, x_j$ を表す．

(i,j) で，文中の i 番目の単語から j 番目の単語までの範囲を表す．

まず，$e_x(r)$ を次のように分解しよう．

$$e_x(r) = \sum_{1 \le i \le j \le k \le n} e_x(r,i,j,k). \tag{2.3}$$

$e_x(r,i,j,k)$ は，ルール $r = A \to B\ C$ の B がスパン (i,j) を，C が $(j+1,k)$ を張ることに対する期待値である．これは二値の確率に対する期待値であるから，それが発生する確率自体に等しく

$$e_x(r,i,j,k) = p(r,i,j,k|x,\theta) = \frac{p(r,i,j,k,x|\theta)}{p(x|\theta)}$$

となる．従って，入力 x の周辺確率 $p(x|\theta)$ 及び 各 (r,i,j,k) 毎に $p(r,i,j,k,x|\theta)$ を計算できれば，式(2.3)によって r の期待値が得られる．

内側外側アルゴリズムは，これらの量を動的計画法によって効率的に計算するアルゴリズムである．これは，各スパン (i,j) 及びその根の記号 A 毎に，二つの量 $I[i,j,A]$（内側確率）と $O[i,j,A]$(外側確率)を計算していく．図 2 に概念図を示す．$I[i,j,A]$ は A を開始記号としたときの $w_i, \ldots, w_j$ に対する周辺確率であるのに対し，$O[i,j,A]$ は (i,j,A) の外側の構造に対する周辺確率となっている．

内側確率が全て求まれば，文に対する周辺確率は $p(x|\theta) = I[1,n,S]$ として得られる．

$p(r,i,j,k,x|\theta)$ は若干複雑だが，まずこれは次のように，x が (r,i,j,k) を含む構文木から生成される確率を表すことに注意する．

$$p(r,i,j,k,x|\theta) = \sum_{z \in \mathcal{Z}(x):(r,i,j,k) \in z} p(z,x|\theta).$$

ここで $(r,i,j,k) \in z$ は r が (i,j,k) の位置で構文木 z 中に出現することを表す．そしてこの確率は，次のように (r,i,j,k) の前後で内側確率と外側確率を用いて分解することができる(θ への依存性は省略した)．

$$\begin{aligned} p(r,i,j,k,x) &= p(x_{1:i-1}, A, x_{k+1:n}) \times p(A \to B\ C) \times p(x_{i:j}|B) \times p(x_{j+1:k}|C) \\ &= O(i,k,A) \times \theta_{A \to B\ C} \times I(i,j,B) \times I(j+1,k,C). \end{aligned}$$

r の右辺の大きさが 1 の場合は省略するが，同様の式を導くことができる．以上が学習の概略であるが，内側確率，外側確率の再帰式など，アルゴリズムのより詳細については Manning and Schütze (1999)などの教科書を参照されたい．

2.3 句構造文法の推定

内側外側アルゴリズムによる期待値計算により，PCFG のパラメータ θ は，CFG (N, Σ, P, S) と特定の初期値 $\theta^{(0)}$ を定めれば推定を行うことができる．

90 年代，この考えに基づき図 1(a)のような句構造文法を教師なしで学習する研究がいくつか行われたものの，得られた文法は言語学者の考える正解とは大きくかけ離れていた(Carroll and Charniak, 1992)．この失敗の原因として，次のような点が考えられる．

(1) 第一に EM アルゴリズムは局所探索法であるため，得られる文法は対数尤度(式(2.1))の大域的最適解ではなく局所解だという点である．木構造の探索は範囲が非常に大きく，この局所解の問題が特に問題となる．Carroll and Charniak (1992)はこの影響を調べており，人工データに対して 300 回の試行でランダムに初期化したモデルは全て違う局所解に陥ったと報告している．

(2) もう一つの問題は，句構造文法の恣意性と表現力の弱さである．PCFG の学習において，固定されている情報は開始記号 S 及び観察された終端記号の列(単語もしくは品詞)のみである．ここでの問題は，ある木構造が与えられたとき，終端記号は多くの情報量を持っているものの，これらと木の中間ノードとの結びつきが，木の上方になるほど指数的に失われていくという点である．これは本質的には，句構造文法で扱う非終端記号が単なる抽象的な記号としてしか振る舞わないことに起因する．例えばモデルに $y_1 \to y_2\ y_3$ というルールが存在したとする．ここで y_1 と終端記号との結びつきを考えると，y_1 は $y_2\ y_3$ を通してでしかこれらと関連を持たず，結果結びつきは急速に失われる．

(3) 最後にこれと関連するが，EM アルゴリズムが見つけやすい構造と言語学者が正しいと考える構造の間には隔たりがあることが指摘されている．例えば正解データから教師あり学習で得たモデルを EM アルゴリズムの初期値として使用した場合，学習を進める毎に尤度(式(2.1))は上昇するものの精度は逆に悪化してしまう(Liang and Klein, 2008)．また英語の文集合に対しモデルが見つけやすい典型的な間違いとして，頻度の高い語の並びを句にまとめてしまうという挙動があげられる．例えば，英語では代名詞，動詞という並びが典型的なため，図 1(a)のような構造ではなく，主語である代名詞と動詞が直接句を形成するような文法が学習されやすい(Pereira and Schabes, 1992)．

次節で述べる依存文法の学習は，学習する PCFG の構造に制限を加えることで上記の 2 と 3 の問題を緩和しようとするものだといえる．1 は非凸関数の最適化に起因する問題であるが，これについても初期値の工夫など様々な改善がなされてきた．

初期の EM アルゴリズムを用いた学習で唯一成功したのは，人手で構築した正解データを用いてあり得る句構造のスパンに制約を課す方法(Pereira and Schabes, 1992)であり，これがその後のコーパス主導の教師あり構文解析へと発展していく(Charniak, 1996; Collins, 1997)．また EM 以外の学習方法として Johnson et al. (2007)はモデルのベイズ化及びサンプリング(マルコフ連鎖モンテカルロ法)に基づく推論を試みているが，状況は変わらなかったと報告している．

3. 依存構造の学習へ

PCFG に基づく文法の推定で初めてある程度の成功を収めたのは，依存文法の推定である(図 1(b))．句構造文法が文の構造を名詞句，動詞句など句同士の階層構造によって表現するのに対し，依存文法は単語間の依存関係によって表現する．各依存関係は head(主辞)から dependent (従属辞)に引かれ，多くの場合 dependent が head を修飾する関係となっている．例えば図 1 (b)において，the は book の dependent である．また通常，文全体の head(文のルート)を表

現するために，文頭に仮想的なノード(\$)を用意し，この右の子を文のルートとする．これにより，後に述べる PCFG への変換(図 5)などが簡略化される．

依存構造木と句構造木は一見大きく異なるものの，両者には透過的な関係がある[1)]．例えば図 1(b)において，the は book の左の子であり，the book が一つの小さな部分木，もしくは句を形成しているとみなせる．本節では依存構造の学習の例を示すが，これは，依存構造を表現する PCFG を定義しそのパラメータを推定するということである．具体的には，まず依存構造木に対する生成モデルを定義し，そのモデルを等価な PCFG に変換する．解析の際には，入力文を PCFG で解析し，得られた木から依存構造を復元すれば良い．

3.1 子の数を考慮に入れたモデル

ここでは研究の転換点となった Klein and Manning (2004)のモデルについて述べる．これは dependency model with valence(DMV)と呼ばれている．このモデルの一つの特徴そして限界として，単語ではなく品詞の上に定義されたモデルであるという点が挙げられる．彼らの研究は主に英語で行っており，扱う品詞の数はおよそ 40 程度である．自然言語の単語は数万から数十万以上であるため，これにより学習しなければならないパラメータの次元数が大幅に削減され，学習が行いやすくなる．ただしこの問題設定にすることで，実用的には，解析の前段階でまず品詞を予測する必要が生じる．文法の純粋な教師なし学習を考えた場合，正解の品詞が全ての文に付与されていることを前提とするのは現実的でない．一つの解決策は，まず単語に対する何らかのクラスタリングを行い，品詞の代わりに単語をクラスタで置き換えてモデルを構築することである．この方向性の研究としては，Headden III et al. (2008)や Bisk et al. (2015)などが存在する．

3.1.1 生成モデル

依存構造木に対して考えうるもっとも単純な生成モデルは，単語間の依存関係のみをモデル化したものであろう．この場合，各依存関係は $p_{\mathrm{A}}(d|h, dir)$ で，依存構造木の確率はこの要素の積でモデル化される．ここで $dir \in \{\leftarrow, \rightarrow\}$ は h(head)から d (dependent)への依存関係の方向である．

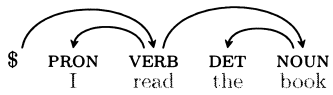
DMV はこのモデルを基本に，木の形を制御する別の要素 $p_{\mathrm{S}}(stop|h, dir, adj)$ を考慮したモデルとなっている．これはベルヌイ分布となっており，$stop \in \{\mathrm{STOP}, \neg\mathrm{STOP}\}$，そして $adj \in \{\mathrm{TRUE}, \mathrm{FALSE}\}$ が条件付け変数で，h が dir 方向に既に子を生成しているかどうかを判断し，まだしていなければ TRUE となる．図 3 に具体的なパラメータの例を示す．NOUN，VERB はそれぞれ名詞，動詞を表す．

具体的なモデルの生成過程を見るため，図 4 に DMV がある依存構造木に与える生成確率を示す．DMV では左側と右側の子はそれぞれ独立に生成される．h が各 d を生成する確率は $p_{\mathrm{S}}(\neg\mathrm{STOP}|h, \cdot, \cdot)$ と $p_{\mathrm{A}}(d|h, \cdot)$ の積で与えられる．最後に，各方向に子の生成を停止する $p_{\mathrm{S}}(\mathrm{STOP}|h, \cdot, \cdot)$ をかけ合わせる．

このモデルは英語の語順を多分に考慮に入れつつ設計されている．Balack Obama talked ... など，英語では NOUN NOUN VERB という並びはよく現れるが，最初に述べた p_{A} のみのモデルでは，NOUN ← VERB に対する重みが強くなった場合，二つの NOUN はどちらも VERB の子になってしまう．ここで正解の解析は NOUN ← NOUN ← VERB である．これに対し，もし $p_{\mathrm{S}}(stop|h, dir, adj)$ が正しく学習されれば，$p_{\mathrm{S}}(\mathrm{STOP}|\mathrm{VERB}, \leftarrow, \mathrm{FALSE})$ の値が大きくなり，英語の動詞(VERB)が左側に通常子を一つ，すなわち動詞に対する主語しか持たないことをモデル化できる．モデルはまた，英語の冠詞(DET)や代名詞(PRON)が通常左右に子を持たないことも捉えらえる．このためには，図 4(b)の 3 行目と 5 行目のパラメータがどれも高くなるように学習がされていれば良い．

パラメータ	具体例	説明
$p_S(stop \mid h, dir, adj)$	$p_S(\neg\text{STOP} \mid \text{VERB}, \rightarrow, \text{TRUE})$	VERB が右側に子を持たない状態から，一つ子を生成する
$p_A(d \mid h, dir)$	$p_A(\text{NOUN} \mid \text{VERB}, \rightarrow)$	右側の具体的な子として NOUN を選ぶ

図 3. DMV の二種類のパラメータとその具体例.



(a)

$$
\begin{aligned}
p(t) = {} & \underbrace{p_A(\text{VERB} \mid \$, \rightarrow)}_{\text{文のルートを VERB に}} \cdot \underbrace{p_S(\neg\text{STOP} \mid \text{VERB}, \rightarrow, \text{TRUE})}_{\text{VERB の最初の右の子を NOUN に}} \cdot \underbrace{p_A(\text{NOUN} \mid \text{VERB}, \rightarrow)}_{} \cdot \underbrace{p_S(\text{STOP} \mid \text{VERB}, \rightarrow, \text{FALSE})}_{\text{VERB の右の子の生成を停止}} \cdot \\
& \underbrace{p_S(\neg\text{STOP} \mid \text{VERB}, \leftarrow, \text{TRUE})}_{\text{VERB の最初の左の子を PRON に}} \cdot \underbrace{p_A(\text{PRON} \mid \text{VERB}, \leftarrow)}_{} \cdot \underbrace{p_S(\text{STOP} \mid \text{VERB}, \leftarrow, \text{FALSE})}_{\text{VERB の左の子の生成を停止}} \cdot \\
& \underbrace{p_S(\text{STOP} \mid \text{PRON}, \rightarrow, \text{TRUE}) \cdot p_S(\text{STOP} \mid \text{PRON}, \leftarrow, \text{TRUE})}_{\text{PRON は左右に子を持たない}} \cdot \\
& \underbrace{p_S(\text{STOP} \mid \text{NOUN}, \rightarrow, \text{TRUE})}_{\text{NOUN は右に子を持たない}} \cdot \underbrace{p_S(\neg\text{STOP} \mid \text{NOUN}, \leftarrow, \text{TRUE})}_{\text{NOUN の最初の左の子を DET に}} \cdot \underbrace{p_A(\text{DET} \mid \text{NOUN}, \leftarrow)}_{} \cdot \underbrace{p_S(\text{STOP} \mid \text{NOUN}, \rightarrow, \text{FALSE})}_{\text{NOUN の左の子の生成を停止}} \cdot \\
& \underbrace{p_S(\text{STOP} \mid \text{DET}, \rightarrow, \text{TRUE}) \cdot p_S(\text{STOP} \mid \text{DET}, \leftarrow, \text{TRUE})}_{\text{DET は左右に子を持たない}}
\end{aligned}
$$

(b)

図 4.（a）品詞に対する依存構造の例（灰色の単語は観測されない）．（b）この依存構造木の DMV のもとでの生成確率.

3.1.2 パラメータの学習

このモデルで特に重要なのは，パラメータの学習が 2.2 節で述べた内側外側アルゴリズムに基づく EM アルゴリズムで行えるという点である．これは DMV による生成過程が等価な PCFG によって表現可能であるという観察に基づく．図 5 に，PCFG の各ルールと DMV のパラメータの対応，そして CFG での解析例を示す．全ての書き換えルールは DMV のパラメータと一対一対応があり，これは PCFG となっている．本 PCFG は常に右の子を全て生成し，その後左の子を生成するが，この方向の固定により，CFG の解析と依存構造木とが一対一に対応する．

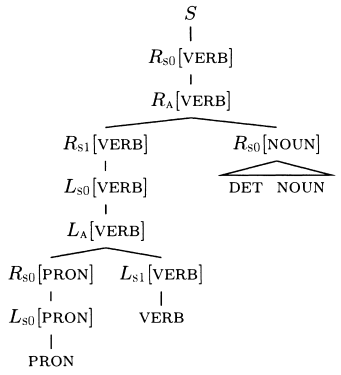
3.1.3 議論

英語での品詞列からの実験で，DMV は英語の基本的な語順を学習できることが示された．評価には，学習したモデルが人手で構築した正解の依存構造木の依存関係をどれだけ復元できたか，という精度を用い，これを文単位でなく単語単位で計算する．DMV は長さ 10 単語以下の文のみで評価した場合におよそ 44%の精度を達成した．英語では，非常に単純なベースラインとして，常に右隣の単語を子とすることで精度 34%が達成できることが知られていた．DMV はこの数値を初めて上回り，PCFG に対する EM アルゴリズムで意味のある構造が学習できることが示された.

評価の問題について少し触れておく．依存構造の教師なし解析の評価は様々な問題点が孕む

書き換えルール	DMV のパラメータ
$S \to R_{\mathrm{s}0}[h]$	$p_{\mathrm{A}}(h\|\$, \to)$
$R_{\mathrm{s}0}[h] \to L_{\mathrm{s}0}[h]$	$p_{\mathrm{s}}(\mathrm{STOP}\|h, \to, \mathrm{TRUE})$
$R_{\mathrm{s}0}[h] \to R_{\mathrm{A}}[h]$	$p_{\mathrm{s}}(\neg\mathrm{STOP}\|h, \to, \mathrm{TRUE})$
$R_{\mathrm{A}}[h] \to R_{\mathrm{s}1}[h]\ R_{\mathrm{s}0}[d]$	$p_{\mathrm{A}}(d\|h, \to)$
$R_{\mathrm{s}1}[h] \to L_{\mathrm{s}0}[h]$	$p_{\mathrm{s}}(\mathrm{STOP}\|h, \to, \mathrm{FALSE})$
$R_{\mathrm{s}1}[h] \to R_{\mathrm{A}}[h]$	$p_{\mathrm{s}}(\neg\mathrm{STOP}\|h, \to, \mathrm{FALSE})$
$L_{\mathrm{s}0}[h] \to h$	$p_{\mathrm{s}}(\mathrm{STOP}\|h, \leftarrow, \mathrm{TRUE})$
$L_{\mathrm{s}0}[h] \to L_{\mathrm{A}}[h]$	$p_{\mathrm{s}}(\neg\mathrm{STOP}\|h, \leftarrow, \mathrm{TRUE})$
$L_{\mathrm{A}}[h] \to R_{\mathrm{s}0}[d]\ L_{\mathrm{s}1}[h]$	$p_{\mathrm{A}}(d\|h, \leftarrow)$
$L_{\mathrm{s}1}[h] \to h$	$p_{\mathrm{s}}(\mathrm{STOP}\|h, \leftarrow, \mathrm{FALSE})$
$L_{\mathrm{s}1}[h] \to L_{\mathrm{A}}[h]$	$p_{\mathrm{s}}(\neg\mathrm{STOP}\|h, \leftarrow, \mathrm{FALSE})$

(a)



(b)

図 5.（a）DMV の PCFG での表現．（b）この文法による図 4(a)の依存構造木の変換．

未解決問題の一つとして知られている(Schwartz et al., 2011; Bisk and Hockenmaier, 2013)．例えば，上で挙げた NOUN NOUN VERB という列を考えると，複合名詞 NOUN NOUN の head がどちらか，というのは恣意的にしか決められないだろう．依存構造の分析からこのような人手による恣意性を取り除くことはできず，一つの正解データの基準をどれだけ復元できるかという評価がどれほど意味があるものなのかは疑わしい．本来は後段の処理で，つまり導出した依存構造木が，機械翻訳や情報抽出などの応用の精度向上にどれほど寄与するかで評価を行うことが客観的な価値判断に有効と考えられるが，そのような研究はほとんど見られない．

最後に DMV のもう一つの重要な貢献である EM の初期値について述べる．言語の重要な特性として，各依存関係の単語間の距離は近いものが好まれる，ということが知られている(Gildea and Temperley, 2010)．DMV ではこの直感を初期値を通じてモデルに埋め込んでいる．具体的には，最初に p_{A} を単語間の距離に反比例する量で定義した後，この正規化されていない分布の上で E ステップを行い，続く M ステップで初期値を得る．実際にはこの初期化が非常に重要であり，後の研究で，この初期化を行わないと精度が 20%以上低下(44%から 21%)することが指摘された(Gimpel and Smith, 2012)．

3.2 学習の工夫

DMV の一定の成功以降，これに対する様々な改良が提案された．1 節で述べたように，多くの研究は，文法に関する様々な仮定を置き，その影響を調べたものが多い．ここではいくつかの代表的な研究をかいつまんで説明する．他の方向性として，正解の品詞が付与されている前提で，品詞間のルールをモデルに埋め込む研究が存在する．一般にこれらのほうがより高い精度を示すが，そちらについては別に 3.3 節でまとめる．

Smith and Eisner (2006)は DMV が初期値を通じて取り込んでいる，短い距離の依存関係を好むという文法の傾向をより明示的にモデルに組み込んでいる．彼らのモデルでは DMV の $p_{\mathrm{A}}(d|h, dir)$ を次で置き換える．

$$p'_{\mathrm{A}}(d|h, dir) \propto p_{\mathrm{A}}(d|h, dir) \cdot e^{\beta|d-h|}$$

ここで $|d-h|$ は，h と d の間に存在する単語の数(距離)である．β (≤ 0) がハイパーパラメータとなっており，これがバイアスの強さを決定する．彼らは更にこの強さを学習中に減衰させるアニーリング法を提案しており，この学習法が様々な言語で有効であることを検証している．

Cohen and Smith (2009)と Berg-Kirkpatrick et al. (2010)は，どちらもモデルは DMV であるが，各パラメータを更に別の対数線形モデルで表現することで，パラメータ同士に相関を持たせている．これらのモデルでは，入力の品詞が与えられたとき，その品詞に対するより粗いカテゴリが利用できることを前提とする．例えば Berg-Kirkpatrick et al. (2010)では，$p_{\text{A}}(d|h, dir)$ を次のようにモデル化する．

$$p'_{\text{A}}(d|h, dir) \propto \exp(\boldsymbol{w} \cdot \boldsymbol{f}(d, h, dir))$$

$\boldsymbol{w}$ は対数線形モデルの重みベクトル，$\boldsymbol{f}$ は DMV のパラメータ毎に特徴ベクトルを抽出する関数である．英語の品詞体系では，代名詞と固有名詞は異なる品詞として区別されるが，両者の振る舞いは似ていることが想定できる．この直感は例えば，$\boldsymbol{f}$ に品詞が粗い名詞に属するか判定する要素を持たせることで，名詞に属する品詞間でパラメータのゆるい相関を持たせることができ，モデル化ができる．Cohen and Smith (2009)はほぼ似た枠組みを，ベイズモデルの事前分布(shared logistic normal prior)によって実現している．実験ではどちらも英語で 63%程度を示すことが分かっており，DMV と比べて大きな改善が行われた．なお，これまで述べたモデルは全て Klein and Manning (2004)の EM の初期値を利用していることに注意する．

Mareček and Straka (2013)は現時点で，品詞間のルールを直接用いない手法の中では最高精度を持つモデルである．これは，依存文法の dependent に対する直感をうまく大規模データから取り出しモデルに組み込んだ研究といえる．彼らが着目したのは句の削減可能性(reducibility)である(Mareček and Žabokrtský, 2012)．彼らは，別の語の dependent になる句は，その句を削除してもしばしば文法的に正しいという性質に着目し，Wikipedia の記事を用いて各品詞 n グラム毎に reducibility を計算，それを DMV の停止確率 p_{s} の計算に組み込んだ．彼らはこのように大規模データからの統計量をうまく利用することで，初期値を工夫せずとも学習がうまくいくことを報告している．

ここまでは，依存構造もしくは品詞に対する言語学的直感をモデルに組み込んだものといえるが，その他様々な学習上の工夫が提案されている．例えば Spitkovsky et al. (2010)は，EM の学習中に短い文から始め徐々に長い文を取り入れる方法，Headden et al. (2009)は大量のランダムな初期化で数回の EM の試行を行い，尤度の高いものを選択する方法，Blunsom and Cohn (2010)は DMV の文法(図 5)を CFG でなく木置換文法(tree substitution grammar)でモデル化することで，CFG の局所性を緩和することに成功している．

3.3 品詞間のルールの組み込み

本節では Naseem et al. (2010)を中心とした，品詞間のルールに対して弱い教師情報を組み込んだモデルについて紹介する．彼女らの手法は，事後分布正則化(posterior reguralization) (Ganchev et al., 2010)に基づき，名詞は動詞の子となりやすい，形容詞は名詞の子となりやすい，といった品詞間の直接的な関係をモデルに組み込む．

EM アルゴリズムは，隠れ変数 $\boldsymbol{z}$ の事後分布の更新(E ステップ)，パラメータ θ の更新(M ステップ)を交互に行う最適化と見なすことができる．まず次のように式(2.1)の下界が導出できる．

$$L(\theta) = \sum_{x \in \boldsymbol{x}} \log \sum_{z \in \mathcal{Z}(x)} p(x, z|\theta) = \sum_{x \in \boldsymbol{x}} \log \sum_{z \in \mathcal{Z}(x)} q(z) \frac{p(x, z|\theta)}{q(z)}$$
$$\geq \sum_{x \in \boldsymbol{x}} \sum_{z \in \mathcal{Z}(x)} q(z) \log \frac{p(x, z|\theta)}{q(z)} = F(q, \theta)$$

ここで二行目への変換は Jensen の不等式を用いた．通常の EM アルゴリズムは，下界 $F(q, \theta)$ を q, θ について交互に最適化する．E ステップでは，

$$q(z) \leftarrow \arg\max_{q(z)} F(q,\theta) = \arg\min_{q(z)} \mathrm{KL}(q(z)||p(z|x,\theta)) = p(z|x,\theta), \tag{3.1}$$

つまり，現在の θ を元に z の事後分布を決める．M ステップでは，

$$\theta \leftarrow \arg\max_{\theta} F(q,\theta)$$

を求めるが，PCFG のような多項分布の場合，これは式(2.2)のような期待値の正規化に帰着する．

事後分布正則化は，上記の E ステップを次のように変化させる．

$$q(z) \leftarrow \arg\max_{q(z)\in\mathcal{Q}(x)} F(q,\theta) = \arg\min_{q(z)\in\mathcal{Q}(x)} \mathrm{KL}(q(z)||p(z|x,\theta)).$$

つまり，事後分布 $q(z)$ の範囲を，特定の空間 $\mathcal{Q}(x)$ に制限する．これは式(3.1)のように閉じた形では解けないが，$\mathcal{Q}$ に凸空間を仮定することで最適解を求めることができる．Naseem et al. (2010) は $\mathcal{Q}$ として，$E_q[f(z)] \leq b$ という制約を用いている．ここで f は依存構造木 z が与えられたとき，事前に定めた品詞間のルールに**属さない**依存関係の割合，b はそのようなルールに属さない依存関係の割合の許される最小値を決めるパラメータである．

このルールを定める $f(z)$ がモデルの振る舞いを決定する．彼女らの実験では，名詞，動詞，形容詞など基本的な品詞に対して，NOUN→VERB などのルール(方向は定めない)を計 13 種類記述し，上記の $\mathcal{Q}$ に対して，b の値を 0.2，つまり 8 割の依存関係が(期待値の上で)定めたルールを満たす必要があると定め，学習を行った．

3.2 節で述べてきた様々な拡張は基本的にヒューリスティックスであり，特定の言語では逆に精度が悪化するなど，効果も言語によっては限定的なものが多かった．それに対しこの手法は，様々な言語を通じて，10 単語以下の文に対して 60%–70%の精度という安定した結果を示した．本研究で示されたのは，入力文の品詞が完全に同定されているという状況であれば，品詞間のルールをモデルに組み込むことで安定した精度が実現できる，という点である．ただしこの仮定が現実的なものかについては疑問が残る．品詞への依存性を高めるほど，高精度の品詞解析器を用意しなければ精度が達成されない，ということを意味するからである．4 節で紹介する Bisk and Hockenmaier (2013)のモデルはこの点を緩和したものといえ，彼らはより少ない言語学上の仮定から Naseem et al. (2010)と同程度の精度を達成することに成功している．

Naseem et al. (2010)に関連する研究として，Grave and Elhadad (2015)は似たアイデアを識別クラスタリング(Xu et al., 2005)の枠組みに適用し，教師なし構文解析をその上で定式化することで生成モデルによるアプローチよりも高い精度を実現できることを示している．

4. 組合せ範疇文法の学習

2.3 節で，句構造文法の EM アルゴリズムによる学習は，句構造の各記号の持つ意味が少ないためうまくいかなったことを述べた．組み合わせ範疇文法(CCG) (Steedman, 2000)などの語彙化文法はこれに対し，各非終端記号は統語的な振る舞いを決めるという点で意味を持っており，任意の記号ではない．この点に着目し，近年，少量の手がかりを人手で与えることでこれらの文法を学習する手法が研究され始めている．

図 6 に CCG による構文解析の例を示す．CCG の解析の基本単位はカテゴリであり，N や S\S などが属する．図では S → N S\N などのルールが存在するが，これは S\N の機能により決まる振る舞いで，このカテゴリは左側の別の N と結合することで S になるという意味を持つ．N/N は右側の N と結合し N になるという意味を持つ．CCG は多数のカテゴリと少数のこのような結合ルール(高々 10 個程度)からなる文法となっている．

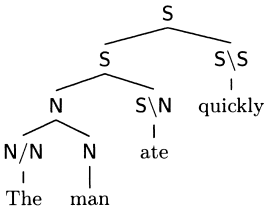


図 6. CCG による構文木.

The	man	ate	quickly
DT	NNS	VBD	RB
N/N	**N**, S/S	**S**, N\N	S\S
(S/S)/(S/S)	(N\N)/(N\N)	S\N	(N\N)\(N\N)
	(N/N)\(N/N)	(S/S)\(S/S)	
		(S\S)/(S\S)	

図 7. Bisk and Hockenmaier (2013)での品詞毎のカテゴリ候補の収集例. DT, NNS, VBD, RB は品詞である. 太字のカテゴリが最初に与えられる知識である.

CFG と異なり，CCG では各単語に割り当てられたカテゴリが統語上の大きな意味を持つ．例えば英語で read などの他動詞には (S\N)/N というカテゴリが割り当てられるが，これは右側の N(目的語)と結合し S\N となり，更に左側に N(主語)をとる，という情報が埋め込まれている．

Bisk and Hockenmaier (2013)はこのような CCG 木に対する生成モデルを提案し，これを教師なし学習することで多数の言語でこれまでの最高精度を達成することを報告している．評価に際しては，CCG の構文木はカテゴリの情報を読み取ることで依存構造木に変換できることを利用する．例えば N/N や S\S など，同じカテゴリを並べたカテゴリは修飾語として振舞うので結合した別の語の dependent とできる．

この手法で鍵となっているのは，訓練中の各単語に対するカテゴリの候補の与え方である．CCG などの語彙化文法は，入力文の各単語のカテゴリが定まると，その上に構築されうる構文木はほぼ決定されるという特徴を持つ(Matsuzaki et al., 2007; Lewis and Steedman, 2014)．言い換えると，統語上のほとんど全ての曖昧性は単語に対するカテゴリ割り当てに集約され，他の部分の曖昧性はほとんど存在しない．Bisk らは CCG のこの特性を，入力の品詞毎にあり得るカテゴリに制約をかけることで効率的に抽出している．彼らが前提として与える知識は，1)文のルートは動詞または名詞であること，そして 2)名詞は動詞の目的語となること，の二点である．この知識をモデルに与えるため，次のような方法で学習を始める前段階として，各品詞毎にあり得るカテゴリに制限をかける．まず，動詞に属する品詞のみに S のカテゴリを許し，名詞に属する品詞のみに N のカテゴリを許す．その後，この情報を元に，他の品詞のカテゴリ候補を拡大していく．例えば，訓練文中で S が割り当てられた単語(動詞)の左に隣接する品詞には S/S(S を左から修飾する)，右に隣接する品詞には S\S(S を右から修飾する)というカテゴリを候補として加える．このような処理を順次繰り返し，品詞毎にカテゴリの候補を拡大していく．図 7 に，さきほどの例文でのカテゴリ候補の拡大例を示す．図では例えば，VBD(動詞の過去形)にまず S が割り当てられ，また左に名詞(N)が存在することから S\N が追加され，更に NNS(名詞)に対して (N\N)/(N\N) という複雑なカテゴリが右側の VBD に対する N\N を基に生成されることなどが見てとれる．

この前処理の後，品詞を入力として，生成モデルのパラメータを変分ベイズ法で推定する．この学習の際に，文全体を張るカテゴリは S(動詞が存在しない場合 N)に制限される．この制限が本質的に重要であり，これによって，実質的に文のルートが動詞であるという制限をモデルに与え，効率的な学習を可能としている．

以上が大まかな枠組みであるが，その後の研究で Bisk and Hockenmaier (2015)はこのモデルを様々な方法で拡張し，詳細なエラー分析を行っている．また Bisk et al. (2015)では，この学習の枠組みが教師なし品詞推定の出力に対しても適用可能であることを示している．通常教師なし品詞推定は単語のクラスタリングを行うものであるため，各カテゴリが名詞に属するのか，形容詞に属するのか，などは分からない．彼らはこれに対し，手法が名詞と動詞の二つの品詞の同定にしか依存していないことから，これらを人手で選定することで最小の労力で手法が適用できることを示している．

Bisk and Hockenmaier (2015)のモデルの拡張，および分析は非常に丁寧に行われており，現時点での教師なし構文解析の限界を示しているものともいえるだろう．彼らの分析によると，現時点で解けていない問題の多くは，単語の**意味**に起因する問題であるという．例えばモデルは “I gave her a gift” という文に対し，“her a gift” を一つの名詞句とする判断をしたと報告している．英語の文法を知らない学習者からすると，“her” が “a gift” を修飾する形容詞と働く可能性も排除できないだろう．これに対し正しい構造，つまり gave が右に目的語を二つとるという構造を得るためには，モデルがこちらを好むような何らかの仕組みもしくはバイアスを外部から組み込む必要があるのではないかと考えられる．

最後に，CCG に基づく他の関連研究についても紹介しておきたい．Garrette et al. (2015)は，品詞を用いずに単語を直接入力とする CCG の学習を提案している．彼らは品詞の情報に頼る代わりに，いくつかの単語に対し，正解のカテゴリが付与された辞書の存在を仮定する．ここから EM 的な学習により，他の単語についてのカテゴリも順次学習することで，高精度を達成できることを示している．二つのアプローチの本質的な違いは，文法に対する事前知識の与え方である．彼らは辞書を利用するが，辞書の構築は人手がかかる作業であることを考えると，Bisk らの手法のほうが汎用性は高いといえるだろう．ただし，先に述べた gave の意味などの問題は，本手法のような直接的な知識の与え方によって解決できる可能性が高い．

5. おわりに

過去 20 年間の間の教師なし構文解析の進展について概説した．90 年代の単純な句構造の学習は失敗に終わったが，その後 Klein and Manning (2004)の品詞に基づく依存構造の学習で研究の方向性を示し，それが様々な方向で拡張された．依存構造の学習において特に重要な研究といえるのは，人手で与えた品詞間のルールを利用する(Naseem et al., 2010)であろう．これは精度の上では CCG に基づく Bisk and Hockenmaier (2013)とほぼ同等であり，現在の品詞または単語列からのみの学習のアプローチの限界を示しているともいえる．どちらも多言語を通して，精度は 6 割もしくは 7 割で頭打ちである．これは，教師なし構文解析は少量の言語学的仮定をモデルに課すことで言語毎の基本的語順を発見できるようにはなったが，単語の意味に起因するようなより深い分析が必要な構文については解くことが難しい状況であることを示唆している．

我々の目標は，そのような深い分析が必要な解析も扱える解析器を，最小の人手の労力によって構築する手段を確立することである．つまり，あらゆるタスク，言語について教師データである構文木などを付与するのは現実的でなく，より効率的な教師情報の与え方を確立したいのである．

過去 20 年間の教師なし構文解析の研究によって，表層的な入力のみからでは学習に限界があることが明らかになった．つまり，より実用的なシステムのためには，何らかの方法で外部知識をモデルに与えてやる必要がある．

このための一つの方向性は，構文解析を別のタスクのための隠れ変数とみなして学習を行う方法であろう．例えば Liang et al. (2011) は，質問応答という非常に限られたドメインに対してではあるが，質問文とその答えのみを入力として，質問文に対するデータベースのクエリを教師なしで学習するモデルを得ることに成功している．このようなアプローチとここで述べた教師なし構文解析との最大の違いは，教師なし構文解析では学習中の構造に対しフィードバックが与えられないという点である．つまり，モデルは構造を探索するが，その構造の言語的良し悪しは全く判断することができない．何らかのフィードバックが与えられれば，それが学習中に必要なバイアスとなりうる．もちろん，このような方法の制約は，モデル化や学習法がタスクに依存し汎用性が低下するという点である．汎用性を残しつつも，外部知識を効率的に取り入れ，深い分析も含めて教師なしに近い状況で学習を行える枠組みを探求するというのが，今後の研究の最大の課題であると考えている．

注．

1) 本稿では以後，依存構造木として図 1(b) のような依存関係の矢印の間に交差が存在しない構造を仮定する．依存構造木の PCFG への変換はこのような制限のもとにおいてのみ可能となる．実際の言語には交差を含む構造も出現するが，多言語にわたってそのような構造の頻度は小さいことが知られている (Kuhlmann, 2013)．

参 考 文 献

Berg-Kirkpatrick, T., Bouchard-Côté, A., DeNero, J. and Klein, D. (2010). Painless unsupervised learning with features, *Proceedings of Human Language Technologies*: *The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 582–590, Los Angeles, California.

Bisk, Y. and Hockenmaier, J. (2013). An HDP model for inducing combinatory categorial grammars, *Transactions of the Association for Computational Linguistics*, **1**, 75–88.

Bisk, Y. and Hockenmaier, J. (2015). Probing the linguistic strengths and limitations of unsupervised grammar induction, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (*Volume 1*: *Long Papers*), 1395–1404, Beijing, China.

Bisk, Y., Christodoulopoulos, C. and Hockenmaier, J. (2015). Labeled grammar induction with minimal supervision, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (*Volume 2*: *Short Papers*), 870–876, Beijing, China.

Blunsom, P. and Cohn, T. (2010). Unsupervised induction of tree substitution grammars for dependency parsing, *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 1204–1213, Cambridge, Massachusetts.

Carroll, G. and Charniak, E. (1992). Two experiments on learning probabilistic dependency grammars from corpora, *Working Notes of the Workshop Statistically-based NLP Techniques*, 1–13, AAAI Press, Palo Alto, California.

Charniak, E. (1996). Tree-bank grammars, *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, 1031–1036.

Chomsky, N. (1986). *Knowledge of Language. Its Nature, Origin, and Use*, Praeger Publications, New York.

Clark, A. (2001). *Unsupervised Language Acquisition*: *Theory and Practice*, Ph.D. Thesis, School of Cognitive and Computing Sciences, University of Sussex.

Cohen, S. and Smith, N. A. (2009). Shared logistic normal distributions for soft parameter tying in unsupervised grammar induction, *Proceedings of Human Language Technologies*: *The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 74–82, Boulder, Colorado.

Collins, M. (1997). Three generative, lexicalised models for statistical parsing, *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, 16–23, Madrid, Spain.

Ganchev, K., Graa, J., Gillenwater, J. and Taskar, B. (2010). Posterior regularization for structured latent variable models, *Journal of Machine Learning Research*, **11**, 2001–2049.

Garrette, D., Dyer, C., Baldridge, J. and Smith, N. (2015). Weakly-supervised grammar-informed bayesian ccg parser learning, *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence* (*AAAI-15*), Austin, Texas.

Gildea, D. and Temperley, D. (2010). Do grammars minimize dependency length?, *Cognitive Science*, **34**(2), 286–310.

Gimpel, K. and Smith, N. A. (2012). Concavity and initialization for unsupervised dependency parsing, *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, 577–581, Montréal, Canada.

Gormley, M. R. and Eisner, J. (2013). Nonconvex global optimization for latent-variable models, *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), 444–454, Sofia, Bulgaria.

Grave, E. and Elhadad, N. (2015). A convex and feature-rich discriminative approach to dependency grammar induction, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (*Volume 1*: *Long Papers*), 1375–1384, Beijing, China.

Headden III, W. P., McClosky, D. and Charniak, E. (2008). Evaluating unsupervised part-of-speech tagging for grammar induction, *Proceedings of the 22nd International Conference on Computational Linguistics* (*Coling 2008*), 329–336, Manchester, U.K.

Headden III, W. P., Johnson, M. and McClosky, D. (2009). Improving unsupervised dependency parsing with richer contexts and smoothing, *Proceedings of Human Language Technologies*: *The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 101–109, Boulder, Colorado.

Hsu, D. J., Kakade, S. M. and Liang, P. S. (2012). Identifiability and unmixing of latent parse trees, *Advances in Neural Information Processing Systems*, **25** (eds. F. Pereira, C. J. C. Burges, L. Bottou and K. Q. Weinberger), 1511–1519, Curran Associates, Inc., Redhook, New York.

Johnson, M., Griffiths, T. and Goldwater, S. (2007). Bayesian inference for PCFGs via Markov chain Monte Carlo, *Proceedings of Human Language Technologies 2007*: *The Conference of the North American Chapter of the Association for Computational Linguistics*, 139–146, Rochester, New York.

Kasami, T. (1965). An efficient recognition and syntax-analysis algorithm for context-free languages, Technical Report, AFCRL-65-758, Air Force Cambridge Research Lab., Cambridge, Massachusetts.

Klein, D. and Manning, C. (2004). Corpus-based induction of syntactic structure: Models of dependency and constituency, *Proceedings of the 42nd Meeting of the Association for Computational Linguistics* (*ACL'04*), *Main Volume*, 478–485, Barcelona, Spain.

Kuhlmann, M. (2013). Mildly non-projective dependency grammar, *Computational Linguistics*, **39**(2),

355–387.

Lari, K. and Young, S. (1990). The estimation of stochastic context-free grammars using the inside-outside algorithm, *Computer Speech & Language*, **4**(1), 35–56.

Lewis, M. and Steedman, M. (2014). A* CCG parsing with a supertag-factored model, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), Doha, Qatar.

Liang, P. and Klein, D. (2008). Analyzing the errors of unsupervised learning, *Proceedings of ACL-08*: *HLT*, 879–887, Columbus, Ohio.

Liang, P., Jordan, M. and Klein, D. (2011). Learning dependency-based compositional semantics, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*: *Human Language Technologies*, 590–599.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*, The MIT Press, Cambridge, Massachusetts.

Mareček, D. and Straka, M. (2013). Stop-probability estimates computed on a large corpus improve unsupervised dependency parsing, *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), 281–290, Sofia, Bulgaria.

Mareček, D. and Žabokrtský, Z. (2012). Exploiting reducibility in unsupervised dependency parsing, *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 297–307, Jeju Island, Korea.

Matsuzaki, T., Miyao, Y. and Tsujii, J. (2007). Efficient HPSG parsing with supertagging and CFG-filtering, *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence* (*IJCAI-07*), 1671–1676.

Naseem, T., Chen, H., Barzilay, R. and Johnson, M. (2010). Using universal linguistic knowledge to guide grammar induction, *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 1234–1244, Cambridge, Massachusetts.

Pereira, F. and Schabes, Y. (1992). Inside-outside reestimation from partially bracketed corpora, *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, 128–135, Newark, Delaware, U.S.A.

Schwartz, R., Abend, O., Reichart, R. and Rappoport, A. (2011). Neutralizing linguistically problematic annotations in unsupervised dependency parsing evaluation, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*: *Human Language Technologies*, 663–672, Portland, Oregon, U.S.A.

Smith, N. A. and Eisner, J. (2006). Annealing structural bias in multilingual weighted grammar induction, *Proceedings of the International Conference on Computational Linguistics and the Association for Computational Linguistics* (*COLING-ACL*), 569–576, Sydney, Australia.

Spitkovsky, V. I., Alshawi, H. and Jurafsky, D. (2010). From baby steps to leapfrog: How "less is more" in unsupervised dependency parsing, *Human Language Technologies*: *The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 751–759, Los Angeles, California.

Steedman, M. (2000). The Syntactic process, *Language, Speech, and Communication*, MIT Press, Cambridge, Massachusetts.

Xu, L., Neufeld, J., Larson, B. and Schuurmans, D. (2005). Maximum margin clustering, *Advances in Neural Information Processing Systems*, **17**, 1537–1544, MIT Press, Cambridge, Massachusetts.

Younger, D. H. (1967). Recognition and parsing of context-free languages in time n^3, *Information and Control*, **10**(2), 189–208.

160 Proceedings of the Institute of Statistical Mathematics Vol. 64, No. 2, 145–160 (2016)

Statistical Models to Induce Latent Syntactic Structures

Hiroshi Noji

Graduate School of Information Science, Nara Institute of Science and Technology

This article describes the advancement of unsupervised syntactic parsing in the past 20 years. Unsupervised parsing aims to obtain the grammar of the language automatically from the input sentences without manually created syntactic trees. The essential point in this task is how to exploit the bias or knowledge of the grammar of the language. In this article, we compare several existing approaches from this perspective and discuss what kind of information we should provide to the model and what can be learned from such knowledge, to guide the future research direction on this area.

Key words: Computational linguistics, unsupervised syntactic parsing.

統計数理(2016)
第 64 巻 第 2 号 161–178

特集「統計的言語研究の現在」
[研究詳解]

言語変化と系統への統計的アプローチ

村脇 有吾†

(受付 2016 年 3 月 3 日；改訂 10 月 7 日；採択 10 月 7 日)

要　　旨

言語変化や諸言語の系統関係の解明といった歴史言語学の課題は，従来は言語学者が人手により取り組んできたが，21 世紀に入る前後から，計算機を用いた統計的手法を適用する事例が増えている．もともと分子生物学分野で開発され，近年言語データに適用されるようになった統計的手法は，年代のような連続値を含んでいたり，不確実性が解候補の組合せ爆発を生むなどの理由から人間が苦手としてきた問題に取り組むことを可能にしつつある．本稿の前半では，特に重要な手法である語彙を手がかりとしたベイズ系統モデルについて，歴史言語学の研究経緯を踏まえつつ，統計的な観点から解説する．ただし，語彙を手がかりとする手法は，インド・ヨーロッパ語族のような既知の語族に対しては一定の成果を上げつつあるが，日本語はモデルを適用する基盤が整っていない．そこで，本稿の後半では，日本語系統論を解決に導く可能性のある手がかりとして言語類型論の特徴に着目する取り組みについて紹介する．

キーワード：言語系統樹，歴史言語学，言語類型論，ベイズ統計．

1. はじめに

我々が話している言語が歴史的にどのような変化を経てきたか，複数の言語が歴史的にどのような関係を持つかといった問題に取り組む分野を歴史言語学とよぶ．言語は人間集団を特徴づける主要な要素であるため，千年のオーダの人類史(例えばヨーロッパにおける人類の定住過程)を解明する上で，歴史言語学は重要な役割を果たす．そのため，人類史の仮説として，歴史言語学だけでなく，集団遺伝学や考古学などの諸分野の知見と整合的なものを求める学際的研究も近年盛んである．

歴史言語学では，21 世紀に入る前後から，計算機を用いた統計的手法を適用する事例が増えている(Forster and Renfrew, 2006)．その特徴は，分子生物学分野で開発されたモデルを言語データに適用することによって主要な成果が得られていることである．そもそも，Darwin (1859)の『種の起源』と Schleicher (1853)によるインド・ヨーロッパ(印欧)語族の系統樹が 19 世紀半ばの同時期に発表されたことに象徴されるように，草創期の進化研究においては，生物と言語の類似性が意識されていた．しかし，その後の両分野の研究は目立った交流がないまま進んだ．統計的手法の適用という点では，生物学分野では 20 世紀後半に順調に研究が進展したのに対して，言語学においては人手による研究手法が主流であり続けた．こうしたなか，生物向けに開発された統計モデルが 1990 年代末から徐々に言語データに適用されはじめ，特に Gray and Atkinson (2003)が印欧祖語の年代推定にベイズ系統モデルを適用したことが大きな話題となった．これがきっかけとなり，系統モデルやその他の生物学由来の統計モデルを言語研究に導入

†京都大学大学院 情報学研究科：〒 606–8501 京都市左京区吉田本町

する事例があいついでいる．

そこで，本稿の前半では，近年の統計的言語研究の中心となっているベイズ系統モデルを紹介する．この統計モデルは，手がかりとして語彙を用いるという点で，伝統的な比較法や，先行する統計的手法である言語年代学と共通しており，実際，これらの研究成果の上に成り立っている．そのため，まずは歴史言語学の従来研究を統計という観点から整理し，その上でベイズ系統モデルを導入する．

なお，歴史言語学的課題に対する統計的取り組みについては，既に言語学者向けの丁寧なチュートリアル(Nichols and Warnow, 2008)が公表されている．しかし，このチュートリアルはベイズ系統モデルの中身についてはほとんど触れていない．一方，モデルの中身については，ベイズ系統推定ソフトウェア BEAST の開発者による網羅的な解説本(Drummond and Bouckaert, 2015)が出ている．しかし，この本は分子生物学のデータを想定しており，言語データは一切登場しない．そこで，本稿では，言語データに軸足を置いてモデルを解説したい(分子生物学における対応物には適宜関連づける)．

現在のベイズ系統モデルには，語彙に基づいた手法であるという点に限界がある．この手法が一定の成功を収めているのは，伝統的な比較法によって大まかな系統関係が既に明らかになっている言語群であり，印欧語族やオーストロネシア語族(台湾から東南アジア島嶼部，太平洋に広がる大語族)などが該当する．一方，日本語と他の言語の系統関係については，100 年以上にわたる諸研究者の尽力にもかかわらず，依然として不明のままである．したがって，このモデルを適用しようにも，必要な基盤が整っていない．

語彙に代わる手がかりとして，筆者は言語類型論の諸特徴に着目しており，本稿の後半ではこれを紹介する．類型論に関する素朴な議論は，比較法と同程度に古くから見られるが，類型論の系統推定への応用は確立されていない．その大きな理由は，語彙とくらべて，類型論の特徴が手がかりとして不確実であり，人手による論証になじまなかったことだと筆者は考えている．そして，この問題は計算機を用いた統計的手法により克服できると見込んでいる．まだまだ始めたばかりの研究だが，これまでの経過と今後の展望を述べたい．

なお，筆者は歴史言語学の体系的な教育を受けた者ではないことを断っておきたい．筆者の専門は計算言語学・自然言語処理とよばれ，この分野では，統計・機械学習を用いた研究が盛んに行われているものの，機械翻訳や質問応答といった応用処理や，それを支えるテキスト解析技術の開発が中心であり，歴史言語学的課題についてはほとんど認知されていない．それでも本稿を書くのは，筆者や本稿の読者にも参入の余地があることを訴えたいからである．発表文献に占める生物系の論文誌の多さが示すように，生物学を背景とする研究者が研究を主導する例が多いが，彼らにとって，生物向けのモデルをそのまま言語データに転用するのではなく，言語独自のモデルを開発する動機はとぼしい．一方の言語学者は，当然ながら言語現象を深く分析し，その性質を考察している．しかし，主流研究が統計とは無縁であったこともあり，統計モデルを用いて問題を解くという発想を受け入れること自体に困難が見られる(Pereltsvaig and Lewis, 2015)．生物学由来のモデルを受け入れた一部の言語学者についても，ほとんどの場合に生物データ向けの既存のソフトウェアパッケージに依存している．そのため，生物に対応物が存在しない言語現象がモデル化から取り残される傾向にある．見方を変えれば，この断絶を埋める研究にいち早く取り組むことで，大きな成果が得られると期待している．

2. 基本的な概念とデータ

2.1 進化と系統樹

言語(や生物種)はそれ自体が複雑なシステムであることから，一部の側面を特徴として取り

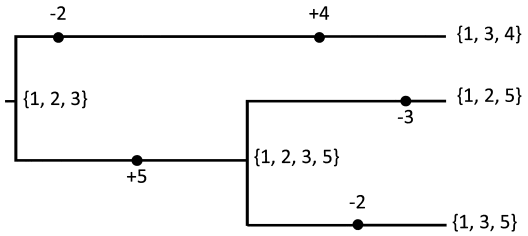


図 1．系統樹の例．

出して分析に用いる．例えば，生物種であれば，羽の有無や歯の形といった形態的特徴や，ゲノムの塩基配列，そこから取り出した STR（縦列型反復配列）や SNP（一塩基多型）などの特徴が用いられる．言語の場合は，語彙的特徴（ある語を持つか否か）が広く用いられているが，本稿ではほかに類型論的特徴にも着目する．特徴にはキリンの首の長さのような連続値もありえるが，本稿では離散値のみを扱う．

いま，言語（種）について，それが持つ諸特徴を取り出し，便宜的に $10110\cdots$ のように並べた列を状態とする．ここでは，簡単のために，ひとまず 2 値特徴を考えている．このとき，特徴に番号を振り，1 が立っている特徴のみに着目することで，同じ特徴を $\{1,3,4,\ldots\}$ のように集合的に表現することも可能である．

進化とは，親から子へと途切れなく状態が受け継がれるが，完璧に複製されるのではなく，次第に変化が蓄積する現象を指す．生物種の場合は遺伝により，言語の場合は第一言語習得により状態が継承される．なお，進化は変化を伴う由来（descent with modification）ともよばれるように，価値判断を含まず，単に変化するという現象のみに着目している．2 値特徴の場合，変化は特徴の誕生，死亡の 2 種類からなる．それぞれ $+4$，-2 のように表すとする．

系統樹は，図 1 のように，複数の言語（種）の歴史的関係を木構造により要約したものである．系統樹を特徴づけるのは分岐である．2 つの言語は，分岐前は完全に同一であり，分岐後は独立に進化するとしている．系統樹には，トポロジーのみを表した時間なし系統樹と，各ノードに年代が紐づいた時間つき系統樹がある．

通常の問題設定では，現代語や文献に記録された古代語，つまり系統樹の葉ノードの状態が観測されている．推定すべきは，観測されていない部分，つまり木のトポロジー，および祖先（祖語）の状態や年代などである．伝統的な歴史言語学では，時間なし系統樹が人手により推定されてきたが，統計的手法では，時間つき系統樹の推定を目的とすることが多い．なお，言語の場合，疎遠な言語同士の関係は一般に不明のままである．共通祖語を持つことが立証できたとき，系統関係が確立されたと言う．

葉ノードの状態が得られたとき，時間なし系統樹相当のものは素朴な手法でも作ることができる．まず，言語対に対して適当に距離が定義できる．仮に変化の速度がほぼ一定と仮定すると，言語対の距離は分岐後の時間におおよそ比例する．したがって，距離に基づく階層的クラスタリングを適用すれば木が得られる．より複雑なモデルを適用する場合でも，推定される系統樹はクラスタリング結果と大まかには似たものとなる．

ただし，系統推定のさまたげとなる厄介な現象がいくつか知られている．系統樹は分岐後の独立進化を仮定しているが，実際には，言語同士の接触により特徴が変化する場合がある．この現象は生物系の用語で水平伝播（horizontal/lateral transmission）とよばれる．言語の場合は語彙の借用が典型例として挙げられる．また，同じ特徴が系統樹上の複数箇所で誕生したり，一度死亡した特徴が復活することも考えられる．これらの現象をそれぞれ成因的相同（homoplasy），

復帰突然変異(back mutation)とよぶ．このような現象が生じうる場合，2 つの言語が同じ特徴を有していたとしても，それが祖語に由来するとは限らない．

2.2 データベースとその整備

統計的研究を行うには，準備として特徴のデータベースの整備が不可欠である．生物データ，特に集団遺伝学で用いられるゲノムデータと比較したとき，言語データの特性として，規模の小ささと高コスト性が挙げられる．

規模については，ヒトゲノムの SNP の場合，10 万のオーダの特徴が得られる(International HapMap Consortium, 2005)．個体数についても，千あるいはそれ以上のオーダで得られ，その数は今後も増え続けると見込まれる．さらに，各個体には，日本人，サルデーニャ人，ヨルバ人といった集団ラベルを割り振ることができる．つまり，日本人という集団は複数の個体によって表現され，個体間のばらつきが進化史の解明の手がかりとして利用できる．

一方，言語の場合，語彙的特徴，類型論的特徴のいずれについても，数は百のオーダにすぎない．比較可能な言語数も，語彙的特徴では十から百，類型論的特徴でも千のオーダで頭打ちである．さらに，言語は集団ごとに 1 つ採取される．言語はコミュニケーションの手段であり，発信者と受信者の双方が理解できなければならないため，集団内での大きなばらつきが期待できないからである．こうした制約から，集団遺伝学で近年発展した諸手法は言語データには適用できない場合が多い．代わって適用されるのは，ヒトとチンパンジーとの共通祖先の年代推定のようなよりマクロな比較や，進化の速度が桁違いなウイルスの系統推定に用いられてきた手法である．

言語データベース整備の高コスト性は，次世代シーケンサのような機械的手段ではなく，もっぱら言語学者が人手でデータを作成していることに起因する．一つの言語の習得だけでも何年もの時間を要するなか，複数の言語から斉一な基準にしたがってデータを採取するには高度な専門知識が欠かせない．おまけに，現状ではデータ整備の機械化の見通しは立たない．言語の話者数には著しい不均衡があり，数千とも言われる世界言語の大半は，電子化されたテキストどころか文字すら確立されていない小言語だからである．そして，そうした小言語が系統推定に重要な役割を果たすことが少なくない．統計・機械学習分野では，音声認識研究の大家 Fred Jelinek のものとされる，「言語学者をクビにするたびに音声認識器の精度が上がる」という発言が知られているが，歴史言語学では，統計的手法を用いる場合でも，依然として言語学者の貢献が欠かせない．

統計や言語処理の研究者にとって好都合なことに，データ整備という高い参入障壁は引き下げられつつある．統計的研究と並行して，2000 年代以降，言語データの整備と公開も急速に進んでいる．特に，マックス・プランク進化人類学研究所をはじめとするマックス・プランクの研究所群からは，後述の WALS (Haspelmath et al., 2005)と APiCS (Michaelis et al., 2013)のほか，Glottolog (Hammarström et al., 2016)などの有用な言語データが公開されている．データ公開による共有には副作用もある．参入障壁の引き下げは，その性質を深く理解しないままデータを扱う研究を生み出す危険がある．実際，類型論のデータベース(Haspelmath et al., 2005)の公開後，言語の特徴と人間や環境の特徴との相関を探す怪しげな研究が数多く発表されている(Roberts and Winters, 2013)．しかし，こうした問題を懸念してデータの公開を控えるよりも，通常の科学的批判を通じて淘汰を行う方が健全だと筆者は考えている．

3. 語彙に基づく系統推定の従来手法

3.1 比較法

歴史言語学の伝統的な手法や，近年のベイズ系統モデルの多くは，言語間の系統推定に語彙

的特徴を用いる．語彙に基づく系統推定の基盤は記号の恣意性である．DOG という意味と「いぬ」という音の結びつきに必然性はない．このことから，DOG を意味する「いぬ」という語が歴史上無関係に複数回発生する可能性は極めて低い．つまり，成因的相同や復帰突然変異の可能性が排除できる．ただし，接触による借用は起こりえる．また，語そのものは引き継いでいても，語形は時間とともに変化するため，ある言語対が持つ語が同一特徴か否かは自明ではない．「名前」と *name*，「骨」と *bone* のような偶然の類似や借用を排除し，祖語から引き継いだ特徴であることを立証する必要がある．そうした語を同源語(cognate)とよぶ．

比較法(英語でも単に comparative method とよばれる)では，同源語の特定に音法則を用いる．歴史的な音の変化は，例外なく規則的に起こることが知られている．結果として，ある言語対が持つ同源語の語形には規則的な音対応が見られる．こうした音対応を立証することで，偶然の類似や借用を排除し，同源語を特定できる．

言語間の系統関係が確立するためには，通常は百のオーダの同源語を特定する必要がある．ただし，重要なのは量そのものではなく，音法則を通じて特徴の特定の質を担保することである．結局のところ，比較法は，対象となる諸言語が祖語から同一特徴を受け継いでいることを示すにすぎない．なお，3 個以上の言語の系統関係については，分岐の前後関係を明らかにする必要がある．そのために，例えば，語彙ではなく個々の音変化そのものを特徴とみなし，言語間での特徴の共有を調べるといったことが行われる(Pellard, 2009)．

歴史言語学では，さらに祖語の語形の再構が試みられるが，観測できる手がかりは不完全であり，再構形は理論上の産物という側面が強い．とはいえ，この作業は語形という記号列の操作であり，人手による論証と親和性が高い．計算機を用いた統計的祖語再構も提案されているが，歴史言語学の成果を追認するにとどまっており，言語学上の新たな知見はとぼしい(Bouchard-Côté et al., 2013)．また，成功事例が報告されているのは，600 以上の言語からなる世界的にも稀な大語族(オーストロネシア語族)のみであり，小規模データに対してはうまく働かないのではないかと推測される．

また，比較法だけでは祖語の年代は推定できない．人間は年代のような連続値を直接扱うのが苦手であり，統計的手法の出番となる．

3.2 言語年代学

これに対し，言語年代学(glottochronology)は，その名の通り，言語対の祖語の年代を推定する統計的手法である．この手法は，考古学における放射性炭素年代測定に触発されて生まれたもので，生物の体内にわずかに含まれる放射性炭素が死後一定割合で減衰していくのと同様に，祖語にあった語彙的特徴も一定割合で失われると仮定する(Swadesh, 1952)．言語年代学の研究は 1940 年代末から 50 年代を中心に行われており，生物学における分子時計(molecular clock)仮説(Zuckerkandl and Pauling, 1965)に先行する．

言語年代学では，準備として基礎語彙を設定する．基礎語彙とは，どんな言語でもそれを表す言葉があるような基本的な概念(100 から 200 項目)である．例えば，WATER, BIG, EYE などが該当し，SNOW のように地域が限定される概念や，MILLION や PAPER のような文化語彙は排除される．また，基礎語彙は借用されにくく，比較的変化しにくいと仮定される．例えば，ある調査によると，英語の一般語彙はおおよそ 50%が借用語だが，基礎語彙に限ると 6%にすぎない(Swadesh, 1952)．各言語について基礎語彙を収集し，次に同源語を特定する．特定には上述の比較法が用いられる．

言語年代学における目標は，言語対 A, B について，それらの祖語 P の年代 t を推定することである．ここでの仮定は，P が持っていた基礎語彙が時間とともに一定割合で失われるというものである．A, B の基礎語彙共有率を c とすると，基礎語彙の残存率 r が与えられたとき，祖

語の年代は

$$t = \frac{\log c}{2 \log r}$$

と求まる(分母の 2 は，P から A，P から B の 2 本の枝に対応)．残存率 r は文献が豊富な言語群を用いて推定しておく．例えば，$r = 0.81$(200 項目基礎語彙で単位は千年)とすると，共有率 $c = 0.43$ のとき，$t \approx 2$(千年)が得られる．

言語年代学は歴史言語学に統計を持ち込んだ先駆的な研究だが，言語学者の間では評判が悪い．特に批判が集中したのは残存率一定という仮定である．例えば，極端な保守性で知られるアイスランド語の場合，古ノルド語と比較すると，残存率が 0.95 を超えており，0.81 という仮定からかけ離れている(Bergsland and Vogt, 1962)．

こうして言語年代学は激しい批判にさらされて衰退し，1990 年代末から 2000 年代にかけて，後発の分子生物学由来の手法で置き換えられることとなった．しかし，基礎語彙データの収集という面では貢献が大きく，近年の統計的研究も言語年代学(より一般には語彙統計学)の成果に依存している．

4. ベイズ系統モデル

言語年代学に代わって 2000 年頃から言語に適用され始めたベイズ系統モデルは，もとは分子生物学のデータを解析するために開発されたものである．分子生物学においても，当初は素朴な階層的クラスタリング手法や，確率モデルを用いる場合でも最尤推定法が用いられてきたが，1990 年代後半からベイズ系統モデルが盛んに研究されるようになった(Huelsenbeck and Ronquist, 2001)．ベイズ系統モデルの利点として，生成モデルであることから結果の解釈が容易なこと，様々な事前知識を柔軟に組み込めること，Markov chain Monte Carlo(MCMC)という理論的裏付けのある推論手法が存在することが挙げられる．

なお，ベイズ系統モデルの言語データへの応用事例として(Gray and Atkinson, 2003)が有名だが，これ以前にも(Gray and Jordan, 2000)が生物学由来の統計的系統モデルを言語データに適用している．ただし，この研究で用いた系統モデルはベイズ以前の階層的クラスタリング手法であり，年代推定は行っていない．

4.1 確率的解釈

ベイズ系統モデルは大掛かりなモデルであることから，準備として，より単純な言語変化の確率的解釈を考える．変化はメトロノームのように一定間隔で起きるのではなく，確率的ゆらぎがあると思われる．仮に一定のものがあるならば，それは確率分布のパラメータであると仮定するのが自然である．

実際のベイズ系統モデルでは，単語の誕生，死亡という 2 種類の変化をモデル化するが，簡単のために，まずは種類に関わりなく変化の回数を数えることとし，その回数と時間との関係をモデル化する．連続時間において独立に発生する離散事象を数えるモデルとしてポワソン過程(Kingman, 1993)が知られている．N_t を時間幅 $[0, t]$ で起きた独立な変化の数とすると，これはポワソン分布，$P(N_t = k) = \frac{e^{-\mu t}(\mu t)^k}{k!}$ に従う．また，2 つの連続した事象の間隔は指数分布 $e^{\mu t}$ に従う．ここではパラメータ μ を変化率とよぶことにする．

ポワソン過程からの複数の試行を図 2 に示す．変化率 μ が一定であっても，一定数の変化に要する時間に確率的ゆらぎがあることが確認できる．上述の言語年代学では，言語対を入力とし，共通祖語からの経過時間 t を点推定していた．これに対し，確率的解釈では，経過時間 t を含む一連の変化をモデルに与え，その確からしさを確率で返す．ある回数の変化が起きるのに要する時間は，点ではなく確率分布により表現される．つまり，考えられる様々な可能性に対

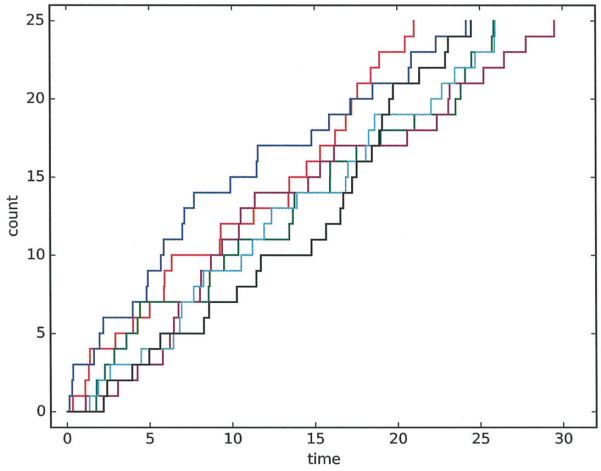


図 2. ポワソン過程からの試行 ($\mu = 1$).

して，その自然さを確率によって評価していることになる．これは，言語対にとどまらず，複数の言語を含む系統樹のモデルを設計する上で重要な性質である．系統樹は複雑な構造であり，局所的には最適でない解釈が全体的な整合性を考えると良いことがありえるが，こうした場合への柔軟な対応が可能となる．

さらに，ベイズモデルにおいては，パラメータ μ に事前分布を設定し，

$$P(\text{時間つき変化過程}, \mu; \alpha) \propto P(\text{時間つき変化過程}\ |\mu) \times P(\mu; \alpha)$$

とすることで，パラメータ割り当てにも確率を与える．後で見るように，系統樹は多くの潜在変数を含むが，パラメータに事前分布を置くことで，パラメータをその他の潜在変数と統一的に扱え，推論時に都合が良い．

掛け算は対数化すると足し算となる．例えば，時間つき変化過程の対数確率は

$$\log P(\text{時間つき変化過程}\ |\mu) + \log P(\mu; \alpha) + C$$

となり，定数項 C を無視すれば，対数確率というスコア 2 つの足し算によって時間つき変化過程の自然さが採点されていることになる．系統樹はより複雑な部分モデルの組み合わせによって構成されるが，基本は同じである．部分モデルは系統樹の構成要素の自然さをスコアによって採点するものであり，それらのスコアを合算すれば系統樹全体のスコアとなる．

4.2 モデル設計と推論

ベイズ系統モデルは，時間つき系統樹をモデル化する．上述の確率的解釈と同様に，モデルは $P(\text{時間つき系統樹}, \theta; \alpha)$ と表される．ここで，θ はモデルのパラメータ群，α はハイパーパラメータ群を表す．パラメータを含む系統樹は複雑な構造であり，いくつかの構成要素（部分モデル）に分解することで構成される．自然な系統樹に相対的に高いスコア（対数確率）を与えるようなモデルの設計が最初の目標となる．ベイズ系統モデルは図 1 のような系統樹を直接スコアで評価しており（後述のように正確には異なる），結果を素直に解釈できることが魅力の一つである．

次に，与えられたモデルのもとで，高いスコアを返すような系統樹とパラメータの組を探す．

この手続きを推論とよぶ．系統樹のうち，葉ノードの状態と年代は観測されている．さらに，いくつかの中間ノードの年代や，場合によっては部分木もモデルに与える．残りは潜在変数であり，連続値を含むことから非加算無限個の候補があるが，とにかく一通り値を割り当てれば，モデルからスコアが得られる．

4.3 部分モデル

部分モデルは現在でも活発に研究されており，その組み合わせには多数の変種がある．部分モデルは大きくは木モデル，置換モデル，時計モデルからなり，他にも各種パラメータの事前分布がある．

木モデルは，ノードの状態を無視し，時間つき木の骨組みを採点する．モデルの例として，Yule 過程，誕生・死亡モデル(birth-death model)，Bayesian skyline モデル等が知られている．しかし，生物学上の問題意識から開発されており，言語系統樹における意味はあまり明らかにされていない．

木モデルに関してむしろ重要なのは，年代較正(calibration)である．内部ノード(例えばインド・イラン祖語)や葉ノード(例えばラテン語)の年代について既知であることを他のノードの年代推定に利用する．生物の場合は主に化石の年代を用いる．年代は点で与えることも可能だが，多くの場合は正規分布のようなソフトな制約を設定する．このとき，平均から離れた年代を該当ノードに割り当てるほどスコアが減点されることになる．例えば，ラテン語の年代の事前分布を $\mathcal{N}(\mu = 2050.0, \sigma = 75.0)$ と置くと，2050BP(before present)から 75 年離れると 0.5，150 年離れると 2.0 の減点(いずれも対数値)が課される．結果として，推論時には，こうしたソフトな制約を満たすような変化率が推定される．

置換モデルは，親から子への状態変化を採点する．ポワソン過程では変化の数を数えたが，置換モデルは，それぞれの 2 値特徴の具体的な値の変化をモデル化する．図 1 のように親から子への枝のどの時刻で変化が起きたかを陽に持つことも可能だが，効率化のために始点と終点の両ノードの値のみに着目する．すなわち，ある親言語のある特徴の値が $x' = i \in \{0, 1\}$ のとき，時間 t 後の言語の特徴 x の値が j である確率 $P(x = j|x' = i, t)$ を設計したい．これは i から始まり，時間 t 後に j となるすべての遷移の確率を積分したものである．

置換モデルは，通常，連続時間マルコフ連鎖によってモデル化される．連続時間マルコフ連鎖はパラメータとして遷移率行列 Q を持つ．2 値特徴の場合，$Q = \left(\begin{smallmatrix} -\alpha & \alpha \\ \beta & -\beta \end{smallmatrix}\right)$，という 2×2 の行列で表される．ここで，$\alpha \geq 0$ は $0 \to 1$ の変化の起こりやすさ，$\beta \geq 0$ は $1 \to 0$ の変化の起こりやすさを表す．遷移率行列 Q を用いると，親から子へのある特徴の変化の確率は

$$P(x = j|x' = i, t) = \exp(tQ)_{i,j}$$

と求まる．図 3 に連続時間マルコフ連鎖の例を示す．遷移率行列 $Q = \left(\begin{smallmatrix} -0.5 & 0.5 \\ 0.25 & -0.25 \end{smallmatrix}\right)$，$i = 0$ とし，$j = 0$ の確率を実線，$j = 1$ の確率を破線で示している．$t = 0$ では初期値のままの $j = 0$ である確率が 1 だが，時間とともに $j = 1$ となる確率が上昇し，定常分布に収束している．遷移率行列の要素の大きさは，定常分布や収束の速さを制御する．

なお，内部ノードの状態を陽に持つのではなく，周辺化によりすべての状態の組み合わせを一度に考慮することも広く行われている．この周辺化は動的計画法により効率的に解けることが知られている(Felsenstein, 1981)．

置換モデルとしては，歴史的には，生物学で遺伝子(ACGT)の置換に対応する 4×4 の遷移率行列が 1960 年代末から研究されてきた．なお，連続時間マルコフ連鎖は $1 \to 0 \to 1$ のように，死亡した特徴が復活する場合も考慮するが，語彙の場合，復帰突然変異は起きないと仮定するのが自然である．この問題に対応するために，復活のないモデルとして，確率的 Dollo モデル

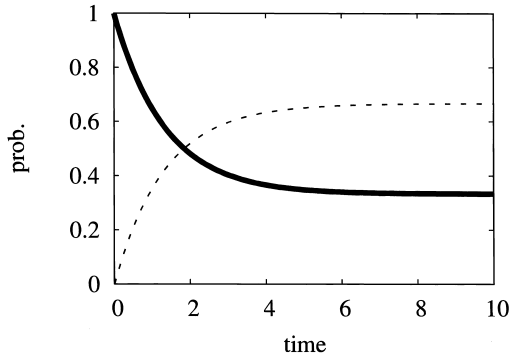


図 3. 連続時間マルコフ連鎖の例.

が提案されている(Nicholls and Gray, 2008).

これに対し, 遷移率を拡張するのが時計モデルである. 従来からの厳密時計モデルでは, 系統樹全体で同じ遷移率が用いられる. しかし, 変化の速度は一定とは限らないという問題意識が生物学においてもあり, 1990 年代以降, 変化率に柔軟性を持たせる緩和時計が研究されてきた. 緩和時計モデルの多くは, $P(x = j|x' = i, t, k) = \exp(r_k tQ)_{i,j}$ のように, 枝 k ごとに異なる係数 r_k を遷移率行列にかけることで実現される. r_k 自体も確率分布から生成されるが, 具体的なモデルは乱立しており, 評価が定まっていない.

4.4 推論

推論時には, 与えられたモデルのもとで, 観測データを入力し, 高いスコアを返すような系統樹とパラメータの組を探す. ただし, モデルの構成要素には複雑な依存関係があり, 最適解は解析的には求められない. そこで, 乱択により近似的に解を探索する. その具体的な手法として, Markov chain Monte Carlo(MCMC)法によるサンプリングが広く用いられている.

系統樹とパラメータの組のうち, 一部は観測されており, 残りは潜在変数である. 潜在変数は連続値を含むことから非加算無限個の候補があり, 離散値に限っても組合せ爆発を起こすため候補の列挙は事実上不可能である. しかし, ともかく一通り値を割り当てればモデルからスコアが得られる. また, 系統樹やパラメータを局所的に変更すると, 新たなスコアが得られる(差分に着目すれば効率的に計算できる). そこで, ひとまず潜在変数を一通り割り当て, 局所的な変更を繰り返すことで, 良い系統樹を近似的に探索するという方針をとる.

MCMC 法の一つである Metropolis-Hastings アルゴリズムでは, スコアが上がった場合はその変更を採用する. スコアが下がった場合は, 多くの場合は変更を破棄して元の割り当てを採用するが, ある確率で変更を採用する. これにより局所解からの脱出が可能となる. 適当な条件が満たされるとき, この操作を無限に繰り返すことで, 最適解にたどり着くことが保証される.

この手続きがサンプリングとよばれるのは, 確率分布からサンプルを得ているとみなせるからである. 簡単な例では, サイコロを振って $3, 4, 5, 3, \ldots$ とサイコロの目を得たとき, それらは多項分布(いわゆる categorical distribution で, ベルヌーイ分布の多項版)からのサンプルである. 連続値の場合も同様に, 正規分布から $1.78\ldots, -0.32\ldots, 0.27\ldots, \ldots$ といったサンプルが得られる. 系統モデルはこれらにくらべてはるかに複雑で, 連続値と離散値の組み合わせからなる確率分布だが, MCMC 法により得られる系統樹とパラメータの組は, この確率分布からのサンプルとみなせる.

こうして系統樹とパラメータの組の複数のサンプルが得られるが，人間が容易に解釈可能な形で要約する必要がある．多くの場合，興味の中心は共通祖語の年代であるため，サンプルから年代を集めてきてヒストグラムを作ればよい．複数の系統樹の要約方法には様々な変種があるが，なかでも最大系統群信頼度木(Heled and Bouckaert, 2013)とよばれる手法がよく用いられる．

4.5 現状と今後の展望

分子生物学由来の統計モデルの適用に対する言語学者の反応は，大半が懐疑的であるという印象を筆者は抱いている．特に一部の言語学者からは激烈な反応が発せられている(Pereltsvaig and Lewis, 2015)．こうした反応のなかでは，これまでの研究経緯を無視したような論文が，言語学者の査読を経ていない(ようにみえる)まま高名な論文誌に掲載されることへのいらだちや，人文系の予算が急速に削減されるなか，計算機を使った研究が科学メディア等でもてはやされていることへの感情的な反発が渾然一体となっている．しかし，統計的手法の研究者の目からは，データベース整備や定性的な議論の面で言語学者の協力が不可欠であることは充分すぎるほど明らかである．したがって，言語学者が説得すべき相手は，統計的手法の研究者よりも，むしろ科学メディアや研究資金提供機関であろう．また，Bouckaert et al. (2012)の著者の一人である Drummond と Pereltsvaig and Lewis (2015)の著者らのブログ上でのかみ合わない議論(http://www.geocurrents.info/cultural-geography/linguistic-geography/mismodeling-indo-european-origin-and-expansion-bouckaert-atkinson-wade-and-the-assault-on-historical-linguistics#disqus_thread)は，伝統的な歴史言語学の教育を受けた者が，統計モデルを用いて問題を解くという発想を受け入れることの難しさを痛感させるものである．これは高等教育の制度設計にかかわる問題であり，一朝一夕には解決できない．

一連の統計的研究の火付け役となった Gray and Atkinson (2003)をはじめとする多くの研究は印欧祖語を対象としている．印欧祖語の故地と年代に関してこれまでに多くの説が提案されてきたが，特に有力なのはクルガン仮説とアナトリア仮説である．クルガン仮説は，考古学的証拠をもとに，5,000–6,000 年前，黒海沿岸のステップ地帯(クルガンはこの地帯にみられる墳墓のこと)に印欧祖語の話者がおり，遊牧民の軍事的征服により各地に広がったとする．一方のアナトリア仮説は，同じく考古学的証拠をもとに，8,000–9,500 年前のアナトリア(現在のトルコのアジア側)を起源とし，農耕とともに拡散したとする．ベイズ系統モデルを用いる Gray らは，一貫してアナトリア仮説を支持している(Gray and Atkinson, 2003; Bouckaert et al., 2012)．しかし，アナトリア仮説は言語学者の間では評判が悪く，これがベイズ系統モデルへの不信感にもつながっているのではないかと筆者は推測している．

もしクルガン仮説が正しいとすると，これまでのモデルのどこに問題があったのだろうか．この問題について，Chang et al. (2015)は意味変化による成因的相同に着目している．伝統的な比較法では，意味変化にかかわらず同源語を追跡するが，言語年代学以降の統計手法では，最初に意味ごとに区切って語を収集する．しかし，ある種の意味変化は通言語的によく起こり，結果として成因的相同を生み出している．例えば，ADULT MALE を意味する現代アイルランド語の *duine*，フランス語の *homme*，ゴート語の *guma* は同源語だが，PERSON から ADULT MALE への意味変化が，古アイルランド語から現代アイルランド語，およびラテン語からフランス語にいたる過程で独立に起きた結果誕生したものである(ゴート語も同様と思われる)．しかし，系統推定を行うと，それらの言語の共通祖先にまでこの語をさかのぼらせるのがより自然な解釈となる．つまり，変化を実際よりも古い段階に持って行き，かつ，時間あたりの変化数を実際よりも低く推定してしまう．

この問題に対処するために，Chang et al. (2015)は，古代語を現代語の過去の状態として使う

というモデル変更を提案している．例えば，現代アイルランド語の ADULT MALE の意味での *duine* の値は 1 だが，古アイルランド語の状態を経るため，時間をさかのぼって他の言語と合流する前に値が 0 となる．これに対し，従来手法では古代語も葉ノード扱いしていたため，内部ノードとして現代語・古代語共通祖語が推定され，その値は 1 になる可能性が高かった．この変更の結果，祖語の年代は約 6,500 年前と推定され，Bouckaert et al. (2012)とくらべて大幅にクルガン仮説に近づいている．

Chang et al. (2015)の第一著者は，計算機科学から言語学に転じたという特異な経歴を持っている．言語現象を丁寧に分析し，それに対応する統計モデルを提案するというこの研究の取り組みは，今後の研究のあり方を示す模範例だと筆者は考えている．

Chang et al. (2015)は従来のモデルの仮定が成り立っていないことを指摘したものだが，そもそも系統モデルは多数の部分モデルの組み合わせであり，それぞれの部分モデルの背後にはデータに関する仮定がある．派手な研究成果が喧伝される一方で，数多くの仮定の検証はまだ充分になされていない．例えば，上述のアイスランド語のように極端に変化が遅い言語を含むデータに対しても緩和時計モデルであれば適合することが期待されるが，これまでの報告を読む限りでは，実際にうまくいっているか判然としない．地道な検証が必要であり，そのための道標として，歴史言語学でこれまでになされてきた議論に有用なものが少なくないと考えている．

仮定のなかで疑うべき最大のものは，系統樹そのものかもしれない．系統樹は理想化にすぎず，系統モデルでは説明できない接触に基づく言語現象が存在することは，歴史言語学においては系統樹の発明当初から認識されてきた(Schmidt, 1872)．これに応じて，系統モデルを推進する Gray らも，接触が系統推定に与える影響について多くの議論を費やしている(Greenhill et al., 2009; Gray et al., 2010)．

特に方言(非常に近い言語)群の関係を考える場合，接触の影響が無視できる範囲を超えていると考えられる．実際，伝統的な方言区画論は現代の特徴に基づく階層的クラスタリングにすぎず，それが時間変化を表す系統樹に対応するという観念は希薄である(東条, 1927)．方言群に強引に系統モデルを適用した報告(Lee and Hasegawa, 2011)もあるが，その結果得られた不可解な系統樹は，少なくとも部分的には接触の影響によって説明できると見られる(Murawaki, 2015b)．方言の語彙的特徴については，むしろ拡散を扱う非統計的モデルが提案されてきた(柳田, 1930; Trudgill, 1974)．統計モデルについては，シミュレーションモデル(Lizana et al., 2011; Murawaki, 2015b)は提案されているものの，実データを扱えるモデルと推論手続きの開発が課題として残っている．接触を考慮すると，モデルの自由度が単なる系統樹にくらべて大幅に上がるからである．

5. 言語類型論に基づく系統推定に向けて

5.1 日本語の起源と言語類型論

現在のところ，語彙的特徴を用いる統計モデルは，日本語と他の言語との系統関係の解明には利用できない．日本語系統論の研究には百年以上の蓄積があるにもかかわらず，日本語と他の言語との間で信頼できる同源語が充分に特定できていないからである(Vovin, 2010)．このことから，逆説的に，仮に同系言語が現存していたとしても，祖語の年代は相当さかのぼるのではないかと推測される(服部, 1999)．

筆者は語彙に代わる手がかりとして言語類型論に注目している(Murawaki, 2015a)．言語類型論とは，世界の諸言語を類型によって分類する分野である．類型の例には，基本語順(SVO, SOV 等)，助数詞の有無，声調の有無がある．これらの特徴を用いると，言語の状態は多値特徴の列で表現できる．

類型論の利点として，語彙とは異なり，日本語を含む任意の言語対が比較できることが挙げられる．実際，日本語と同系の言語の有力候補として，朝鮮語，さらにはツングース，モンゴル，テュルクを核とするいわゆるアルタイ諸語が挙げられてきたが，その根拠となったのは，「語頭に r 音が立たない」，「*have* 型の所有動詞を持たない」といった類型論上の類似である．また，類型論のいくつかの特徴は，語彙とくらべて歴史的に安定的だと推測される（Nichols, 1992; 松本, 2007）．

しかし，歴史言語学では，系統関係は語彙的特徴によって確立されるもので，類型論上の類似は決定的な証拠にならないという見方が支配的である．アルタイ諸語の場合も，従来指摘された特徴は実は世界的にありふれており，該当言語群に固有のものではないことが指摘されている（松本, 2007）．

進化という観点から類型論的特徴の性質を見直すと，欠点としてまず挙げられるのは，成因的相同や復帰突然変異が広範囲に起こりえることである．例えば，ある言語対の基本語順の値がいずれも SVO だとしても，その特徴を共通祖先から引き継いだとは限らない．SVO 語順は歴史上無関係に複数回誕生したと考えられるし，一度失われたとしても復活しうるからである．つまり，語彙と違って類型論的特徴は手がかりとして不確実性が高く，人手による論証になじまない．見方を変えれば，計算機を用いた統計的手法が活躍できそうな未開拓地が広がっている．まずは，データに成り立つ統計的性質を明らかにするところからはじめる必要があるだろう．

これまでにも類型論の系統推定への利用を試みる報告もいくつかあるが（Tsunoda et al., 1995; Dunn et al., 2005; Longobardi and Guardiano, 2009），語彙に基づく手法とくらべて少ない．系統推定の手がかりとしての類型論的特徴の有効性についても，肯定的な報告（Dunn et al., 2005; Longobardi and Guardiano, 2009）とやや否定的な報告（Greenhill et al., 2010; Dunn et al., 2011）が混在し，結論が出ていない．

従来，類型論に対する統計的取り組みの障害となっていたのは，データ整備に言語学の高度な知識が必要となることである．例えば，動詞のように自明に思える概念であっても，世界中の言語を収集すると，悩ましい事例が出てくる．かつては言語学者が個別にデータを収集していた（角田, 1991）が，現在は組織的なデータ収集の成果として World Atlas of Language Structures（WALS）とよばれるデータベース（Haspelmath et al., 2005）が公開されており，望むなら統計モデルの設計に専念することも可能な環境が整いつつある．

5.2 類型論的特徴の変化の経路

類型論的特徴はどのように変化するのだろうか．そもそも，語彙の交代とくらべると，例えば語順の変化がどのように起きるかは直感的に想像しづらい．基本語順が SOV から SVO に変化するのは一大変化であり，言語システム全体に複雑な影響があると思われる．

これまでの系統モデルの研究では，語彙的特徴の場合と同様に，特徴ごとに独立な置換モデルを仮定することが多い（Teh et al., 2008; Daumé III, 2009; Maurits and Griffiths, 2014）．しかし，類型論的特徴の場合，特徴間の依存関係は無視できる範囲を超えているのではないかと筆者は推測している．

実際，言語類型論の従来研究では，類型論の特徴間の依存関係が大きな関心事であった．例えば，数詞（Q）と名詞（N）の語順と形容詞（A）と名詞（N）の語順を考えると，QN，NQ と AN，NA の組み合わせにより，図 4 のように 4 通りの状態を取りえる．しかし，世界の言語を見ると，（QN, AN）型および（NQ, NA）型という一貫した語順を持つ言語が多く，（QN, NA）型の言語も存在するが，（NQ, AN）型の言語は非常に稀である．つまり，「NQ ならば NA」という含意関係がよく成り立つ．さらに歴史的変化を考え，もし（QN, AN）型から（NQ, NA）型への変化があったとすると，途中で（NQ, AN）型ではなく，（QN, NA）型を経由したと解釈するのが自然である

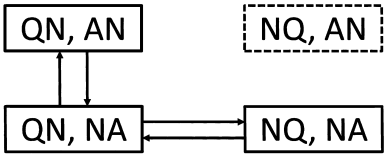


図 4. 2 つの特徴からなる状態とその遷移.

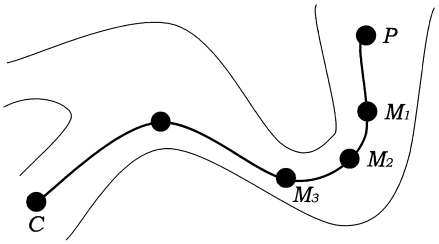


図 5. 祖語 P から言語 C への多様体(帯)上の遷移(太線).

(Greenberg, 1978). ここで, 特徴ごとに独立な変化を仮定すると, (NQ, AN)型の祖語を推定するおそれがある.

依存関係に関する統計的研究としては, 特徴対の含意関係をデータから自動発見するモデル(Daumé III and Campbell, 2007)が提案されているが, 統計が真価を発揮するのは, むしろ 3 つ以上の特徴間の関係のモデル化だと筆者は考えている. 類型論データベースに記述された特徴の数は百のオーダであり, このような組み合わせは人間の手に余るが, 計算機であれば扱える.

特徴間の依存関係を一般化して考えると, 特徴列が作る高次元空間のなかで, 言語が占めうる部分空間はごく一部だと推測される. また, 言語がある程度漸進的に変化してきたことを考慮すると, ある祖語 P から言語 C への経路について, P, C はもちろん, 途中状態 $M_1, M_2, \ldots$ も, 実際に人間が話した言語であるからには自然でなければならない. だとすると, この部分空間は相互に孤立した部分空間の集まりではなく, 図 5 のように, 局所的には近傍と繋がりあった多様体のような構造を持っていると推測できる. この多様体上であれば言語として自然, そこから外れていれば不自然とみなせる.

多値特徴列からなる状態 x の自然さを判定するモデルは, エネルギーに基づく学習(LeCun et al., 2006)と同様の枠組みで作ることができる(Murawaki, 2015a). つまり, モデルのパラメータを θ とすると, 関数 $f(x;\theta) = d \in \mathbb{R}$ が, 自然な特徴の組み合わせに高い確率を, そうでない組み合わせに低い確率を返すようにするのが目標となる. そのために適切なモデルを設計したうえで, パラメータ θ を類型論データベースに登録された実在の諸言語から学習する問題として定式化できる.

関数 f は言語学で言うところの共時的な様態を学習していることになるが, 上述の議論の通り, 歴史的変化にも応用できる. すなわち, 言語 C およびその祖語 P はもちろん, 途中状態 $M_1, M_2, \ldots$ に対しても, f が高い確率を返すはずである. また, 進化はある程度漸進的であり, M_t は M_{t-1} の近傍に位置するはずである. この 2 つの条件を考慮することで, 言語変化の経路がかなりの程度絞り込めるのではないかと見込んでいる.

5.3 言語接触の影響

類型論的特徴の場合，語彙的特徴以上に系統樹の仮定が疑われる．実際，言語類型論では，地域的特徴とよばれる接触の影響に多くの議論が費やされてきた．特徴の変化のモデル化と並行して，系統樹の妥当性も検証していく必要がある．

系統関係の親疎にかかわらず，多くの地域的特徴を共有する言語群は言語連合とよばれる．特に有名なのがバルカン言語連合であり，ギリシア語，アルバニア語，東南スラヴ諸語など，印欧語族のなかでも比較的系統的に遠い言語が含まれている．統計モデルとしては，系統推定に言語連合を組み込んだベイズモデルが提案されている(Daumé III, 2009)．このモデルでは，個々の言語は(1)系統樹に沿った進化の結果と(2)言語連合からの生成の確率的混合としてモデル化されている．ただし，モデルの自由度を抑えて推論可能とするために，言語連合は時間不変なクラスタとしてモデル化されている点が不自然である．

接触に関わる現象で，より扱いやすいものとして，筆者はクレオール形成に着目している(Murawaki, 2016)．クレオールは複数の言語の影響下で成立したとみられる一群の言語で，その多くがヨーロッパによる植民地化の影響を受けた大西洋・インド洋沿岸に分布している．クレオール形成過程は論争の絶えない課題だが，有力な仮説によると，文法が極端に単純化したピジンがまず生まれ，その後子供がピジンを母語として獲得し，その過程で複雑な意思疎通が可能なほど文法が発達することによって成立するという．しかし，この際に起きる言語普遍的な構造再編がクレオールを特徴づけているという説と，語彙提供言語や基層言語などとよばれる言語の影響が強いという説が入り乱れている．

これらの説を踏まえると，クレオールは，(1)語彙提供言語 L，(2)基層言語 S，(3)構造再編 R という 3 種類の確率的混合としてモデル化できる．具体的には，各クレオールについて混合比 $\theta = (\theta_{\mathrm{L}}, \theta_{\mathrm{S}}, \theta_{\mathrm{R}})$ を導入する．この混合比にしたがって 3 種類のいずれかの特徴を確率的に選んだ結果，各クレオールが形成されたと仮定する．この混合比をデータから推定することで，上記の仮説を検証する．つまり，集団遺伝学における Admixture 解析(Pritchard et al., 2000)や，言語処理におけるトピックモデル LDA(Latent Dirichlet Allocation)(Blei et al., 2003)に似た混合モデルによってクレオール形成が分析できる．

このモデルをクレオールの類型論データベース(Michaelis et al., 2013)に適用したところ，クレオール形成において構造再編が無視できない影響を持っているという結果を得た．また，クレオールの持つ諸特徴から語彙提供言語や基層言語の影響を差し引いたものと日本語を比較したところ，日本語はクレオール的ではないことが確認できた．日本語系統論のなかには，日本語混成言語説との関係においてクレオールに注目する議論があるが，日本語の場合，近い過去にクレオール形成はなかったと推測できる．ただし，クレオール形成よりもより穏健な言語接触が日本語の形成に影響を及ぼした可能性はあり，その解明は今後の課題として残っている．

6. おわりに

本稿では，歴史言語学的課題に対する近年の統計的研究について，これまでの研究経緯を含めて簡単に紹介するとともに，今後の方向性を議論した．歴史言語学的課題への取り組みは，長年人手による手法が主流であった．人間は記号の離散的な操作や，一歩一歩積み上げるような論証は得意だが，一方で，年代のような連続値を含む問題や，不確実性が解候補の組合せ爆発を生む問題は苦手である．こうした問題については計算機を用いる統計的手法が適しており，分子生物学由来の統計的手法がこれまでにない成果を生んできた．

しかし，主に生物データ向けのソフトウェアパッケージを転用することで研究が進んできたため，生物に対応物が存在しない言語現象がモデル化から取り残される傾向にあるように筆者

は感じている．その一方で言語資源の組織的な整備が進んでおり，統計モデルによる仮説の検証が容易に行える環境が整いつつある．したがって，言語現象を丁寧に分析し，それに対応する統計モデルを提案していけば，大きな成果が期待できる．本稿をきっかけに，こうした課題に取り組む研究者が一人でも増えれば幸いである．

参 考 文 献

Bergsland, K. and Vogt, H. (1962). On the validity of glottochronology, *Current Anthropology*, **3**(2), 115–153.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent Dirichlet allocation, *Journal of Machine Learning Research*, **3**, 993–1022.

Bouchard-Côté, A., Hall, D., Griffiths, T. L. and Klein, D. (2013). Automated reconstruction of ancient languages using probabilistic models of sound change, *Proceedings of the National Academy of Sciences*, **110**(11), 4224–4229.

Bouckaert, R., Lemey, P., Dunn, M., Greenhill, S. J., Alekseyenko, A. V., Drummond, A. J., Gray, R. D., Suchard, M. A. and Atkinson, Q. D. (2012). Mapping the origins and expansion of the Indo-European language family, *Science*, **337**(6097), 957–960.

Chang, W., Cathcart, C., Hall, D. and Garrett, A. (2015). Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis, *Language*, **91**(1), 194–244.

Darwin, C. (1859). *The Origin of Species by Means of Natural Selection or, the Preservation of Favored Races in the Struggle for Life*, John Murray, London.

Daumé III, H. (2009). Non-parametric Bayesian areal linguistics, *Proceedings of Human Language Technologies*: *The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 593–601.

Daumé III, H. and Campbell, L. (2007). A Bayesian model for discovering typological implications, *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 65–72.

Drummond, A. J. and Bouckaert, R. R. (2015). *Bayesian Evolutionary Analysis with BEAST*, Cambridge University Press, Cambridge.

Dunn, M., Terrill, A., Reesink, G., Foley, R. A. and Levinson, S. C. (2005). Structural phylogenetics and the reconstruction of ancient language history, *Science*, **309**(5743), 2072–2075.

Dunn, M., Greenhill, S. J., Levinson, S. C. and Gray, R. D. (2011). Evolved structure of language shows lineage-specific trends in word-order universals, *Nature*, **473**(7345), 79–82.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach, *Journal of Molecular Evolution*, **17**(6), 368–376.

Forster, P. and Renfrew, C. (eds.) (2006). *Phylogenetic Methods and the Prehistory of Languages*, McDonald Institute for Archaeological Research, Cambridge.

Gray, R. D. and Atkinson, Q. D. (2003). Language-tree divergence times support the Anatolian theory of Indo-European origin, *Nature*, **426**(6965), 435–439.

Gray, R. D. and Jordan, F. M. (2000). Language trees support the express-train sequence of Austronesian expansion, *Nature*, **405**(6790), 1052–1055.

Gray, R. D., Bryant, D. and Greenhill, S. J. (2010). On the shape and fabric of human history, *Philosophical Transactions of the Royal Society B*: *Biological Sciences*, **365**(1559), 3923–3933.

Greenberg, J. H. (1978). Diachrony, synchrony and language universals, *Universals of Human Language* (ed. Joseph H. Greenberg), Volume 1, Stanford University Press, Stanford, California.

Greenhill, S. J., Currie, T. E. and Gray, R. D. (2009). Does horizontal transmission invalidate cultural

phylogenies?, *Proceedings of the Royal Society B*: *Biological Sciences*, **276**(1665), 2299–2306.

Greenhill, S. J., Atkinson, Q. D., Meade, A. and Gray, R. D. (2010). The shape and tempo of language evolution, *Proceedings of the Royal Society B*: *Biological Sciences*, **277**(1693), 2443–2450.

Hammarström, H., Forkel, R., Haspelmath, M. and Bank, S. (eds.) (2016), *Glottolog 2.7*, Max Planck Institute for the Science of Human History, Jena.

Haspelmath, M., Dryer, M., Gil, D. and Comrie, B. (eds.) (2005). *The World Atlas of Language Structures*, Oxford University Press, Oxford.

服部四郎 (1999).『日本語の系統』, 岩波書店, 東京.

Heled, J. and Bouckaert, R. R. (2013). Looking for trees in the forest: Summary tree from posterior samples, *BMC Evolutionary Biology*, **13**(1), 1–11.

Huelsenbeck, J. P. and Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees, *Bioinformatics*, **17**(8), 754–755.

International HapMap Consortium (2005). A haplotype map of the human genome, *Nature*, **437**(7063), 1299–1320.

Kingman, J. F. C. (1993). *Poisson Processes*, Oxford University Press, Oxford.

LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M. and Huang, F. J. (2006). A tutorial on energy-based learning, *Predicting Structured Data* (eds. Gökhan Bakır, et al.), MIT Press, Cambridge, Massachusetts.

Lee, S. and Hasegawa, T. (2011). Bayesian phylogenetic analysis supports an agricultural origin of Japonic languages, *Proceedings of the Royal Society B*: *Biological Sciences*, **278**(1725), 3662–3669.

Lizana, L., Mitarai, N., Sneppen, K. and Nakanishi, H. (2011). Modeling the spatial dynamics of culture spreading in the presence of cultural strongholds, *Physical Review E*, **83**(6), 066116.

Longobardi, G. and Guardiano, C. (2009). Evidence for syntax as a signal of historical relatedness, *Lingua*, **119**(11), 1679–1706.

松本克己 (2007).『世界言語のなかの日本語: 日本語系統論の新たな地平』, 三省堂, 東京.

Maurits, L. and Griffiths, T. L. (2014). Tracing the roots of syntax with Bayesian phylogenetics, *Proceedings of the National Academy of Sciences*, **111**(37), 13576–13581.

Michaelis, S. M., Maurer, P., Haspelmath, M. and Huber, M. (eds.) (2013). *APiCS Online*, Max Planck Institute for Evolutionary Anthropology, Leipzig.

Murawaki, Y. (2015a). Continuous space representations of linguistic typology and their application to phylogenetic inference, *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, 324–334.

Murawaki, Y. (2015b). Spatial structure of evolutionary models of dialects in contact, *PLoS ONE*, **10**(7), 1–15.

Murawaki, Y. (2016). Statistical modeling of creole genesis, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*.

Nicholls, G. K. and Gray, R. D. (2008). Dated ancestral trees from binary trait data and their application to the diversification of languages, *Journal of the Royal Statistical Society*: *Series B* (*Statistical Methodology*), **70**(3), 545–566.

Nichols, J. (1992). *Linguistic Diversity in Space and Time*, University of Chicago Press, Chicago and London.

Nichols, J. and Warnow, T. (2008). Tutorial on computational linguistic phylogeny, *Language and Linguistics Compass*, **2**(5), 760–820.

Pellard, T. (2009). Ōgami: Éléments de description d'un parler du sud des Ryūkyū, Ph.D. Thesis, Ecole des Hautes Etudes en Sciences Sociales (EHESS) (in French).

Pereltsvaig, A. and Lewis, M. W. (2015). *The Indo-European Controversy*: *Facts and Fallacies in Historical Linguistics*, Cambridge University Press, Cambridge.

Pritchard, J. K., Stephens, M. and Donnelly, P. (2000). Inference of population structure using multilocus genotype data, *Genetics*, **155**(2), 945–959.

Roberts, S. and Winters, J. (2013). Linguistic diversity and traffic accidents: Lessons from statistical studies of cultural traits, *PLoS ONE*, **8**(8), 1–13.

Schleicher, A. (1853). Die ersten Spaltungen des indogermanischen Urvolkes, *Allgemeine Monatsschrift für Wissenschaft und Literatur*, **3**, 786–787 (in German).

Schmidt, J. (1872). *Die Verwandtschaftsverhältnisse der indogermanischen Sprachen*, Hermann Böhlau, Weimar (in German).

Swadesh, M. (1952). Lexicostatistic dating of prehistoric ethnic contacts, *Proceedings of American Philosophical Society*, **96**, 452–463.

Teh, Y. W., Daumé III, H. and Roy, D. (2008). Bayesian agglomerative clustering with coalescents, *Advances in Neural Information Processing Systems 20*, 1473–1480.

東条 操 (1927).『国語の方言区画』, 東京堂出版, 東京.

Trudgill, P. (1974). Linguistic change and diffusion: Description and explanation in sociolinguistic dialect geography, *Language in Society*, **3**, 215–246.

角田太作 (1991).『世界の言語と日本語』, くろしお出版, 東京.

Tsunoda, T., Ueda, S. and Itoh, Y. (1995). Adpositions in word-order typology, *Linguistics*, **33**(4), 741–762.

Vovin, A. (2010). *Koreo-Japonica*, University of Hawai'i Press, Honolulu.

柳田國男 (1930).『蝸牛考』, 刀江書院, 東京.

Zuckerkandl, E. and Pauling, L. (1965). Evolutionary divergence and convergence in proteins, *Evolving Genes and Proteins*, **97**, 97–166.

Statistical Approaches to Language Change and Linguistic Phylogenies

Yugo Murawaki

Graduate School of Informatics, Kyoto University

Since around the turn of the twenty-first century, there has been a growing trend to employ computer-intensive statistical methods to answer historical linguistic questions, such as language change and phylogenies of extant and documented languages. Although these questions have traditionally been addressed manually by linguists, manual analysis has limitations. Because human inference is based on logic, humans are unable to estimate continuous values (e.g., dating the common ancestor of extant languages). They are also bad at inherent uncertainty because it leads to a combinatorial explosion. Computational statistics provides powerful ways to solve these problems.

The current trend can be characterized by the fact that key results have been achieved with statistical methods originally developed in the field of molecular biology. Although historical linguistics itself has a record of adopting statistical models, the new statistical techniques have been developed largely independently of historical linguistics. Therefore their scientific foundations have yet to be fully understood by linguistic communities. We also observe that since most recent statistical studies on linguistic questions depend on ready-to-use software packages that are designed to address biological questions, linguistic phenomena that lack exact counterparts in biology tend to be left untouched.

In light of this, we first overview the new statistical models while relating them to the research history of historical linguistics. After reviewing the concept of evolution, the comparative method that exploits regular sound changes, and ill-fated glottochronology, we explain the essence of recently developed Bayesian phylogenetic models.

Since most phylogenetic models use lexical traits, they can be applied only if the group of languages in question has a sufficient number of shared lexical traits. Unfortunately, this is not the case in Japanese, and we have no choice but to seek for different kinds of traits. Later in this paper, we describe novel approaches based on typological traits, which we believe have the potential to trace the origin of the Japanese language.

Key words: Linguistic phylogeny, historical linguistics, linguistic typology, Bayesian statistics.

統計数理(2016)
第 64 巻 第 2 号 179–200

特集「統計的言語研究の現在」
[原著論文]

条件付き確率場の理論と実践

岡崎 直観†

(受付 2016 年 6 月 17 日；改訂 8 月 25 日；採択 9 月 14 日)

要　　旨

自然言語処理のタスクの多くは，入力から出力のラベルを予測する問題として定式化できる．言語は構造を持つと考えられるので，入力や出力に単語列や木などの構造を持たせることで，さらに多くのタスクが予測問題として定式化できる．本稿では，系列ラベリング問題，すなわち入力と出力が系列データの場合の条件付き確率場を解説する．条件付き確率場は，多クラスロジスティック回帰を系列データに適用するため，ラベル列のマルコフ性を仮定した素性関数を導入し，動的計画法でラベル列の予測とパラメータの学習を効率化している．そこで，ロジスティック回帰の素性関数，確率的勾配降下法による学習，正則化などの基礎理論を復習し，条件付き確率場の理論全体を説明する．また，能動学習，部分的に正解が付与された訓練データからの学習，深層ニューラルネットワークの適用など，条件付き確率場の最近の研究動向や実践について概観する．

キーワード：条件付き確率場，ロジスティック回帰，確率的勾配降下法．

1. はじめに

自然言語処理は，言葉を操る賢いコンピュータを実現することを究極の目標としているが，その目標到達への道程は長い．一方で，自然言語処理は情報検索，機械翻訳，質問応答，自動要約，対話生成，評判分析，ソーシャルネットワーク分析など，幅広い応用を生み出している．これらを支えているのは，品詞タグ付け(形態素解析)，チャンキング，固有表現抽出，構文解析，共参照解析，意味役割付与などの基盤解析技術である．

このように，自然言語処理は多種多様なタスクに取り組んでいるが，その多くのタスクは入力 x から出力 $\hat{y}$ を予測する問題として定式化できる．入力 x に対する出力 y の「良さ」をスコア付けする関数を $s(x, y)$ と書くと，与えられた入力に対して，可能な出力の集合 $\mathcal{Y}$ から最適な出力 $\hat{y}$ を選ぶ問題は，次式で与えられる．

$$\hat{y} = \operatorname*{argmax}_{y \in \mathcal{Y}} s(x, y) \tag{1.1}$$

例えば，入力文を単語列 $\boldsymbol{x}$ で表し，その文がポジティブ(+1)な内容であるか，ネガティブ(−1)な内容であるか，どちらでもないか(0)，を判定する評判分析タスクは，次式で定式化される．

$$\hat{y} = \operatorname*{argmax}_{y \in \{+1,0,-1\}} s(\boldsymbol{x}, y) \tag{1.2}$$

関数 $s(x, y)$ はナイーブベイズ，パーセプトロン，ロジスティック回帰，サポートベクトルマシ

†東北大学大学院 情報科学研究科：〒 980–8579 宮城県仙台市青葉区荒巻字青葉 6–6–05

ン，ニューラルネットワークなどの手法を用い，訓練データから自動的に導出することが多い．

言語は構造を持つので，出力にも単語列や木などの構造を持たせることにすると，さらに多くのタスクが予測問題として定式化される．例えば，M 個の単語の列 $\boldsymbol{x} = (x_1, x_2, \ldots, x_M)$ から M 個の品詞ラベルの列 $\boldsymbol{y} = (y_1, y_2, \ldots, y_M)$ を予測すると，品詞タグ付け(part-of-speech tagging)が実現する．予測するラベルの種類を B-NP, I-NP, B-VP などのチャンクタグに変更すると，浅い句構造解析(shallow parsing)をモデル化できる．このように，系列 $\boldsymbol{x}$ を入力として系列 $\hat{\boldsymbol{y}}$ を出力する問題は，系列ラベリング(sequential labeling)と呼ばれる．

$$\hat{\boldsymbol{y}} = \operatorname*{argmax}_{\boldsymbol{y} \in \mathcal{Y}^M} s(\boldsymbol{x}, \boldsymbol{y}) \tag{1.3}$$

ここで，出力されるラベルの集合を $\mathcal{Y}$ と書くと，予測されるラベル列 $\hat{\boldsymbol{y}}$ は $\mathcal{Y}^M$ の中から選ぶ．また，スコア付け関数 $s(\boldsymbol{x}, \boldsymbol{y})$ は，隠れマルコフモデル，構造化パーセプトロン，条件付き確率場などで学習する．

本稿では，系列データにおける条件付き確率場の理論と実践を解説する．条件付き確率場の本質は，多クラスロジスティック回帰を系列データに適用するため，ラベル列のマルコフ性を仮定した素性関数を導入し，動的計画法でラベル列の予測とパラメータの学習を効率化した点にある．そこで，本稿では条件付き確率場の導入として，ロジスティック回帰，および多クラスロジスティック回帰を復習する．これらの理論をベースに条件付き確率場を説明し，能動学習や部分アノテーションなどの拡張や，最近の研究動向を概観する．

2. 線形二値分類器

線形二値分類器(linear binary classifier)は，入力 x に対してスコア $s(x) \in \mathbb{R}$ を計算し，その正負によって二値のラベル $\hat{y} \in \{+1, -1\}$ を推定する．

$$s(x) = \boldsymbol{w}^{\intercal} \boldsymbol{\phi}(x) \tag{2.1}$$

$$\hat{y} = \begin{cases} +1 & (s(x) = \boldsymbol{w}^{\intercal} \boldsymbol{\phi}(x) \geq 0) \\ -1 & (\text{それ以外の場合}) \end{cases} \tag{2.2}$$

ただし，$\boldsymbol{\phi}(x) \in \mathbb{R}^d$ は x を素性ベクトル(feature vector)で表現したものである(d は素性空間の次元数)．$\boldsymbol{w} \in \mathbb{R}^d$ は重みベクトル(weight vector)と呼ばれ，訓練データに適合するように学習で求める．式(2.2)は，入力の素性ベクトルと重みベクトルの内積値を計算し，その正負でラベルを推定するという，単純明快な分類ルールである．以降では，素性関数 $\boldsymbol{\phi}(x)$ の設計と，重みベクトル $\boldsymbol{w}$ の学習について説明する．

2.1 素性空間

式(2.2)からも明らかなように，分類器は入力 x を素性ベクトル $\boldsymbol{\phi}(x)$ を通して観測し，そのラベルを予測する．したがって，高性能の分類器を開発するためには，入力の特徴をよく反映し，ラベルの予測に効きそうな素性空間(feature space)を設計する必要がある．残念ながら，よい素性を設計するための汎用的な方法はないが，素性の作り方には一定の慣習がある．

題材として，文章中の英単語 x が人名の一部であるか($y = +1$)，そうでないか($y = -1$)を推定するタスクを考える．単語の先頭が大文字から始まる場合，その単語は固有名詞である可能性が高くなる．例えば，素性ベクトルの 2 次元目として，次式で表現される素性関数 $\phi_2(x)$ を定義する．

$$\phi_2(x) = \begin{cases} 1 & (x \text{ が大文字で始まる場合}) \\ 0 & (\text{それ以外の場合}) \end{cases} \tag{2.3}$$

他にも，全ての文字が小文字か(人名の可能性は低い)，全てが大文字の単語か(人名の可能性は低い)，x が *Mar* という綴りで始まるか(*Mary*, *Maria*, *Margaret* のように女性の名前である可能性が高い)，文章中で直前の単語が *Ms* か(人名の可能性は高い)など，x の特徴を表す様々な事象を素性として定義する．このように，x の特徴を表す事象(条件)を考え，その条件の成立時のみ 1 を返すような素性関数をたくさん定義し，それぞれを特徴ベクトルの各次元に割り当てる．入力 x に対して素性関数 $\phi_k(x)$ が 0 以外の値を返すとき，入力 x に対して素性 ϕ_k が「発火」したと呼ぶ．

先ほど，「単語が *Mar* という綴りで始まるか」という素性を説明したが，「単語が *ie* という綴りで終わるか(*Stephanie*, *Marie*, *Julie*)」や「単語が *ard* という綴りで終わるか(*Richard* や *Edward*)」など，人名の推定に効きそうな特徴はたくさんある．また，人名ではない英単語を $y = -1$ と予測することも重要であるので，「人名らしくない単語の特徴」にも着目すべきである．しかしながら，これらの素性を人間が網羅的に発見し，定義することは困難である．

そこで，入力の特徴を学習データから掘り起こし，素性関数を定義するアプローチもよく採用される．例えば，学習データに含まれる単語から「長さ 3 の接頭辞を取り出す」というルールを設計し，抽出された全ての接尾辞に対して素性関数を定義することで，「単語が *Mar* という綴りで始まる」「単語が *Ale* という綴りで始まる」などの素性関数が自動的に導出される．素性関数を取り出すルールは，素性テンプレート(feature template)と呼ばれる．

単語 x の長さ $2, 3, 4$ の接頭辞や接尾辞，単語 x の前後に出現する単語など，様々な素性テンプレートを設計・適用することで，人名の予測に効きそうな素性を系統的に作成できる．一方，素性テンプレートは機械的な処理なので，人名の予測に役に立たない素性も多く作り出されてしまう．しかし，線形二値分類器の学習では，それぞれの素性関数 $\phi_i(x)$ の有用性が重み w_i で調整され，役に立たない素性の重みは 0 に近くなると期待できる．ゆえに，役に立たない素性を生成するリスクを気にせず，網羅性重視で素性テンプレートを設計することが多い．

ところで，どのような入力に対しても，常に 1 を返すような素性を導入しておくと，原点を通らない分離平面を表現することができる．例えば，入力 x によらず常に $\phi_1(x) = 1$ を返すような素性を定義すると，$y = +1$ と予測する条件は次式で表される(説明のため，素性ベクトルの 1 次元目を用いたが，次元のインデックス番号は任意である)．

$$\boldsymbol{w}^\intercal \boldsymbol{\phi}(x) = w_1 + \sum_{k=2}^{K} w_k \phi_k(x) \geq 0 \Leftrightarrow \sum_{k=2}^{K} w_k \phi_k(x) \geq -w_1 \tag{2.4}$$

したがって，1 次元目の重み w_1 は分類時の閾値を表しており，この閾値をも学習で自動的に調節することができる．w_1 はバイアス項(bias term)と呼ばれ，素性の重みベクトルと分けて明示的に記述されることもある．

式(2.3)は，ある条件を満たした場合は 1 を，それ以外は 0 を返す関数として設計されている．ところが，線形二値分類器の枠組みでは，素性関数は 0 や 1 だけでなく，実数値を返してもよい．したがって，式(2.3)の素性関数が 1 の代わりに 2 や 0.1 の値を返しても問題ない．実際には，ある素性関数の値を常に定数倍しても，対応する重みがその定数倍を打ち消すように調整されることが多いので，1 以外の素性値を用いることは稀である．

ただし，ある特徴が出現した回数や，ある特徴に関連する事象が別のコーパス中で出現する回数など，入力 x に応じて素性値を変化させたい場合は，実数値を返す素性を定義する．例えば，大規模なコーパスにおいて，単語 x の直前に *Ms* という語が出現した回数(出現頻度)の対

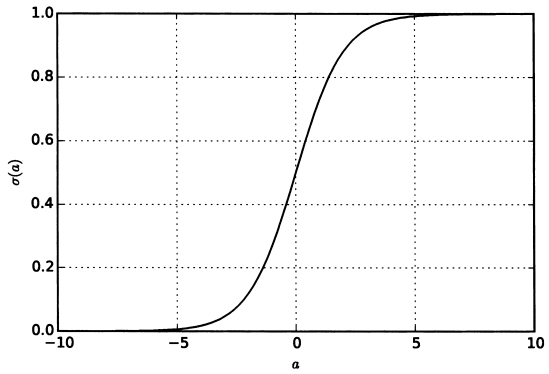


図 1. シグモイド関数 $\sigma(a)$.

数を素性値として用いることで，単語 x の名前らしさを連続値で示唆する素性を設計できる．

$$\phi(x) = \log(x \text{ がコーパスにおいて } Ms \text{ の後に出現する回数}) \tag{2.5}$$

なお，素性値は線形二値分類器のスコアに線形の影響を与える．式(2.5)では，少数の高頻度語の影響を抑えるため，出現回数の対数をとっている．このような非線形の変換は，実数値を返す素性の側に盛り込んでおく必要がある．

2.2 ロジスティック回帰

ロジスティック回帰(logistic regression)は線形二値分類器の一種で，入力 x に対するラベル $y \in \{+1, -1\}$ の条件付き確率 $p(y|x)$ を，シグモイド関数 σ(sigmoid function)でスコア $s(x)$ を確率値に変換することにより計算する．

$$p(y|x) = \sigma(ys(x)) = \frac{1}{1 + \exp\left(-y\boldsymbol{w}^\intercal \boldsymbol{\phi}(x)\right)} \tag{2.6}$$

シグモイド関数 $\sigma(a) = \frac{1}{1+\exp(-a)}$ は，$(-\infty, +\infty) \to (0, 1)$ の単調増加関数で，図 1 のような形状をしている．$p(+1|x) = \sigma(s(x))$ であるから，入力の素性ベクトルと重みベクトルの内積値 $s(x)$ を図 1 の横軸にとり，$y = +1$ と予測する確率 $p(+1|x)$ を縦軸から求めていることに相当する．図 1 からも明らかなように，$s(x)$ が大きければ $p(+1|x)$ は 1 に近づき，$s(x)$ が小さければ $p(+1|x)$ は 0 に近づく．また，

$$1 - \sigma(a) = \frac{\left\{1 + e^{-a}\right\} - 1}{1 + e^{-a}} = \frac{e^{-a}}{1 + e^{-a}} = \frac{e^a e^{-a}}{e^a(1 + e^{-a})} = \frac{1}{1 + e^a} = \sigma(-a) \tag{2.7}$$

であるから，

$$p(-1|x) = \sigma(-s(x)) = 1 - \sigma(s(x)) = 1 - p(+1|x) \tag{2.8}$$

が成り立つ．式(2.8)より，$y = +1$ と予測する確率と，$y = -1$ と予測する確率の和が 1 になることが確認できる．

なお，入力 x を $y = +1$ と予測する確率が 0.5 を上回る条件を求めると，式(2.2)と整合することが確認できる．

$$p(+1|x) \geq 0.5 \Leftrightarrow \frac{1}{1 + \exp(-\boldsymbol{w}^\intercal \boldsymbol{\phi}(x))} \geq 0.5 \Leftrightarrow \boldsymbol{w}^\intercal \boldsymbol{\phi}(x) \geq 0 \tag{2.9}$$

2.3 ロジスティック回帰モデルの学習

入力 $x^{(i)}$ に対して正解の二値ラベル $y^{(i)}$ を付与した N 件の訓練事例 $\mathcal{D} = (x^{(i)}, y^{(i)})_{i=1}^{N}$ がある．ロジスティック回帰モデルの学習とは，学習データの各訓練事例の入力 $x^{(i)}$ に対して，その正解ラベル $y^{(i)}$ を再現(正しく予測)できるように，重みベクトル $\boldsymbol{w}$ を調整することである．ある訓練事例 $(x^{(i)}, y^{(i)})$ に関して，ロジスティック回帰モデルの予測の正しさは $p(y^{(i)}|x^{(i)})$ で見積もることができる．すなわち，$p(y^{(i)}|x^{(i)})$ が 1 に近ければモデルの予測が正しく，0 に近ければモデルの予測が間違っていることになる．

そこで，全ての学習事例に対して $p(y^{(i)}|x^{(i)})$ を計算し，その確率の積を求める．

$$\prod_{i=1}^{N} p(y^{(i)}|x^{(i)}) \tag{2.10}$$

式(2.10)は，学習データ $\mathcal{D}$ におけるモデルの尤もらしさを表すので，尤度(likelihood)と呼ばれる．したがって，式(2.10)を重みベクトル $\boldsymbol{w}$ の関数とみなし，式(2.10)を最大化するような重みベクトル $\boldsymbol{w}^*$ を求めると，訓練事例を再現できる分類器を学習したことになる．このように，尤度関数を最大化することでモデルのパラメータを学習することを最尤推定(maximum likelihood estimation)と呼ぶ．ただ，式(2.10)は，小さい数(確率値)の積を計算するため，数値計算ではアンダーフローなどの問題を引き起こす．そこで，式(2.10)の対数をとった対数尤度(log likelihood)を最大化することが多い．

$$\mathcal{L}^{\mathrm{LR}} = \log \prod_{i=1}^{N} p(y^{(i)}|x^{(i)}) = \sum_{i=1}^{N} \log p(y^{(i)}|x^{(i)}) \tag{2.11}$$

まとめると，ロジスティック回帰モデルの学習は，以下の目的関数 $E^{\mathrm{LR}}(\boldsymbol{w})$ を最小化することに帰着する．

$$E^{\mathrm{LR}}(\boldsymbol{w}) = -\mathcal{L}^{\mathrm{LR}} = -\sum_{i=1}^{N} \log p(y^{(i)}|x^{(i)}) \tag{2.12}$$

幸いなことに，この目的関数は凸関数であり，大域最適解 $\boldsymbol{w}^*$ を持つ．式(2.12)を最小化する手法として，L-BFGS 法や確率的勾配降下法(stochastic gradient descent)が代表的である．ここでは，理論と実装が簡単な確率的勾配降下法を説明する．

確率的勾配降下法は勾配降下法(gradient descent)をベースにしている．勾配降下法は，次の漸化式を $t = 1, 2, \ldots, T$ に関して適用し，式(2.12)の解を求める．

$$\boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(t)} - \eta^{(t)} \frac{\partial E^{\mathrm{LR}}(\boldsymbol{w})}{\partial \boldsymbol{w}^{(t)}} = \boldsymbol{w}^{(t)} + \eta^{(t)} \frac{\partial \mathcal{L}^{\mathrm{LR}}}{\partial \boldsymbol{w}^{(t)}} \tag{2.13}$$

重みベクトルの初期値 $\boldsymbol{w}^{(1)}$ は $\mathbf{0}$ など，適当な値に設定すればよい．$\eta^{(t)}$ は学習率と呼ばれ，$1/t$, $1/\sqrt{t}$, $1/(t_0 + t)$ など，反復が進むにつれて減衰していくように設定する($t_0 > 0$ はハイパー・パラメータ)．式(2.13)は，t 回目の反復において $\boldsymbol{w}^{(t)}$ を $\mathcal{L}^{\mathrm{LR}}$ の最急方向(steepest direction)に向け，幅 $\eta^{(t)}$ だけ動かすという更新式を表している．反復を終了する回数 T は，ハイパー・パラメータとして予め設定したり，収束判定などで漸化式の適用時に自動的に決定する．例えば，反復において重みベクトルの変化量の 2-ノルムが初めて $\epsilon > 0$ を下回った時を収束と判定するには，次式を用いる．

$$\|\boldsymbol{w}^{(T+1)} - \boldsymbol{w}^{(T)}\|_2 < \epsilon \tag{2.14}$$

ところで，訓練事例ごとの対数尤度

$$l(x, y) = \log p(y|x) \tag{2.15}$$

を定義すると，勾配は

$$\frac{\partial \mathcal{L}^{\mathrm{LR}}}{\partial \boldsymbol{w}} = \sum_{i=1}^{N} \frac{\partial l(x^{(i)}, y^{(i)})}{\partial \boldsymbol{w}} \tag{2.16}$$

のように，各訓練事例の勾配の和として書ける．したがって，式(2.13)の勾配 $\frac{\partial \mathcal{L}^{\mathrm{LR}}}{\partial \boldsymbol{w}^{(t)}}$ を求めるには，学習データ $\mathcal{D}$ に含まれている全ての訓練事例 $(x^{(i)}, y^{(i)})$ に関する対数尤度 $\frac{\partial l(x^{(i)}, y^{(i)})}{\partial \boldsymbol{w}^{(t)}}$ を計算し，その和を計算することになる．しかしながら，学習データの訓練事例数が多くなると，式(2.16)の計算に時間がかかるだけでなく，重みベクトル $\boldsymbol{w}$ を更新する間隔が長くなり，収束が遅くなる．

確率的勾配降下法は，勾配が式(2.16)のように各事例の勾配の和で求められるとき，勾配 $\frac{\partial \mathcal{L}^{\mathrm{LR}}}{\partial \boldsymbol{w}^{(t)}}$ の代わりに各事例 $(x^{(i)}, y^{(i)})$ の勾配 $\frac{\partial l(x^{(i)}, y^{(i)})}{\partial \boldsymbol{w}}$ を用いる．すなわち，式(2.13)の代わりに次の漸化式を用いる．

$$\boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(t)} + \eta^{(t)} \frac{\partial l(x^{(i)}, y^{(i)})}{\partial \boldsymbol{w}^{(t)}}, \quad i = t \bmod N \tag{2.17}$$

なお，$t \bmod N$ は t を訓練事例数 N で割ったときの余りである．

式(2.17)の更新式を得るため，訓練事例 (x, y) 毎の対数尤度 $l(x, y)$ を重みベクトル $\boldsymbol{w}$ で微分する．$l(x, y)$ を合成関数とみなすと，

$$\frac{\partial l(x, y)}{\partial \boldsymbol{w}} = \frac{\partial l(x, y)}{\partial p(y|x)} \frac{\partial p(y|x)}{\partial s(x, y)} \frac{\partial s(x, y)}{\partial \boldsymbol{w}}. \tag{2.18}$$

式(2.15)より，

$$\frac{\partial l(x, y)}{\partial p(y|x)} = \frac{\partial}{\partial p(y|x)} \log p(y|x) = \frac{1}{p(y|x)}. \tag{2.19}$$

次に，$p(y|x)$ を $s(x, y)$ で微分する（簡略化のために $s = s(x, y)$ と略記する）．最終行の変形では，式(2.6)と式(2.7)を用いる．

$$\begin{aligned} \frac{\partial p(y|x)}{\partial s(x, y)} &= \frac{\partial}{\partial s} \left(\frac{1}{1 + e^{-ys}} \right) \\ &= (-1) \cdot \frac{1}{(1 + e^{-ys})^2} \cdot e^{-ys} \cdot (-y) \\ &= y \cdot \frac{1}{1 + e^{-ys}} \cdot \frac{e^{-ys}}{1 + e^{-ys}} = y \cdot p(y|x) \left\{ 1 - p(y|x) \right\} \end{aligned} \tag{2.20}$$

また，

$$\frac{\partial s(x, y)}{\partial \boldsymbol{w}} = \frac{\partial}{\partial \boldsymbol{w}} \left(\boldsymbol{w}^{\intercal} \boldsymbol{\phi}(x) \right) = \boldsymbol{\phi}(x) \tag{2.21}$$

ゆえに，

$$\frac{\partial l(x, y)}{\partial \boldsymbol{w}} = \frac{1}{p(y|x)} \cdot y \cdot p(y|x) \left\{ 1 - p(y|x) \right\} \cdot \boldsymbol{\phi}(x) = y \left\{ 1 - p(y|x) \right\} \boldsymbol{\phi}(x). \tag{2.22}$$

最終的に，式(2.17)の更新式は次式で表される．

$$\boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(t)} + \eta^{(t)} y^{(i)} \{ 1 - p(y^{(i)}|x^{(i)}) \} \boldsymbol{\phi}(x^{(i)}), \quad i = t \bmod N \tag{2.23}$$

理論的な説明が長くなったが，確率的勾配降下法によるロジスティック回帰モデルの学習は

Algorithm 1 の擬似コードで表される．この擬似コードでは，学習データに対する反復回数を T 回と定めているが，式(2.14)を使って収束を判定してもよい．また，学習率 η の計算も，$1/t$, $1/\sqrt{t}$, $1/(t_0+t)$ などに変更してもよい．

Algorithm 1: 確率的勾配降下法によるロジスティック回帰モデルの学習
入力: 学習データ $\mathcal{D} = (x^{(i)}, y^{(i)})_{i=1}^{N}$
出力: 重みベクトル $\boldsymbol{w}$

$t \leftarrow 1$;
$\boldsymbol{w} = \boldsymbol{0}$;
for epoch $\leftarrow 1$ **to** T **do**
  for $i \leftarrow 1$ **to** N **do**
    $\eta \leftarrow 1/t$(※他の計算方法でも可);
    $q \leftarrow 1 - p(y^{(i)}|x^{(i)})$(※現在の重みベクトル $\boldsymbol{w}$ を用いて);
    $\boldsymbol{w} \leftarrow \boldsymbol{w} + \eta y^{(i)}(1-q)\boldsymbol{\phi}(x^{(i)})$;
    $t \leftarrow t + 1$;
  end
end

学習率 η を 1 に固定して考えると，Algorithm 1 は各訓練事例 $(x^{(i)}, y^{(i)}) \in \mathcal{D}$ に対して，以下の処理を行うと解釈できる．

(1) 現在の重みベクトル $\boldsymbol{w}$ を使い，訓練事例の入力 $x^{(i)}$ からラベル $y^{(i)}$ を予測する確率(事例の尤度) q を計算する．
(2) 重みベクトル $\boldsymbol{w}$ を次式で更新する: $\boldsymbol{w}^{\mathrm{new}} \leftarrow \boldsymbol{w}^{\mathrm{old}} + y^{(i)}(1-q)\boldsymbol{\phi}(x^{(i)})$.

$y^{(i)} = +1$ のとき，更新後の重みベクトル $\boldsymbol{w}^{\mathrm{new}}$ は次式で与えられる．

$$\boldsymbol{w}^{\mathrm{new}} = \boldsymbol{w} + (1-q)\boldsymbol{\phi}(x^{(i)}) \tag{2.24}$$

基本的に，重みベクトル $\boldsymbol{w}$ に訓練事例の素性ベクトルを $\boldsymbol{\phi}(x^{(i)})$ を足し込むことになり，その量は $(1-q)$ で調整される．ここで，$(1-q)$ は予測の外れ度合いを表すので，予測が大幅に外れているときは重みベクトルの更新幅が大きくなり，予測がそれほど外れていないときは，重みベクトルの更新幅が小さくなる．

更新後の重みベクトル $\boldsymbol{w}^{\mathrm{new}}$ を使ってスコア $s^{\mathrm{new}}(x^{(i)})$ を再計算する．

$$\begin{aligned} s^{\mathrm{new}}(x^{(i)}) &= (\boldsymbol{w} + (1-q)\boldsymbol{\phi}(x^{(i)}))^{\intercal}\boldsymbol{\phi}(x^{(i)}) \\ &= \boldsymbol{w}^{\intercal}\boldsymbol{\phi}(x^{(i)}) + (1-q)(\boldsymbol{\phi}(x^{(i)}))^{\intercal}\boldsymbol{\phi}(x^{(i)}) \geq s(x^{(i)}) \end{aligned} \tag{2.25}$$

したがって，訓練事例 $(x^{(i)}, y^{(i)})$ に関して重みベクトルを更新することで，スコア $s(x^{(i)})$ が上昇するので，この事例のラベルを $+1$ と予測しやすくなることが分かる．

$y^{(i)} = -1$ のときも同様に，

$$\begin{aligned} s^{\mathrm{new}}(x^{(i)}) &= (\boldsymbol{w} - (1-q)\boldsymbol{\phi}(x^{(i)}))^{\intercal}\boldsymbol{\phi}(x^{(i)}) \\ &= \boldsymbol{w}^{\intercal}\boldsymbol{\phi}(x^{(i)}) - (1-q)(\boldsymbol{\phi}(x^{(i)}))^{\intercal}\boldsymbol{\phi}(x^{(i)}) \leq s(x^{(i)}) \end{aligned} \tag{2.26}$$

であるから，スコア $s(x^{(i)})$ を減少させる作用があり，この事例のラベルを -1 と予測しやすく

なる．

2.4 正則化

ここで，式(2.12)の目的関数が最小となる条件について考えてみたい．これは，式(2.10)の尤度が最大となる時であるから，全ての学習事例の尤度が 1 となる場合である．しかし，式(2.6)から明らかなように，ロジスティック回帰モデルで $p(y|x) = 1$ となるのは，$s(x) \to \pm\infty$ の時だけである．素性ベクトル $\boldsymbol{\phi}(x)$ の大きさを ∞ とすることはあり得ないので，$s(x) \to \pm\infty$ を成立させるには，重みベクトル $\boldsymbol{w}$ の大きさが ∞ になる必要がある．

一般的に，重みベクトル $\boldsymbol{w}$ が大きくなると，素性ベクトル $\boldsymbol{\phi}(x)$ の僅かな差でもスコア $s(x)$ が大きく変動するため，学習データのノイズに対する耐性が低下する．また，重みベクトルの要素の分散が大きくなるため，それぞれの訓練事例のみで発火する素性に依存してスコア付けを行うようになり，過学習を引き起こす．そこで，実際にロジスティック回帰モデルを学習するときは，目的関数に正則化項(regularization term)を追加し，重みベクトル $\boldsymbol{w}$ が大きくなり過ぎないように制御する．例えば，L2-正則化(l_2-regularization)では重みベクトル $\boldsymbol{w}$ の 2-ノルムをペナルティ項として目的関数に導入する．

$$E^{\mathrm{L2LR}}(\boldsymbol{w}) = -\mathcal{L}^{\mathrm{LR}} + \frac{1}{2}C\|\boldsymbol{w}\|_2^2 = -\sum_{i=1}^{N} \log p(y^{(i)}|x^{(i)}) + \frac{1}{2}C\|\boldsymbol{w}\|_2^2 \tag{2.27}$$

ただし，$C > 0$ は L2-正則化の強さを調整するハイパー・パラメータである．C を大きくすると，訓練事例への適合よりも $\boldsymbol{w}$ の 2-ノルムの大きさを抑える作用が強くなる．C を小さくすると，訓練事例への適合を重視する傾向が強くなる．

式(2.27)の目的関数を $\boldsymbol{w}$ で微分し，確率的勾配降下法を適用すると，重みベクトル $\boldsymbol{w}$ の更新式は次式で表される．

$$\boldsymbol{w}^{(t+1)} = (1 - \eta^{(t)}C)\boldsymbol{w}^{(t)} + \eta^{(t)}y^{(i)}\{1 - p(y^{(i)}|x^{(i)})\}\boldsymbol{\phi}(x^{(i)}), \quad i = t \bmod N \tag{2.28}$$

したがって，L2-正則化付きでロジスティック回帰モデルを学習するには，Algorithm 1 の重みベクトルの更新式を式(2.28)に変更するだけでよい．正則化を行わなかった場合の更新式(2.23)と比較すると，各反復で重みベクトル $\boldsymbol{w}$ を $(1 - \eta^{(t)}C)$ 倍するという処理が加わり，重みベクトルを縮小させようとする力が働く．

また，目的関数に重みベクトル $\boldsymbol{w}$ の 1-ノルムをペナルティ項として加えた L1-正則化(l_1-regularization)もよく用いられる．

$$E^{\mathrm{L1LR}}(\boldsymbol{w}) = -\mathcal{L}^{\mathrm{LR}} + C|\boldsymbol{w}| = -\sum_{i=1}^{N} \log p(y^{(i)}|x^{(i)}) + C|\boldsymbol{w}| \tag{2.29}$$

ここで，$C > 0$ は L1-正則化の強さを調整するハイパー・パラメータである．L1-正則化は，素性の重みの値を 0 に落とそうとする作用があることが知られている．学習の結果，素性の重みが 0 になったということは，その素性は無くてもよいと学習アルゴリズムが判断したことになるため，L1-正則化は学習と素性選択を同時に行う方法としても知られている．ただ，式(2.29)の目的関数は $w_k = 0$ において微分不可能であるため，そのままでは確率的勾配降下法を適用できない．しかし，FOrward-Backward Splitting(FOBOS)などの手法を用いることにより，Algorithm 1 に近い流れの擬似コードで重みベクトルを学習できる（Duchi and Singer, 2009）．

3. 線形多クラス分類器

本節では，入力 x に対して，L 個のラベル $\mathcal{Y} = \{1, 2, \ldots, L\}$ の中から 1 つのラベル $y \in \mathcal{Y}$ を

推定する線形多クラス分類器(linear multi-class classifier)を説明する．説明を簡単にするため，ラベル y は正の整数を取ることにするが，$y=1$ は「人名」，$y=2$ は「組織名」など，ラベルの番号に任意の分類カテゴリを割り当ててよい．線形二値分類器では単一の重みベクトル $\boldsymbol{w} \in \mathbb{R}^d$ を用いたが，線形多クラス分類器では L 個のラベルに対応する重みベクトル $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_L$ を用いる．線形二値分類器と同様に，入力素性ベクトル $\boldsymbol{\phi}(x) \in \mathbb{R}^d$ と重みベクトル $\boldsymbol{w}_y \in \mathbb{R}^d$ との内積を，入力 x をラベル y に分類するスコア $s(x,y)$ と定義する．

$$s(x,y) = \boldsymbol{w}_y^\intercal \boldsymbol{\phi}(x) \tag{3.1}$$

そして，入力 x に対するラベル $\hat{y}$ を次式で推定する．

$$\hat{y} = \operatorname*{argmax}_{y \in \mathcal{Y}} s(x,y) = \operatorname*{argmax}_{y \in \mathcal{Y}} \boldsymbol{w}_y^\intercal \boldsymbol{\phi}(x) \tag{3.2}$$

式(3.4)は，候補となる全てのラベル $y \in \mathcal{Y}$ に関してスコア $s(x,y)$ を計算し，最も高いスコアを与えたラベル $\hat{y}$ に分類するという，単純明快なルールである．素性関数 $\boldsymbol{\phi}(x)$ の設計は，2.1 節で説明した方針のままでよい．

線形多クラス分類器の原理は，式(3.1)と(3.2)で理解しておけばよいが，より一般的には次のように定式化される．

$$s(x,y) = \boldsymbol{\lambda}^\intercal \boldsymbol{f}(x,y) = \sum_{k=1}^{K} \lambda_k f_k(x,y) \tag{3.3}$$

$$\hat{y} = \operatorname*{argmax}_{y \in \mathcal{Y}} s(x,y) = \operatorname*{argmax}_{y \in \mathcal{Y}} \boldsymbol{\lambda}^\intercal \boldsymbol{f}(x,y) \tag{3.4}$$

この定式化での素性空間の次元数を K とすると，$\boldsymbol{\lambda} \in \mathbb{R}^K$ は重みベクトル，$\boldsymbol{f}(x,y) \in \mathbb{R}^K$ は入力 x とラベル y に関する素性ベクトルを表す．

式(3.1)と(3.2)よりも式(3.3)と(3.4)の定式化の方が一般性が高く，表現力も高いため，本稿では式(3.3)と(3.4)の定式化を採用する．しかし，式(3.3)と(3.4)の定式化では，素性空間が入力 x だけでなくラベル y にも依存するため，最初は分かりづらい．イメージを掴んでもらうため，式(2.3)に示した線形二値分類器の素性関数を多クラス分類に拡張し，ラベル $y=1$ の時のみに発火する素性関数の例を示す．

$$f_2(x,y) = \begin{cases} 1 & (x \text{ が大文字で始まる} \wedge y = 1) \\ 0 & (\text{それ以外の場合}) \end{cases} \tag{3.5}$$

f_2 は，x が大文字で始まり，かつラベルを $y=1$ と予測する場合に発火する素性である．さらに，「x が大文字で始まる」という特徴に関して，予測するラベル $y = 1, 2, \ldots, L$ 毎に素性関数を $f_2, f_3, \ldots, f_{L+1}$ などと定義する．すると，$\lambda_2, \lambda_3, \ldots, \lambda_{L+1}$ は「x が大文字で始まる」という特徴から，それぞれラベル $y = 1, 2, \ldots, L$ を予測するときの重みを表すようになる．

式(3.3)と(3.4)の定式化は，$\boldsymbol{f}(x,y)$ と $\boldsymbol{\lambda}$ を次のように定義すると，式(3.1)と(3.2)の定式化と等価になる．

$$\boldsymbol{f}(x,y) = \underbrace{\mathbf{0} \oplus \cdots \oplus \mathbf{0}}_{y-1} \oplus \boldsymbol{\phi}(x) \oplus \underbrace{\mathbf{0} \oplus \cdots \oplus \mathbf{0}}_{L-y} \tag{3.6}$$

$$\boldsymbol{\lambda} = \boldsymbol{w}_1 \oplus \boldsymbol{w}_2 \oplus \cdots \oplus \boldsymbol{w}_L \tag{3.7}$$

ただし，$\mathbf{0} \in \mathbb{R}^d$，$\oplus$ はベクトルの連結を表す．すなわち，素性空間を d 次元ずつ L 個のグループに分け，素性空間の各グループをラベル $y \in \mathcal{Y}$ に対応付ける($K = dL$)．重みベクトル $\boldsymbol{\lambda}$ は

各ラベルに対応した重みベクトル $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_L$ を並べたものである．素性関数 $\boldsymbol{f}(x, y)$ は，ラベル y に対応する素性空間グループで入力 x の素性ベクトル $\boldsymbol{\phi}(x)$ を展開し，その他のグループでは $\mathbf{0}$ とする．したがって，$s(x, y) = \boldsymbol{\lambda}^\intercal \boldsymbol{f}(x, y) = \boldsymbol{w}_y^\intercal \boldsymbol{\phi}(x)$ となり，式(3.1)と式(3.3)が一致する．

実際に多クラス分類器を構築するときは，素性関数 $\boldsymbol{f}(x, y)$ を直接設計することは稀である．代わりに，入力 x に関する素性ベクトル $\boldsymbol{\phi}(x)$ を定義し，式(3.6)の変換を用いて多クラス分類用の素性関数 $\boldsymbol{f}(x, y)$ を自動的に導出することが多い．したがって，多クラス分類器のツールを使うだけであれば素性関数 $\boldsymbol{f}(x, y)$ を意識する必要はない．しかし，式(3.3)と(3.4)の定式化を用いると，次式のように複数のラベル($y = 1$ or 2)で共通に発火する素性を定義できる(素性の次元の番号 2045 は適当である)．

$$f_{2045}(x, y) = \begin{cases} 1 & (x \text{ が大文字で始まる} \wedge (y = 1 \vee y = 2)) \\ 0 & (\text{それ以外の場合}) \end{cases} \tag{3.8}$$

例えば，$y = 1$ を人名，$y = 2$ を組織名であることにすると，これらの固有名詞で共通に発火する素性を定義したことになる．

3.1 多クラスロジスティック回帰

(二値分類である)ロジスティック回帰では，素性ベクトルと重みベクトルの内積で計算されるスコアを，シグモイド関数(sigmoid function)を使って確率値に変換していた．多クラスロジスティック回帰では，ラベル y 毎に計算されるスコア $s(x, y)$ にソフトマックス関数(softmax function)を適用し，入力 x に対してラベル y が予測される条件付き確率を求める．

$$p(y|x) = \frac{\exp s(x, y)}{\sum_{y' \in \mathcal{Y}} \exp s(x, y')} = \frac{\exp\left(\boldsymbol{\lambda}^\intercal \boldsymbol{f}(x, y)\right)}{\sum_{y' \in \mathcal{Y}} \exp\left(\boldsymbol{\lambda}^\intercal \boldsymbol{f}(x, y')\right)} \tag{3.9}$$

式(3.9)の分母は，分配関数(partition function)と呼ばれる．

$$Z(x) = \sum_{y' \in \mathcal{Y}} \exp\left(\boldsymbol{\lambda}^\intercal \boldsymbol{f}(x, y')\right) \tag{3.10}$$

なお，多クラスロジスティック回帰モデルは，最大エントロピー法(maximum entropy modeling)(Berger et al., 1996)とも呼ばれる．

多クラスロジスティック回帰モデルの学習では，(二値分類の)ロジスティック回帰モデルの理論とアルゴリズムをそのまま流用できる．言い換えれば，各事例の対数尤度の勾配さえ求めることができれば，確率的勾配降下法による最尤推定や事後確率最大化(正則化)の手順は全く同じである．ここでは，最尤推定について説明する．学習データ $\mathcal{D} = (x^{(i)}, y^{(i)})_{i=1}^N$ の各事例 $(x^{(i)}, y^{(i)})$ の対数尤度を $l(x^{(i)}, y^{(i)}) = \log p(y^{(i)}|x^{(i)})$ とすると，学習データ全体の対数尤度 $\mathcal{L}^{\mathrm{MLR}}$，最小化すべき目的関数 $E^{\mathrm{MLR}}(\boldsymbol{\lambda})$ は，

$$\mathcal{L}^{\mathrm{MLR}} = \log \prod_{i=1}^N p(y^{(i)}|x^{(i)}) = \sum_{i=1}^N \log p(y^{(i)}|x^{(i)}) = \sum_{i=1}^N l(x^{(i)}, y^{(i)}) \tag{3.11}$$

$$E^{\mathrm{MLR}}(\boldsymbol{\lambda}) = -\mathcal{L}^{\mathrm{MLR}} = -\sum_{i=1}^N l(x^{(i)}, y^{(i)}). \tag{3.12}$$

確率的勾配降下法を適用するため，各事例の対数尤度 $l(x, y)$ を展開する．

$$
\begin{aligned}
l(x,y) &= \log p(y|x) \\
&= \boldsymbol{\lambda}^{\mathsf{T}} \boldsymbol{f}(x,y) - \log \sum_{y' \in \mathcal{Y}} \exp\left(\boldsymbol{\lambda}^{\mathsf{T}} \boldsymbol{f}(x,y')\right) \\
&= \sum_{k=1}^{K} \lambda_k f_k(x,y) - \log \sum_{y' \in \mathcal{Y}} \exp\left(\sum_{k=1}^{K} \lambda_k f_k(x,y')\right)
\end{aligned} \tag{3.13}
$$

そして，式(3.13)を λ_k に関して偏微分する．第 1 項では k に関する部分のみを考えればよい．第 2 項には合成関数，対数関数，指数関数の微分公式を適用する．

$$
\begin{aligned}
\frac{\partial l(x,y)}{\partial \lambda_k} &= \frac{\partial}{\partial \lambda_k} \left\{ \sum_{k=1}^{K} \lambda_k f_k(x,y) - \log \sum_{y' \in \mathcal{Y}} \exp\left(\sum_{k=1}^{K} \lambda_k f_k(x,y')\right) \right\} \\
&= f_k(x,y) - \frac{\sum_{y' \in \mathcal{Y}} \left\{ \exp\left(\sum_{k=1}^{K} \lambda_k f_k(x,y')\right) \cdot f_k(x,y') \right\}}{\sum_{y'' \in \mathcal{Y}} \exp\left(\sum_{k=1}^{K} \lambda_k f_k(x,y'')\right)}
\end{aligned} \tag{3.14}
$$

$$
= f_k(x,y) - \sum_{y' \in \mathcal{Y}} p(y'|x) f_k(x,y') \tag{3.15}
$$

ゆえに，k 番目の素性に関する確率的勾配降下法の更新式は，

$$
\lambda_k^{(t+1)} = \lambda_k^{(t)} + \eta^{(t)} \left\{ f_k(x^{(i)}, y^{(i)}) - \sum_{y' \in \mathcal{Y}} p(y'|x^{(i)}) f_k(x^{(i)}, y') \right\}, \quad i = t \bmod N \tag{3.16}
$$

式(3.15)および(3.16)の解釈を考えてみたい．簡単のため，素性 f_k はどれか 1 つのラベルに関してのみ発火し(1 を返す)，それ以外のラベルに関しては発火しない(0 を返す)こととする．もし，訓練事例 $(x^{(i)}, y^{(i)})$ に関して素性 f_k が発火する場合，$f_k(x^{(i)}, y^{(i)}) = 1$ かつ $\forall y' \neq y^{(i)} : f_k(x^{(i)}, y') = 0$ であるから，

$$
\frac{\partial l(x^{(i)}, y^{(i)})}{\partial \lambda_k} = 1 - p(y^{(i)}|x^{(i)}) \geq 0. \tag{3.17}
$$

式(3.16)に当てはめて考えると，予測誤差に応じて素性 f_k の重みが増加する．これは，f_k が訓練事例に関して発火しているため，その信頼度を増やそうとしていると解釈できる．

一方，訓練事例 $(x^{(i)}, y^{(i)})$ に関して素性 f_k は発火しないものの，正解のラベル $y^{(i)}$ を別のラベル $\tilde{y}$ に置き換えると発火する場合を考える．これは，素性 f_k は入力 $x^{(i)}$ の特徴を捉えようとするが，正解とは異なるラベル $\tilde{y} \neq y^{(i)}$ を予測しようとする状況である．このとき，$f_k(x^{(i)}, \tilde{y}) = 1$ かつ $\forall y' \neq \tilde{y} : f_k(x^{(i)}, y') = 0$，よって $f_k(x^{(i)}, y^{(i)}) = 0$ であるから，

$$
\frac{\partial l(x^{(i)}, y^{(i)})}{\partial \lambda_k} = 0 - p(\tilde{y}|x^{(i)}) \leq 0 \tag{3.18}
$$

訓練事例 $(x^{(i)}, y^{(i)})$ は，入力 $x^{(i)}$ に対してラベル $\tilde{y}$ を予測することは間違いであることを示唆しているため，$p(\tilde{y}|x^{(i)})$ は予測誤差を表す．式(3.16)に当てはめて考えると，予測誤差に応じて素性 f_k の重みが減少する．これは，f_k が訓練事例に関して発火せず，正解とは異なるラベルで発火してしまうため，その信頼度を下げようとしていると解釈できる．

なお，訓練事例 $(x^{(i)}, y^{(i)})$ の入力 $x^{(i)}$ に対して，どのようなラベル $y' \in \mathcal{Y}$ に関しても素性 f_k が発火しない場合，$\forall y' \in \mathcal{Y} : f_k(x^{(i)}, y') = 0$ であるから，

$$
\frac{\partial l(x^{(i)}, y^{(i)})}{\partial \lambda_k} = 0. \tag{3.19}
$$

これは，訓練事例 $(x^{(i)}, y^{(i)})$ の入力 $x^{(i)}$ とは全く関連がない素性 f_k に関しては，その重み λ_k を更新しないことを表している．

Algorithm 2 に，確率的勾配降下法による多クラスロジスティック回帰モデルの学習の擬似コードを示す．

Algorithm 2: SGD による多クラスロジスティック回帰モデルの学習
入力: 学習データ $\mathcal{D} = (x^{(i)}, y^{(i)})_{i=1}^{N}$
出力: 重みベクトル $\boldsymbol{\lambda} \in \mathbb{R}^K$

$t \leftarrow 1$;
$\boldsymbol{\lambda} = \mathbf{0}$;
for epoch $\leftarrow 1$ **to** T **do**
  for $i \leftarrow 1$ **to** N **do**
    $\eta \leftarrow 1/t$(※他の計算方法でも可);
    for $k \leftarrow 1$ **to** K **do**
      $\lambda_k \leftarrow \lambda_k + \eta \left\{ f_k(x^{(i)}, y^{(i)}) - \sum_{y' \in \mathcal{Y}} p(y'|x^{(i)}) f_k(x^{(i)}, y') \right\}$;
    end
    $t \leftarrow t + 1$;
  end
end

4. 条件付き確率場

多クラスロジスティック回帰では，ラベル y は単一の確率変数であった．これに対し，条件付き確率場(conditional random fields)（Lafferty et al., 2001)では，複数の確率変数を同時に予測する．予測される確率変数は入力 x に依存するだけでなく，系列や木構造などのグラフ構造上でマルコフ性に従うと仮定する．入力 x も構造を持つと仮定してもよいが，入力と出力の構造は一致しなくてもよい．

自然言語の単語や文は無秩序に並んでいるのではなく，何らかの規則性を持っている．その規則性を分類モデルに取り込むことで，タスクの予測精度の向上が期待できる．例えば，英語では「文頭の単語は名詞になりやすい」「冠詞の後には名詞や形容詞が続きやすい」などの規則性があり，この規則性は単語列から品詞列を推定するのに役立つ．

本稿では，条件付き確率場の代表例として，系列 $\boldsymbol{x} = x_1, \ldots, x_M$ が与えられた時，ラベル列 $\boldsymbol{y} = y_1, \ldots, y_M$ を予測する系列ラベリング(sequential labeling)問題を扱う(M は系列の要素数)．先ほどの品詞タグ付けの例では，入力 $\boldsymbol{x}$ が単語列，ラベル列 $\boldsymbol{y}$ が品詞ラベル列となる．ここで，隣り合うラベル同士 y_m, y_{m+1} $(m = 1, \ldots, M-1)$ の依存関係(線形連鎖一次マルコフ性)を考慮し，ラベル列 $\boldsymbol{y}$ を予測することを考える．品詞タグ付けの例では，「直前の品詞の予測結果が冠詞ならば現在の単語の品詞は名詞や形容詞になりやすい」などの依存関係を考慮することになる．

条件付き確率場の予測モデルは，多クラス線形分類器の入力 x を系列 $\boldsymbol{x}$ に，ラベル y をラベル列 $\boldsymbol{y}$ に置き換えたものになる．線形多クラス分類器(式(3.3))と同様に，入力系列 $\boldsymbol{x}$ のラベル系列 $\boldsymbol{y}$ を予測するスコア $s(\boldsymbol{x}, \boldsymbol{y})$ を，素性ベクトル $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^K$ と重みベクトル $\boldsymbol{\lambda} \in \mathbb{R}^K$ の内積で計算する．素性関数 $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y})$ は，系列 $\boldsymbol{x}$ とラベル列 $\boldsymbol{y}$ の全体から素性ベクトルを取り出す関数である．

$$s(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{\lambda}^\mathsf{T} \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{k=1}^{K} \lambda_k f_k(\boldsymbol{x}, \boldsymbol{y}) \tag{4.1}$$

式(3.2)と同様に，入力系列 $\boldsymbol{x}$ に対するラベル列 $\hat{\boldsymbol{y}}$ を次式で推定する．

$$\hat{\boldsymbol{y}} = \operatorname*{argmax}_{\boldsymbol{y} \in \mathcal{Y}} s(\boldsymbol{x}, \boldsymbol{y}) = \operatorname*{argmax}_{\boldsymbol{y} \in \mathcal{Y}} \boldsymbol{\lambda}^\mathsf{T} \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}) \tag{4.2}$$

ここで，各ラベル $y_m \in \boldsymbol{y}$ が取りうるラベルの集合を $\mathcal{Y}$ とすると，ラベル列の候補集合は $\mathcal{Y} = \mathcal{Y}^M$ である．

式(3.2)で考慮すべきラベルの候補数は $|\mathcal{Y}|$ であったが，式(4.2)ではラベル列の候補は $|\mathcal{Y}|^M$ 通りである．例えば，ラベルの種類数 $|\mathcal{Y}| = 9$ とし，長さ $M = |\boldsymbol{x}| = 10$ のラベル列を予測する場合でも，ラベル列の候補数は $3,486,784,401$ に膨れ上がる．したがって，式(4.1)をラベル列の全候補 $\mathcal{Y}$ に関して計算し，最大のスコアを与えたラベル系列 $\hat{\boldsymbol{y}}$ を求めることは事実上不可能である．式(4.2)を効率よく求めるアルゴリズムは，4.2 節で紹介する．

式(3.9)と同様に，ラベル列 $\boldsymbol{y}$ 毎に計算されるスコア $s(\boldsymbol{x}, \boldsymbol{y})$ にソフトマックス関数を適用し，入力系列 $\boldsymbol{x}$ に対してラベル列 $\boldsymbol{y}$ が予測される条件付き確率を求める．

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{\exp s(\boldsymbol{x}, \boldsymbol{y})}{\sum_{\boldsymbol{y'} \in \mathcal{Y}} \exp s(\boldsymbol{x}, \boldsymbol{y'})} = \frac{\exp \left(\boldsymbol{\lambda}^\mathsf{T} \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y})\right)}{\sum_{\boldsymbol{y'} \in \mathcal{Y}} \exp \left(\boldsymbol{\lambda}^\mathsf{T} \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y'})\right)} \tag{4.3}$$

ここで，式(4.3)の分母(分配関数)の計算には，ラベル列の全候補 $\boldsymbol{y'} \in \mathcal{Y}$ に関する内積とその和が必要である．

$$Z(\boldsymbol{x}) = \sum_{\boldsymbol{y'} \in \mathcal{Y}} \exp \left(\boldsymbol{\lambda}^\mathsf{T} \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y'})\right) \tag{4.4}$$

式(4.4)を愚直に計算すると，指数オーダー($|\mathcal{Y}|^M$ 回)の内積計算と加算演算が必要になり，分配関数の値を求めることも事実上不可能である．式(4.4)を効率よく計算する方法は，4.3 節で説明する．

条件付き確率場の学習では，(多クラス)ロジスティック回帰モデルの学習の理論とアルゴリズムをそのまま流用できる．すなわち，各訓練事例の対数尤度さえ求めることができれば，確率的勾配降下法による最尤推定や事後確率最大化(正則化)の手順は全く同じである．学習データ $\mathcal{D} = (\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)})_{i=1}^N$ の各事例 $(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)})$ の対数尤度を $l(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}) = \log p(\boldsymbol{y}^{(i)}|\boldsymbol{x}^{(i)})$ とすると，学習データ全体の対数尤度 $\mathcal{L}^{\mathrm{MLR}}$ は，

$$\mathcal{L}^{\mathrm{MLR}} = \log \prod_{i=1}^{N} p(\boldsymbol{y}^{(i)}|\boldsymbol{x}^{(i)}) = \sum_{i=1}^{N} \log p(\boldsymbol{y}^{(i)}|\boldsymbol{x}^{(i)}) = \sum_{i=1}^{N} l(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}) \tag{4.5}$$

確率勾配降下法による重みベクトル $\boldsymbol{\lambda}$ の更新式は，式(2.17)と同一である．

$$\boldsymbol{\lambda}^{(t+1)} = \boldsymbol{\lambda}^{(t)} + \eta^{(t)} \frac{\partial l(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)})}{\partial \boldsymbol{\lambda}^{(t)}}, \quad i = t \bmod N \tag{4.6}$$

そこで，式(4.6)の更新式で必要な勾配を求める．多クラスロジスティック回帰から条件付き確率場に拡張する際に変更したのは，入力とラベルの型だけなので，式(3.15)の導出を再利用すると，重み λ_k に関する偏微分は次式で与えられる．

$$\frac{\partial l(\boldsymbol{x}, \boldsymbol{y})}{\partial \lambda_k} = f_k(\boldsymbol{x}, \boldsymbol{y}) - \sum_{\boldsymbol{y'} \in \mathcal{Y}} p(\boldsymbol{y'}|\boldsymbol{x}) f_k(\boldsymbol{x}, \boldsymbol{y'}) \tag{4.7}$$

式(4.7)の第 2 項は，ラベル列の全候補 $\boldsymbol{y'} \in \mathcal{Y}$ において現在の確率モデルが素性関数 f_k を発火

させる期待値である．ラベル列の全候補 $\boldsymbol{y}' \in \mathcal{Y}$ が必要であることに加えて，素性関数 f_k が発火するかどうかが $\boldsymbol{y}'$ に依存しているため，この値を求めることも容易ではない．

まとめると，条件付き確率場は多クラスロジスティック回帰の入力 x を入力系列 $\boldsymbol{x}$ に，ラベル y をラベル列 $\boldsymbol{y}$ に置換しただけのモデルであるが，計算上の困難が 3 つある．

(1) スコアを最大にするラベル列 $\hat{\boldsymbol{y}}$ の予測: 式(4.2)
(2) 分配関数 $Z(\boldsymbol{x})$ の計算: 式(4.4)
(3) 素性関数のモデルにおける発火の期待値: 式(4.7)

これらを効率よく計算するための鍵は，いずれも，ラベル列の一次マルコフ性に基づく素性関数 $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y})$ の設計と，動的計画法である．以降では，これらについて説明していく．

4.1 線形連鎖 1 次マルコフ素性

本稿では，入力系列 $\boldsymbol{x}$ が与えられた時，ラベル列に関して一次マルコフ性を仮定する．すなわち，位置 m のラベル y_m の予測結果は，入力 x_m，隣接するラベル y_{m-1}，および y_{m+1} に依存すると仮定する．この依存関係は，x_m, y_{m-1}, y_m を引き数とした素性関数で記述できる．例えば，現在の単語が “Ltd” で，直前の単語が組織名の途中(I-ORG)で，現在の単語も組織名(I-ORG)である場合に発火する素性は，次式で表される．

$$\pi_k^{(\mathrm{xyy})}(x_m, y_{m-1}, y_m) = \begin{cases} 1 & (x_m \text{ が “Ltd”} \wedge y_{m-1} \text{ が I-ORG} \wedge y_m \text{ が I-ORG}) \\ 0 & (\text{それ以外の場合}) \end{cases} \tag{4.8}$$

なお，この素性関数は 1 つの入力 x_m，2 つのラベル y_{m-1}, y_m に依存するので，(xyy) という印を付けてある．

また，x_m，y_m，y_{m-1} のどれかに依存しない素性を考えてもよい．例えば，直前のラベル y_{m-1} に関わらず，現在の単語 “Ltd” と組織名(I-ORG)の依存関係を表現する素性は，次式で表される．

$$\pi_k^{(\mathrm{xy})}(x_m, y_m) = \begin{cases} 1 & (x_m \text{ が “Ltd”} \wedge y_m \text{ が I-ORG}) \\ 0 & (\text{それ以外の場合}) \end{cases} \tag{4.9}$$

さらに，位置 m を任意とし，隣接するラベル間の事象のみを記述した素性は，次式で表される．

$$\pi_k^{(\mathrm{yy})}(y_{m-1}, y_m) = \begin{cases} 1 & (y_{m-1} \text{ が B-PERSON} \wedge y_m \text{ が I-PERSON}) \\ 0 & (\text{それ以外の場合}) \end{cases} \tag{4.10}$$

線形連鎖一次マルコフ素性に基づく条件付き確率場では，式(4.8)，式(4.9)，式(4.10)のいずれかの型を持つ素性関数を定義することになる．

ところで，線形多クラス分類器では，入力 x と予測ラベル y の両方で条件付けされた素性関数 $\boldsymbol{f}(x, y)$ を直接設計することは稀で，入力 x に関する素性ベクトル $\boldsymbol{\phi}(x)$ を定義し，素性関数 $\boldsymbol{f}(x, y)$ を自動的に導出することを説明した．条件付き確率場でも，入力 x に関する素性ベクトル $\boldsymbol{\phi}(x)$ の設計に注力し，式(4.8)，式(4.9)，式(4.10)の形の素性関数を自動的に導出するのが一般的である．

例えば，式(4.9)の形は多クラス線形分類器のものと同一であり，(3)節で説明した方法で自動的に導出できる．また，式(4.10)の形の素性は，すべてのラベルの組み合わせ $(y_{m-1}, y_m) \in \mathcal{Y}^2$ に対して定義するか，訓練事例に出現する隣り合うラベルの組み合わせに対して定義すればよい．式(4.8)の形の素性は，以上の方法を組み合わせることによって導出できる．このように，条件付き確率場のツールを使うだけであれば，素性関数を直接設計しなくてもよい．

さて，式(4.8)の形の素性関数を K_1 個定義し，その全ての素性関数の値を K_1 次元ベクトルで返す素性関数ベクトルを $\boldsymbol{\pi}^{(\mathrm{xyy})}(x_m, y_{m-1}, y_m) \in \mathbb{R}^{K_1}$ と書く．同様に，式(4.9)と式(4.10)の素性関数をそれぞれ，K_2 個，K_3 個定義し，素性関数ベクトルで表現したものを，それぞれ，$\boldsymbol{\pi}^{(\mathrm{xy})}(x_m, y_m) \in \mathbb{R}^{K_2}$，$\boldsymbol{\pi}^{(\mathrm{yy})}(y_{m-1}, y_m) \in \mathbb{R}^{K_3}$ と書く(ただし，$K = K_1 + K_2 + K_3$)．線形連鎖一次マルコフ素性は，式(4.8)，式(4.9)，式(4.10)のいずれかで表されるので，時刻 m における素性関数ベクトル $\boldsymbol{\pi}(x_m, y_{m-1}, y_m) \in \mathbb{R}^K$ は次式で表される．

$$\boldsymbol{\pi}(x_m, y_{m-1}, y_m) = \boldsymbol{\pi}^{(\mathrm{xyy})}(x_m, y_{m-1}, y_m) \oplus \boldsymbol{\pi}^{(\mathrm{xy})}(x_m, y_m) \oplus \boldsymbol{\pi}^{(\mathrm{yy})}(y_{m-1}, y_m) \tag{4.11}$$

$\boldsymbol{\pi}(x_m, y_{m-1}, y_m)$ は，位置 m における素性ベクトルを表すので，局所素性ベクトルと呼ぶ．この記法に基づくと，線形連鎖一次マルコフ素性による大域素性ベクトルは次式で表される．

$$\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{m=1}^{M} \boldsymbol{\pi}(x_m, y_{m-1}, y_m) \tag{4.12}$$

ただし，$m = 1$ のとき，y_0 は系列の先頭を表す特殊なシンボル BOS とし，x_1 と y_1 のみに依存する素性とするか，系列の先頭のラベルを捉える素性としてもよい．同様に，系列の末尾のラベルを捉える素性を導入してもよい．

4.2 ラベル列の推定

式(4.2)を効率よく求める方法を説明するため，以下の量を定義する．

$$r_{\boldsymbol{x}}(m, i, j) = \boldsymbol{\lambda}^{\mathsf{T}} \boldsymbol{\pi}(x_m, i, j) \tag{4.13}$$

$r_{\boldsymbol{x}}(m, i, j)$ は，入力系列 $\boldsymbol{x}$ が与えられた時，$y_{m-1} = i$ かつ $y_m = j$ の時に発火する素性関数の重みの和である．

式(4.11)，式(4.12)，式(4.13)より，式(4.1)は次のように展開できる．

$$s(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{\lambda}^{\mathsf{T}} \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{\lambda}^{\mathsf{T}} \sum_{m=1}^{M} \boldsymbol{\pi}(x_m, y_{m-1}, y_m) = \sum_{m=1}^{M} r_{\boldsymbol{x}}(m, y_{m-1}, y_m) \tag{4.14}$$

これは，系列全体の大域素性ベクトル $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y})$ を先に求めてから重み $\boldsymbol{\lambda}$ との内積を計算する代わりに，先に各時刻 $m = 1, 2, \ldots, M$ において局所素性ベクトル $\boldsymbol{\pi}(x_m, y_{m-1}, y_m)$ と重みベクトル $\boldsymbol{\lambda}$ の内積を求め，その和を計算しても結果が同じであることを示している．式(4.14)の計算は大変そうに見えるが，素性関数は式(4.8)，式(4.9)，式(4.10)のいずれかの形をしているため，入力 $\boldsymbol{x}$ の時刻 m における特徴量に対応した素性や，ラベルのペアに関する素性の次元番号を列挙し，その次元番号に対応する重み取り出し，その和を計算するだけでよい．

図 2 は，式(4.14)の計算をラティスとして図示したものである．横軸は時刻 $m = 1, 2, 3$，縦軸はラベル y を表し，図中の各ノード (m, j) は時刻 m のラベル $y_m = j$ である事象を表している．また，各点をつなぐ線(エッジ)は，$r_{\boldsymbol{x}}(m, y_{m-1}, y_m)$ の値を示している．この図を用いると，$m = 1$ から $m = 3$ に至る任意の経路がラベル列 $\boldsymbol{y}$ を表し，その経路上の重みの和がスコア $s(\boldsymbol{x}, \boldsymbol{y})$ である．したがって，式(4.2)を求める問題は，図 2 で $m = 1$ から $m = 3$ に至る経路の中で，スコアが最大となる経路を求める問題に帰着する．

この問題は，最短経路問題と同様に，以下の漸化式で求めることができる．

$$\psi_{\boldsymbol{x}}(m, j) = \begin{cases} r_{\boldsymbol{x}}(1, \mathsf{EOS}, j) & (m = 1 \text{ のとき}) \\ \max_{i \in \mathcal{Y}} \{\psi_{\boldsymbol{x}}(m-1, i) + r_{\boldsymbol{x}}(m, i, j)\} & (m > 1 \text{ のとき}) \end{cases} \tag{4.15}$$

ここで，$\psi_{\boldsymbol{x}}(m, j)$ は図 2 の左端からノード (m, j) に至る経路の中で，最大のスコアを記録した

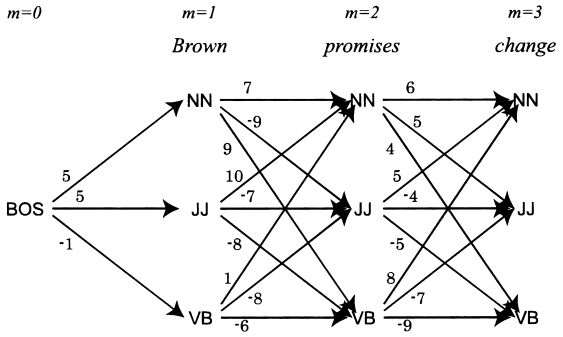


図 2. ラティスで表現した素性の重み．エッジの数字は $r_{\boldsymbol{x}}(m,i,j)$ の値を表す．

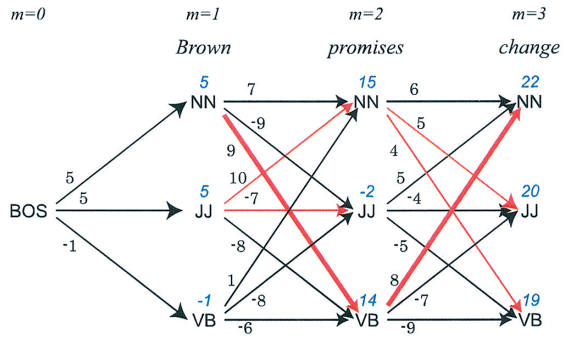


図 3. ビタビ・アルゴリズムの動作例．ノードは各時点 m におけるラベル y_m を表す．エッジの数字は $r_{\boldsymbol{x}}(m,i,j)$ の値を表す．ノードの左上の青色の数字は $\psi_{\boldsymbol{x}}(m,j)$ を，赤色のエッジはその値がどこに由来するのかを示している．

ものである．式(4.15)は，ノード $(m-1,i)$ を経由してノード (m,j) に至る経路のスコアの最大値は，$\psi_{\boldsymbol{x}}(m-1,i)+r_{\boldsymbol{x}}(m,i,j)$ で与えられることを利用し，経路のスコアを最大にする経由地 $(m-1,i)$ を選ぶことで，ノード (m,j) に至る経路のスコアの最大値を求めている．

式(4.15)の漸化式を使い，$\psi_{\boldsymbol{x}}(m,j)$ の値を $m=1$ から $m=M$ まで求めたとき，スコアが最大となる経路の終点 $(M,\hat{y}_M)$ は次式で表される．

$$\hat{y}_M = \operatorname*{argmax}_{i\in\mathcal{Y}} \psi_{\boldsymbol{x}}(M,i) \tag{4.16}$$

そして，この終点に至るまでの経路を逆向きに辿ることで，式(4.2)の解を求めることができる．すなわち，$m=M-1$ から $m=1$ まで，以下の漸化式を適用していけばよい．

$$\hat{y}_m = \operatorname*{argmax}_{i\in\mathcal{Y}} \{\psi_{\boldsymbol{x}}(m,i)+r_{\boldsymbol{x}}(m+1,i,\hat{y}_{m+1})\} \tag{4.17}$$

実際には，式(4.17)を計算しなくても，式(4.15)の漸化式を適用する際に選ばれた経由地への逆向きのリンクを保持しておけば，終点 $(M,\hat{y}_M)$ から逆向きのリンクを辿っていくだけでよい．この計算過程を示したのが図 3 である．図中のノード (m,i) の上に $\psi_{\boldsymbol{x}}(m,i)$ の値を示し，その値がどのノードから来たのか，赤色の矢印で示してある．式(4.16)に従い，$m=3$ のラベルを

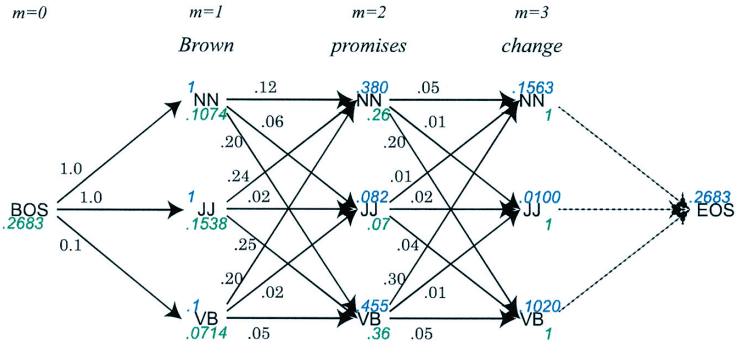


図 4. 前向き・後ろ向きアルゴリズムの計算例．ノードは各時点 m におけるラベル y_m を表す．エッジの数字は $\check{r}_{\boldsymbol{x}}(m,i,j)$ の値を表す．ノードの左上の青色の数字は $\alpha_{\boldsymbol{x}}(m,j)$ を，右下の緑色の数字は $\beta_{\boldsymbol{x}}(m,i)$ を表す．式(4.22)で求めた分配関数の値を BOS の右下に，式(4.24)で求めた分配関数の値を EOS の左上に示した．前向き・後ろ向きのどちらを使っても，分配関数の値が同じになることを確認できる．

NN と決定したら，そのノードに至る矢印を逆向きに辿ることで，式(4.2)の解，すなわち予測されるラベル列 $\hat{\boldsymbol{y}}$ を求めることができる．図 3 の場合は，NN VB NN が解となる．このアルゴリズムは，ビタビ・アルゴリズム(Viterbi algorithm)と呼ばれる．

4.3 分配関数の計算

次に，分配関数(式(4.4))を効率よく求める方法を説明する．まず，$r_{\boldsymbol{x}}(m,i,j)$ の指数を取った関数 $\check{r}_{\boldsymbol{x}}(m,i,j)$ を定義する．

$$\check{r}_{\boldsymbol{x}}(m,i,j) = \exp r_{\boldsymbol{x}}(m,i,j) \tag{4.18}$$

分配関数の定義を $\check{r}_{\boldsymbol{x}}(m,i,j)$ の式として書き直す．

$$\begin{aligned} Z(\boldsymbol{x}) &= \sum_{y_1,\dots,y_t \in \mathcal{Y}^L} \exp s(\boldsymbol{x},\boldsymbol{y}) \\ &= \sum_{y_1,\dots,y_t \in \mathcal{Y}^L} \prod_{m=1}^{M} \check{r}_{\boldsymbol{x}}(m,y_{m-1},y_m) \end{aligned} \tag{4.19}$$

式(4.19)を展開し，共通部分を $m=1$ から括り出すと，次のようになる．

$$Z(\boldsymbol{x}) = \sum_{y_1=1}^{L} \check{r}_{\boldsymbol{x}}(1,\mathsf{BOS},y_1) \sum_{y_2=1}^{L} \check{r}_{\boldsymbol{x}}(2,y_1,y_2) \sum_{y_3=1}^{L} \dots \sum_{y_M=1}^{L} \check{r}_{\boldsymbol{x}}(M,y_{M-1},y_M) \tag{4.20}$$

ここで，式(4.20)の形に着目すると，右側($m=M$)から左側($m=1$)に向かって，$\check{r}_{\boldsymbol{x}}(m,i,j)$ の和を求め，その値を左側に伝播させながら，和の計算を進めればよいことが分かる．

この計算過程を図示したのが図 4 である．あるノード (m,i) の右側から終端($m=M$)に至る経路上のエッジにある $\check{r}_{\boldsymbol{x}}(m,i,j)$ の積を計算し，全ての経路で和を取ったものを $\beta_{\boldsymbol{x}}(m,i)$ と定義すると，以下の漸化式が成り立つ．

$$\beta_{\boldsymbol{x}}(m,i) = \begin{cases} 1 & (m=M \text{ のとき}) \\ \displaystyle\sum_{j=1}^{L} \check{r}_{\boldsymbol{x}}(m,i,j)\beta_{\boldsymbol{x}}(m+1,j) & (m<M \text{ のとき}) \end{cases} \tag{4.21}$$

この漸化式で各ノードの $\beta_{\boldsymbol{x}}(m,i)$ の値を求めたものを，ノードの下側に緑色で示した．式(4.20)との対比により，分配関数は次式で求まることが分かる．

$$Z(\boldsymbol{x}) = \beta_{\boldsymbol{x}}(0, \mathsf{BOS}) \tag{4.22}$$

一方，式(4.19)を展開するとき，$m = M$ から括り出すと，次のようになる．

$$Z(\boldsymbol{x}) = \sum_{y_M=1}^{L} \sum_{y_{M-1}=1}^{L} \check{r}_{\boldsymbol{x}}(M, y_{M-1}, y_M) \sum_{y_{M-2}=1}^{L} \cdots \sum_{y_1=1}^{L} \check{r}_{\boldsymbol{x}}(1, y_0, y_1)$$

ここで，左端($m = 1$)からノード (m,j) に至る経路上にあるエッジの $\check{r}_{\boldsymbol{x}}(m,i,j)$ の積を計算し，全ての経路で和を取ったものを $\alpha_{\boldsymbol{x}}(m,j)$ と定義すると，以下の漸化式が得られる．

$$\alpha_{\boldsymbol{x}}(m,j) = \begin{cases} \check{r}_{\boldsymbol{x}}(1, \mathsf{BOS}, j) & (m = 1 \text{ のとき}) \\ \displaystyle\sum_{i=1}^{L} \check{r}_{\boldsymbol{x}}(m,i,j)\alpha_{\boldsymbol{x}}(m-1,i) & (1 < m \text{ のとき}) \end{cases} \tag{4.23}$$

この漸化式で各ノードの $\alpha_{\boldsymbol{x}}(m,i)$ の値を求めたものを，ノードの上側に青色で示した．式(4.20)との対比により，分配関数は次式で求めることもできる．

$$Z(\boldsymbol{x}) = \sum_{i=1}^{L} \alpha_{\boldsymbol{x}}(M,i) \tag{4.24}$$

分配関数の値を求めるだけであれば，$\alpha_{\boldsymbol{x}}(m,i)$ と $\beta_{\boldsymbol{x}}(m,i)$ のどちらを用いてもよい．ただ，次節で説明する周辺確率の計算には，これらの両方の値が必要になる．このように，$\alpha_{\boldsymbol{x}}(m,i)$, $\beta_{\boldsymbol{x}}(m,i)$ の値を求めるアルゴリズムのことを，前向き・後ろ向きアルゴリズムと呼ぶ．

4.4 素性関数のモデルにおける発火の期待値

最後に，式(4.7)の計算方法を説明する．式(4.11)を用い，大域素性ベクトルを局所素性ベクトルに変換すると，式(4.7)は以下のように変形できる．

$$\begin{aligned} \frac{\partial l(\boldsymbol{x},\boldsymbol{y})}{\partial \lambda_k} &= f_k(\boldsymbol{x},\boldsymbol{y}) - \sum_{\boldsymbol{y'} \in \mathcal{Y}} p(\boldsymbol{y'}|\boldsymbol{x}) f_k(\boldsymbol{x},\boldsymbol{y'}) \\ &= \sum_{m=1}^{M} \left\{ \pi_k(x_m, y_{m-1}, y_m) - \sum_{i \in \mathcal{Y}} \sum_{j \in \mathcal{Y}} p(i,j|\boldsymbol{x},m) \pi_k(x_m, i, j) \right\} \end{aligned} \tag{4.25}$$

ここで，$p(i,j|\boldsymbol{x},m)$ は，与えられた入力系列 $\boldsymbol{x}$ に対して，現在のモデルで出力の系列を予測したとき，時刻 $m-1$ のラベルが i で，時刻 m のラベルが j と予測される確率(周辺確率)を表す．したがって，任意の m, i, j に関して，$p(i,j|\boldsymbol{x},m)$ が計算できればよい．

ここで，$\alpha_{\boldsymbol{x}}(m,i)$ が左端($m = 1$)からノード (m,i) に至る経路のスコアの和，$\beta_{\boldsymbol{x}}(m,i)$ がノード (m,i) から右端($m = M$)に至るスコアの和であることを利用すると，$p(i,j|\boldsymbol{x},m)$ を次式で求めることができる．

$$\begin{aligned} p(i,j|\boldsymbol{x},m) &= \frac{(m-1,i)\text{—}(m,j) \text{ を通過する全経路のスコアの和}}{Z(\boldsymbol{x})} \\ &= \frac{\alpha_{\boldsymbol{x}}(m-1,i)\check{r}_{\boldsymbol{x}}(m-1,i,j)\beta_{\boldsymbol{x}}(m,j)}{Z(\boldsymbol{x})} \end{aligned} \tag{4.26}$$

ゆえに，4.3 節で分配関数を計算する時に用いた $\alpha_{\boldsymbol{x}}(m,i)$ および $\beta_{\boldsymbol{x}}(m,i)$ の値を保存しておけば，周辺確率 $p(i,j|\boldsymbol{x},m)$ の値を簡単に求めることができる．

4.5 その他の研究動向

2001 年の登場以降，条件付き確率場は様々な言語処理タスクに適用されてきた．典型例として，浅い句構造解析（Sha and Pereira, 2003），固有表現抽出（McCallum and Li, 2003），形態素解析（Kudo et al., 2004），情報抽出（Sarawagi and Cohen, 2005），構文解析（Finkel et al., 2008），ゾーニング（Hirohata et al., 2008）などがある．条件付き確率場を実装したツールとして，CRF++ [1], CRFsuite [2], Wapiti [3] などが有名である．

高性能の系列予測器を作るには，大量の訓練データを用意しなければならない．そこで，条件付き確率場の理論をうまく利用して，訓練データの構築を支援する研究もある．坪井ら（Tsuboi et al., 2008）は，入力系列の一部のみに正解ラベルを付与した訓練データや，正解を一つに絞り込むことができずに複数の正解ラベルが付与された訓練データから，重みベクトルを学習する手法を提案した．本手法は後に中国語の単語分割に適用され（Liu et al., 2014），CRFsuite のソースコードをベースにした実装が公開されている [4]．

Settles and Craven（2008）は，少量の訓練データと正解が付与されていない大量の事例があるとき，どの事例に正解を付与すべきかを示唆する能動学習（active learning）の戦略を比較・検討した．この研究では，少量の訓練データで学習したモデルを使い，正解が付与されていない事例の入力 $\boldsymbol{x}$ に対してラベル列 $\hat{\boldsymbol{y}}$ を予測し，予測されたラベル列の確率推定値 $p(\hat{\boldsymbol{y}}|\boldsymbol{x})$ が低い事例の正解を付与したり（least confidence），ラベルの周辺確率から計算されるラベル列のエントロピーの高い事例に対して正解を付与する戦略（token entropy）などを検討している．CRFsuite では，予測されたラベル列の確率は `-p`，`--probability` オプションで，予測されたラベル列の周辺確率は `-i`，`--marginal` オプションで得ることができる．また，CRFsuite の API（`Tagger::marginal`）を呼び出すことで，全てのラベルに関する周辺確率を計算できる．これらを活用することで，Settles らが比較・検討した能動学習の戦略を実装できる．

本稿では，ラベルに関する 1 次マルコフ性を仮定し，隣り合うラベルの依存関係を考慮した．しかし，言語のデータを扱っていると，2 単語先や 3 単語先など，距離が離れたラベルの依存関係をモデル化したくなることがある．最も単純な解決策は，2 次（連続する 3 個のラベル）や 3 次（連続する 4 個のラベル）など，2 次以上のマルコフ性を仮定した素性関数を導入することである．ただ，素性の次数を l とすると，ビタビ・アルゴリズムや前向き・後ろ向き・アルゴリズムの計算量は $|\mathcal{Y}|^{l+1}$ であるため，次数を上げると計算量が急増してしまう．代わりに，連続する複数のラベルをひと塊として扱うセミ・マルコフ条件付き確率場（Sarawagi and Cohen, 2005）が提案されている．また，入力列から出力列を予測するまでの間に，潜在変数のラベル列を導入することで，離れた距離の依存関係をモデル化しようとする研究もある（Morency et al., 2007）（Sun et al., 2008）．なお，CRFsuite の実装を拡張して，高次のマルコフ性，セミ・マルコフ性，木構造予測などを実現した実装 [5] も公開されている．

最近では，Recurrent Neural Network（RNN）や Long Short-Term Memory（LSTM）などの深層ニューラルネットワークに条件付き確率場のアイディアを導入し，品詞タグ付けや固有表現抽出を実現した研究も報告されている（Zhou and Xu, 2015）（Lample et al., 2016）．これらの研究は，潜在変数のラベル列の代わりに中間層の分散表現を用いるとともに，LSTM の記憶セルなどで離れた距離の依存関係をモデル化している．また，深層ニューラルネットワークに基づく条件付き確率場では，タスクに依存した素性関数を人間が設計しなくても，文字や単語の分散表現を学習することにより，特徴抽出を自動化できる．実際，Lample et al.（2016）は文字に関する素性や辞書などの外部知識を用いずに，深層ニューラルネットワークだけで獲得した素性を用いて，固有表現抽出の最高性能を達成できることを報告した．

5. おわりに

本稿では，系列データにおける条件付き確率場の理論と実践を解説した．条件付き確率場は，多クラスロジスティック回帰に基づいているため，これらの理論や学習方法を復習した．また，ラベル列のマルコフ性を仮定した素性関数と，ビタビ・アルゴリズムや前向き・後ろ向き・アルゴリズムなどの動的計画法でラベル列の予測とパラメータの学習を効率化する方法を詳説した．

注．

1) https://taku910.github.io/crfpp/
2) http://www.chokkan.org/software/crfsuite/
3) https://wapiti.limsi.fr/
4) https://github.com/ExpResults/partial-crfsuite
5) https://github.com/WladimirSidorenko/CRFSuite

参 考 文 献

Berger, A., Pietra, S. D. and Pietra, V. D. (1996). A maximum entropy approach to natural language processing, *Computational Linguestics*, **22**(1), 39–71.

Duchi, J. and Singer, Y. (2009). Efficient online and batch learning using forward backward splitting, *Journal of Machine Learning Research*, **10**, 2899–2934.

Finkel, J. R., Kleeman, A. and Manning, C. D. (2008). Efficient, feature-based, conditional random field parsing, *Proceedings of ACL-08*: *HLT*, 959–967.

Hirohata, K., Okazaki, N., Ananiadou, S. and Ishizuka, M. (2008). Identifying sections in scientific abstracts using conditional random fields, *Proceedings of the Third International Joint Conference on Natural Language Processing* (*IJCNLP 2008*): *Volume-I*, 381–388.

Kudo, T., Yamamoto, K. and Matsumoto, Y. (2004). Applying conditional random fields to Japanese morphological analysis, *Proceedings of EMNLP 2004*, 230–237.

Lafferty, J. D., McCallum, A. and Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data, *Proceedings of the Eighteenth International Conference on Machine Learning* (*ICML 2001*), 282–289.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K. and Dyer, C. (2016). Neural architectures for named entity recognition, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies* (*NAACL-HLT 2016*), 260–270.

Liu, Y., Zhang, Y., Che, W., Liu, T. and Wu, F. (2014). Domain adaptation for CRF-based Chinese word segmentation using free annotations, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), 864–874.

McCallum, A. and Li, W. (2003). Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons, *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, 188–191.

Morency, L.-P., Quattoni, A. and Darrell, T. (2007). Latent-dynamic discriminative models for continuous gesture recognition, *Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR '07*), 1–8.

Sarawagi, S. and Cohen, W. W. (2005). Semi-markov conditional random fields for information extraction, *Advances in Neural Information Processing Systems 17* (*NIPS 2005*), 1185–1192.

Settles, B. and Craven, M. (2008). An analysis of active learning strategies for sequence labeling tasks,

Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, 1070–1079.

Sha, F. and Pereira, F. (2003). Shallow parsing with conditional random fields, *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 134–141.

Sun, X., Morency, L.-P., Okanohara, D., Tsuruoka, Y. and Tsujii, J. (2008). Modeling latent-dynamic in shallow parsing: A latent conditional model with improved inference, *Proceedings of the 22nd International Conference on Computational Linguistics* (*Coling 2008*), 841–848.

Tsuboi, Y., Kashima, H., Mori, S., Oda, H. and Matsumoto, Y. (2008). Training conditional random fields using incomplete annotations, *Proceedings of the 22nd International Conference on Computational Linguistics* (*Coling 2008*), 897–904.

Zhou, J. and Xu, W. (2015). End-to-end learning of semantic role labeling using recurrent neural networks, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (*ACL-IJCNLP 2015*) (*Volume 1*: *Long Papers*), 1127–1137.

200 Proceedings of the Institute of Statistical Mathematics Vol. 64, No. 2, 179–200 (2016)

Theory and Practice of Conditional Random Fields

Naoaki Okazaki

Graduate School of Information Sciences, Tohoku University

Most tasks of Natural Language Processing are formalized as a prediction problem of an output for a given input. Assuming that an input and output have a structure such as a sequence and tree, which is a natural assumption for a language, we can formalize more tasks as the prediction problem. This paper explains Conditional Random Fields (CRF) where an input and output are in the form of a sequence. In order to apply the multi-class logistic regression to the sequential labeling problem, CRF introduces feature functions that assume the Markov property for a label sequence and facilitates an efficient inference and parameter estimation by using dynamic programming. Therefore, this paper reviews the fundamental theories of logistic regression, feature functions, training with stochastic gradient descent, regularization, etc., and describes the overall theory of CRF. In addition, it covers recent research topics and practices including active learning for CRF, learning from partially-annotated supervision data, and models with deep neural networks.

Key words: Conditional Random Fields, logistic regression, stochastic gradient descent.

統計数理(2016)
第 64 巻 第 2 号 201–231

特集「統計的言語研究の現在」
[原著論文]

言語理解研究における眼球運動データ及び読み時間データの統計分析

新井 学[1]・Douglas Roland[2]

(受付 2016 年 4 月 4 日;改訂 8 月 20 日;採択 9 月 15 日)

要　　旨

言語理解に関する実験的研究は科学技術の進歩と共に過去 30 年ほどで飛躍的に前進した.以前には導入の困難だった眼球運動測定機もその低価格化と共に広く普及し,現在では世界の多くの研究室で眼球運動測定研究が行われている.しかし,このように量的データの収集が容易になった一方で,このような研究で得られるデータの量は機器の性能向上と共に増大しており,その分析方法は統計解析理論の前進,および様々な分析ツールの開発により複雑化している.そこで本稿では,言語理解研究における眼球運動測定実験,中でも視覚世界実験と読み実験によるデータ,そして自己ペース読み課題を用いた読み時間のデータに対して,現在広く利用されていて,かつ特別なリソースを必要としない分析方法を解説する.主に線形混合モデル及び一般化線形混合モデルを用いたデータ解析手法を中心に紹介し,これらのモデルを慎重に且つ論理的な手順をもって適用することは今までのデータの集約を必要とした分散分析などの手法と比べて多くの利点があることを説明する.

キーワード:線形混合モデル,一般化線形混合モデル,眼球運動,視覚世界パラダイム,自己ペース読み課題,読み時間.

1. 言語理解研究におけるデータ分析手法の背景

言語理解に関する実験的研究は急速な科学技術の進歩と共に過去約 30 年の間で飛躍的に前進した.特にパーソナルコンピューターを用いた反応時間計測によって,量的データを扱う時間計測(chronometric)研究が容易に実行できるようになり,実験的研究に対する敷居はかなり下がったと言える.更に以前には導入の難しかった眼球運動測定研究も機器の低価格化と共に広く普及し,現在では世界の多くの研究者・研究室で眼球運動測定研究が行われている.しかしこのように量的データの収集が容易になったのとは対照的に,そこで得られるデータ量は機器の性能向上と共に増大し,それらデータの分析手法は統計解析理論の前進と,様々な分析ツールの開発と共に困難さを増している.そこで本稿では言語理解研究における眼球運動データ,そして自己ペース読み課題を用いた読み時間のデータ分析について,現時点で用いられている方法をそのメリット・デメリットと共に紹介する.眼球運動データでは「視覚世界パラダイム」(Visual World Paradigm; 以下 VWP)と呼ばれる絵刺激上の注視を調査する方法と,モニターに提示した文を読む際の眼球運動を計測する方法に絞って解説する.本稿ではこれに加えて自己

[1] 成城大学 経済学部:〒 157–8511 東京都世田谷区成城 6–1–20;manabu-arai@seijo.ac.jp
[2] 東京大学大学院 総合文化研究科:〒 153–8902 東京都目黒区駒場 3–8–1;doug.roland@gmail.com

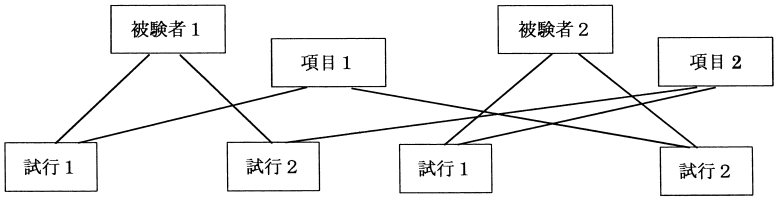


図 1. 複数の被験者と項目による階層的データの構造.

ペース読み課題による読み時間データを含め，各実験手法の特徴を考慮した異なるアプローチを検討する．その中でも主に，近年急速に普及している「線形混合モデル」(Linear Mixed-Effects Model,「階層線形モデル」(Hierarchical Linear Model)とも呼ばれる)を利用した分析方法を中心に説明する．

分散分析を代表とする従来の分析方法では，複数の試行から得られたデータを被験者ごと，もしくは項目ごとに平均値を計算するデータ集約(aggregation)を行った後で，特定の説明変数の有意性を判定するために t 検定などの検定を行っていた．このデータ集約を必要とする理由はデータに階層性が存在するからである．本稿で対象とする研究では通常各被験者が複数の試行に参加し，1 人の被験者より複数の(通常項目の数だけ)データが得られる．つまり図 1 のように，被験者 1 から得られる各試行データは，被験者 2 の各試行データとは独立していて，入れ子(ネストされた)構造になっている．つまり 2 段階のサンプリングによるデータ構造を持っていることになる．さらに，通常複数の項目を用意して，各項目に対して複数の(被験者の数だけ)データが得られるため，項目についても，被験者とはまた別の，入れ子になったデータ構造を考慮しなければならない．

この構造によって，各被験者(または各項目)内のデータは，他の被験者(項目)のデータよりも類似する．従来の方法では，このいわゆる「集団内類似性」を考慮に入れることができないために，データ集約を必要としていた(詳しくは清水, 2014 を参照)．そして被験者と項目，つまり調査された人間および文の二つの要因によってデータの類似性が生まれるという事実を考慮するため，被験者ごとの平均値と，刺激文ごとの平均値に基づく 2 つの分析，いわゆる F1 と F2 分析，が行われてきた．しかし，この方法では片方の分析だけで有意差が見られた場合など解釈に困るため，Clark (1973)によって F1 と F2 の結果から被験者・項目両方に一般化できる効果かどうか判定する $minF'$ を計算し判定を行うことが提案された．しかし，この方法は保守的である等の指摘もあり，現実には F1 と F2 の結果報告だけに留まる，または $minF'$ の報告も併記するが無視されるケースが多く，その結果，被験者・項目両方に一般化可能な効果の検定という意味では課題が残ったままであった(Raaijmakers et al., 1999)．

それとは対照的に，線形混合モデルではデータの階層構造をそのままモデル化することができるため，データの集約を必要としない．つまり各試行のデータ(図 1 における最下位レベルデータ)をそのまま従属変数として分析することができる．さらには被験者と項目という 2 つのランダム効果を同時にモデル化できるため，一つのモデルによってデータの集約なしに言語研究特有のデータ分析における問題を解決することができる．さらには，眼球運動測定実験のように各試行から複数のデータが得られるような(つまり図 1 の最下位レベルの各試行の下にさらに「データ 1」「データ 2」のようなレベルが存在する)，さらに多くのレベルが存在する階層的データもモデル化することが可能である．このように線形混合モデルではデータを失うことなくデータ構造に合ったモデル構築が可能である．

線形混合モデルはその名の通り線形モデル(Linear Model)を拡張したものである．つまり，最

も単純な回帰分析の式($y = ax + b + e$)，つまり y という変数の値を，直線の切片(b)と変数 x の傾き(a)と実測値とのズレ(残差またはエラー e)の組み合わせから予測する式の x にあたる部分を，任意の数の説明要因(固定効果)とランダム効果とに分け，それらを階層的に同時に混ぜる(混合)ことができるようにしたものが線形混合モデルである(詳しくは Baayen, 2008 等を参照)．従来の t 検定や分散分析も分類上は最小二乗法(Ordinary Least Squares)を用いた線形モデルに属する．そのため，これらすべての分析は「線形」で y の値を回帰している以上，共通して残差(e)が正規分布(Normal または Gaussian distribution)に従うことを前提としている．

線形混合モデルが従来の手法と大きく異なるのは，最小二乗法が適用できず，自分で実際のデータに対して最もあてはまりのよい(尤度の高い)モデルを探索し選択する点である．特に，ランダム効果(実験を行う前にその効果を予測することができない要因．固定効果は逆に，事前に効果が予測される要因)に対しても柔軟なパラメータの設定が可能であり，非常に詳細なモデル構築が可能である．たとえば読み時間における個人差(実験における説明変数とは関係のない読みのスピードの差)をランダム効果(ランダム切片と呼ばれる)として指定し，さらに説明変数の効果の大きさにおける被験者間の差を追加要因(ランダムスロープと呼ばれる)として指定することなどが可能である．これはつまり，それぞれの実験のデザインによってどのような効果(固定効果とランダム効果両方)が起こりうるか考慮し，実際のデータ構造をできる限り適切にモデル化する必要があることを意味する．

本稿では，それぞれの種類のデータの特性を考慮し，現時点で妥当だと思われる方法について具体的に説明する．注意する点として本稿で紹介している分析手法は必ずしも最も優れている方法ではなく，あくまで数ある有効な方法の一つとして，メリット・デメリットを含めて紹介している．現実には個別のケースに合わせて最も適切だと考えられる方法を各自採用していただきたい．また本稿では自分のデータに応用する際の手助けとなるよう，フリーの統計解析ソフトである R のコードを適宜示している．

2. 眼球運動測定研究

2.1 眼球運動の基礎情報

人が文を読む時，または静的な視覚的刺激(絵や写真)を処理する時の眼球運動は，スムーズに平面上を移動しているのではなく，一つの場所での停留(fixation)と急速なスピードで別の場所へ移動するジャンプ(フランス語でジャンプを意味するサッカードと呼ばれる)を繰り返している．サッカードは典型的に 20–35 ms ほどで非常に短くこの移動中には視覚的な情報はほとんど得られないことがわかっている．そして停留は平均で 200–250 ms ほどだが，その時間はその時処理している情報によって 150 ms から 500 ms ほどまで変動する．つまり，停留がいつ，どこで，どのくらいの長さで起こったのか調べることで，実時間に沿った文理解における認知処理を調査することができる(Rayner and Pollatsek, 1989)[1)]．眼球運動は一般的な読みにおける文理解の処理を理解するのに最も優れた方法であり(Rayner, 1998)，近年では読みだけでなく，視覚世界パラダイムを用いて音声言語の理解の処理を理解するのにも眼球運動測定が利用されており，その重要性は益々高まってきていると言える．次のセクションでは読みと視覚世界パラダイムの二つの手法による眼球運動データの分析方法を紹介する．

2.2 眼球運動測定・視覚世界パラダイムによる注視データ(カテゴリー変数)分析

2.2.1 実験デザインとデータの構造

具体的な VWP を用いた実験デザインとして，Kamide et al. (2003)に似た日本語による実験デザインを仮に想定する．この仮想実験での被験者は図 2(a)か図 2(b)のどちらかの絵を見なが

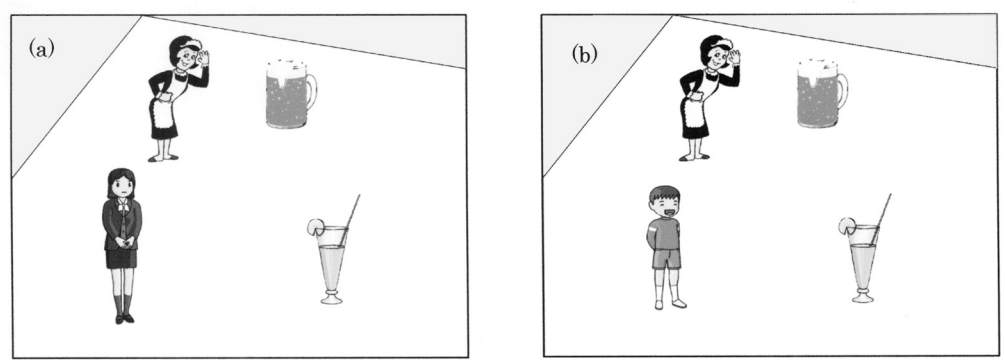


図 2. VWP 実験の刺激絵例.

ら，音声刺激を聞いて理解する．

音声刺激は以下の(1a)または(1b)のどちらかの条件の実験文が再生される．

(1a)非制限条件：喫茶点でウェイトレスが OL につめたく冷えたビールを運んだ．
(1b)制限条件：喫茶点でウェイトレスが子供につめたく冷えたジュースを運んだ．

刺激絵にはビールとジュースの絵が含まれていて，与格の「〜に」に当たる情報を聞いた時点で，被験者はウェイトレスが運ぶ対象物となる直接目的語を予測すると仮定する．その際，その予測が語彙情報及びその語彙に基づく現実世界の知識(普通 OL はビールもジュースも飲めるが，子供はジュースしか飲めない)を即座に反映するのであれば，(1a)の条件では次に来る情報に対して制限がないが(非制限条件)，(1b)の条件では制限があるため(制限条件)，後者の方がよりジュースに注視が集まるはずだという仮説が成立する．この場合，興味対象となる対象物がビールとジュースというように複数あるので，単純に特定の興味対象への注視時間を従属変数とすることができない．そのため分析では「特定の対象物を他の対象物に比べてどれくらい見ていたか」という割合を計算する[2)]．そしてここでの興味の対象は，試行の開始から終わりまで全ての眼球運動ではなく，認識された言語情報の処理に対する反応として起こる注視であるので，まず各音声刺激でその予測的眼球運動のキューとなる言語情報のオンセット時間をマークし眼球運動データとリンクさせる必要がある．この仮想実験ではそのキューになるのは与格名詞句であるので，この語句のオンセット(もしくは助詞「に」のオンセット)の時間をチェックして(実際には眼球運動が言語情報を反映するまでにかかる時間(およそ 180〜200 ms)を加える(Matin et al., 1993))，その時間から，「ビール」か「ジュース」のオンセットまでの時間の幅を分析対象とする「時間枠」として設定し，この枠の中でジュースに向けられた注視とそれ以外への注視の比率を計算し，分析を行う．修飾句「つめたく冷えた」を含める理由としては分析対象の時間枠が短くなりすぎると有効な注視が観測されにくくなるためである(後述するが時間枠の幅が狭ければ狭いほどデータは二項変数分布に近似する)．

元々眼球運動測定器によって記録された眼球運動データは各サンプリングデータが記録された時間情報と注視のあった画面上の位置を示す座標軸情報で構成されている．これを絵刺激上の対象物ごとに区切ったテンプレートと照合する事で，表 1 が示すように座標軸情報からどの対象物を見ていたかを示すカテゴリーデータ('AOI' の変数)が得られる．これによって特定の対象物への注視は，その対象物に注視があったか否か，という二項変数として扱うことができる．今までは，この二項変数データに基づく割合，つまり被験者ごとに試行数全体の内どの位ある対象物を見たか(e.g., Tanenhaus et al., 1995)，または各試行で得られた複数のサンプリン

表 1. 眼球測定器によって出力されるサンプリングデータの例.

subject	item	factor1	factor2	Record Time	X-axis	Y-axis	AOI1	AOI2	AOI3	background
1	1	1	2	255787	993	528	0	1	0	0
1	1	1	2	255790	993	528	0	1	0	0
1	1	1	2	255793	993	528	0	1	0	0
1	1	1	2	255797	993	528	0	1	0	0
1	1	1	2	255800	993	528	0	1	0	0

グデータの内どのくらいターゲットを見ていたか，という割合を求めることで 0 から 1 の間の数値を取る値に変換し，その値を連続変数として分散分析などの検定テストに適用している例が多く見られたが，この方法には明らかな問題がある．Jaeger (2008)に詳しいように，連続変数を分散分析などのパラメトリック手法で分析するには，その変数は母集団が正規分布に従い境界値を持たないことが仮定されるが，割合を計算した場合，0 から 1 の間の数値しか得られない．また正規分布を前提とした線形回帰モデルなどの分析を適用した場合，その推定値と実際の値とのズレ(残差)は平均値とは独立してランダムに起きるはずであり，エラーの分布は正規分布に従わなければならないが(つまり線形混合モデルの式における各試行レベルでのエラー(e)が $e \sim N(0, \sigma^2)$ に従う)，割合を計算した場合にはそのエラーの分散は平均値に依存し，割合が 0.5 をピークとして分散が最も高くなる．つまりデータサンプルの割合の平均値が 0.5 から離れれば離れるほど分散は低く見積もられる(結果として標準誤差も低く見積もられ，第 1 種の過誤(タイプ 1 エラー)が起こりやすくなる)．Jaeger (2008)はこの原因によって割合のスケールで行った分析では実際には主効果しか存在しないデータにおいて誤った交互作用が検出されることがあることを報告している．

2.2.2 対数オッズに対する線形混合モデルと一般化線型モデル

割合をそのまま従属変数として使うことができない問題に対して広く知られている解決方法はオッズを計算し対数変換することである(Agresti, 2002; Barr, 2008)．オッズは，ある出来事が起きた回数(たとえば上の VWP の例で一定時間内でビールへの注視が記録された回数)の，そのイベントの起こらなかった回数(ビールへの注視が記録されなかった回数)に対する比率であり，説明変数の乗法的な効果を説明するのに適している．これによって VWP では一定時間内で別の場所を見ていた注視に比べてどのくらいターゲット対象物を見ていたかを表すことができる．そして，そのオッズを対数変換したロジット(logit)とよばれる値を，正と負の境界を持たない，回帰分析上都合のよい加法的な効果で説明される従属変数として線形混合モデルに加える．このロジット値を従属変数として線形混合モデルを適用することは，二項変数に対してロジットリンク関数を用いて行う一般化線形混合モデル(混合ロジスティック回帰)を用いるのと概念上は同義になる．割合からロジットへは以下のように変換できる．

$$\eta = \ln\left(\frac{\phi}{1-\phi}\right)$$

注意すべき点として眼球運動データは機器の視線探知ロスであったり被験者のまばたきがあったりと欠損値が少なからず起こる．そのため，全体の試行数のうち何回ターゲットを見ていたかという割合を計算すると，これら欠損値は「ターゲットを見ていなかった」試行としてカウントされてしまうが，実際には何を見ていたのか不明であるため分析に含まれるべきではない．そのため，特定の時間枠内におけるロジットの計算では，以下の式で分母の n は背景を含め画面上に注視が記録されたデータポイントの合計で，y はターゲットに向けられた注視があったデータポイントの合計に当たる．実際には 0 の対数は定義されていないため，分母分子両方に

表 2. 試行ごとに AOI1 の経験ロジット(‘logit’ の変数)を計算したサンプルデータ.

subject	item	factor1	factor2	AOI1	AOI2	AOI3	background	sum	logit
1	1	1	1	87	0	284	0	371	-1.18
1	2	1	2	157	253	80	0	490	-0.75
1	3	2	1	496	0	0	0	496	6.90
1	4	2	2	120	623	0	0	743	-1.64
1	5	1	1	505	0	249	0	754	0.71
1	6	1	2	0	0	492	12	504	-6.92

0.5(ゼロではない最小値の半分)を足してロジットの計算を行う経験ロジット(empirical logit)(η')が計算される(McCullagh and Nelder, 1989)[3].

$$\eta' = \ln\left(\frac{y+0.5}{n-y+0.5}\right)$$

n に対するもう一つの考え方としては分析対象を刺激絵の中の興味対象の対象物 2 つに絞るという方法も考えられる.つまり興味の対象として A と B という二つの対象物(先の例ではビールとジュース)のうち割合としてどちらの方をより多く見たかといった問題に答えるため,対象物 A への注視を y として,n は対象物 A と対象物 B 両方への注視の合計を取ることができる.そして,A と B への注視量の比率を対数変換し(上の式で y を対象物 A への注視の合計,$n-y$ を対象物 B への注視の合計とする),経験ログ比(empirical log-ratio)を計算することができる(e.g., Arai et al., 2007).この場合,A と B 以外の対象物を注視していた時のデータは考慮されないので,それら対象物への注視量が条件間で違いがなかったか確認する必要がある.

上記のように各試行において経験ロジットまたは経験ログ比の値を計算することで,カテゴリー変数が適切な連続変数へと変換され,表 2 が示すように,各行が各試行に該当するように変換される(つまり図 1 の階層的データと同じ構造となる).そしてこのロジット(logit)または log-ratio を従属変数として適切な固定効果とランダム効果をモデル化し分析を行う.以下の R コードはある眼球運動データ(dat)に含まれるロジットに対して説明変数(X, Z)とランダム効果(subject, item)を含めた線形混合モデルの例を表している.ロジットの分散は平均値に依存しているので以下のようにロジットの値に重みを加えた線形混合モデルを適用することが勧められている(Barr, 2008).

```
#重み (wts) の計算.AOI1 をターゲット対象物として,各試行の時間枠内で AOI1 への注視があったデータポイントの合計,sum は背景を含む絵刺激で記録された注視のあったデータポイントの合計
dat$wts<−1/(dat$AOI1+0.5)+1/(dat$sum−dat$AOI1+0.5)
#重み付けした経験ロジットに対する線形混合モデルの R コード
library(lme4)　#lmer 関数に必要な lme4 パッケージを呼び込む
m0<−lmer(logit∼X*Z+(1+X*Z|subject)+(1+X*Z|item), weights=1/wts, data=dat)
summary(m0)　#結果の表示
```

この式では,logit が従属変数,X と Z が固定因子,括弧内の subject と item がそれぞれ被験者と項目のランダム効果に対応する.X と Z の間の*は乗算を意味し,両者の間に主効果のみではなく交互作用が含まれていることを意味する.括弧内の 1 が切片を意味し(つまり(1|subject)であればランダム切片のみの指定となる),パイプ(|)の前に指定された X*Z はランダムスロープを意味し,*でつながれているため,3 つのランダムスロープ(X と Z と X:Z)が被験者に対し

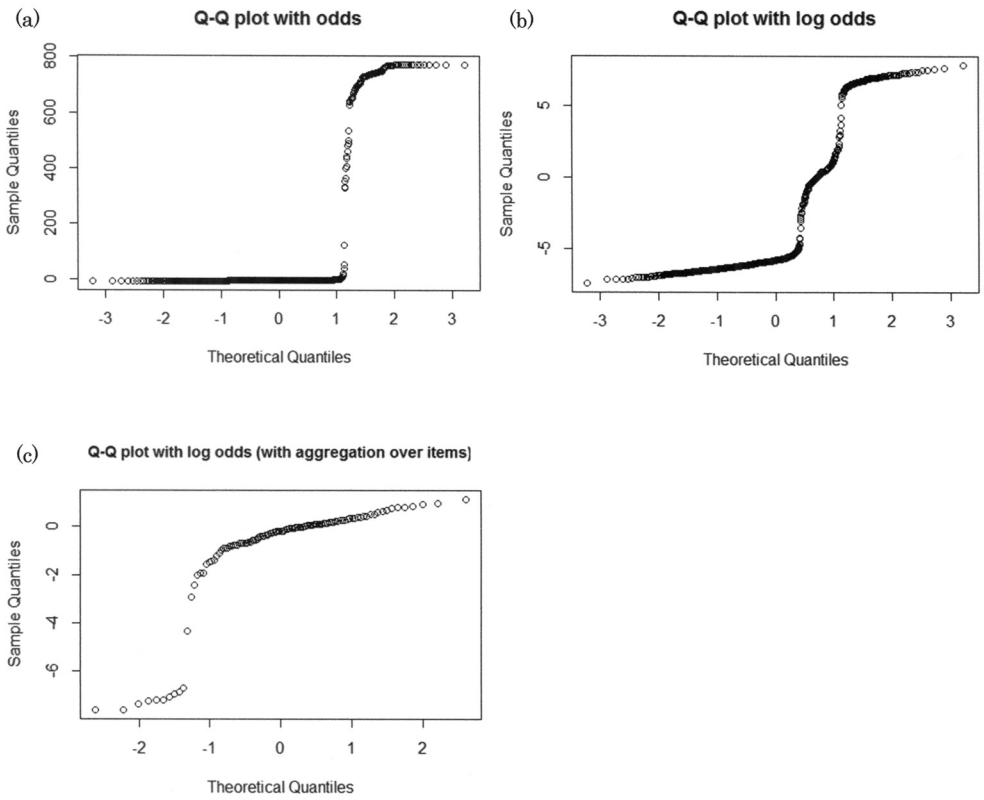


図 3. VWP 実験の眼球運動データから計算されたオッズ(a)，ロジット(b)，そして被験者ごとにデータ集約して計算したロジット(c)に対する線形混合モデルの残差の Normal Q-Q Plot.

て指定されていることを意味する．このようにして元々カテゴリー変数だった眼球運動データに対して線形混合モデルを用いて分析することができるが，このモデルが実際のデータに対してあてはまりが良いか否かは別の問題である．図 3(a)，(b)は Arai et al. (2015)で報告されている眼球運動データ(実験 1)の実際の値から計算されたオッズと対数変換されたオッズ(＝ロジット)に対して線形混合モデルを適用し，その残差が正規分布に従うか否かチェックするために Q-Q プロット(Normal Quantile-Quantile Plot)を描画している．このグラフは，正規分布の分位数(Normal Quantile)と，実際のデータの確率分布の分位数を比較し，視覚的に正規分布からのズレを確認することができる．データがもし正規分布に完全に従う場合右斜め 45° の直線上にデータが乗る．グラフからわかるようにロジットに線形混合モデルを適用した場合，単なるオッズに適用した時に比べて残差はより正規分布に近づくが，正規分布に従っているとは言い難い(実際にシャピロ・ウィルク(Shapiro-Wilk)検定を行うと，有意な差が検定される)．このように両グラフから分散の不均一性(heteroscedasticity)を観測することができる．

```
qqnorm(resid(m0))
shapiro.test(resid(m0))
```

Barr (2008)は，各時間枠内の各眼球運動データサンプルの間には依存性が存在するので，これらをまとめて分析すると個々の観測データにおける独立性の前提に反すると指摘している．つまり，ある時点で対象物 A に対して注視が記録され，次にデータが記録される時点(1000 Hz で記録されていたら 1000 分の 1 秒後)で別の対象物に注視が記録されることは事実上不可能である．この各サンプリングデータ間の依存関係を無視して分析を行うことによって正しい標準誤差が計算されず第 1 種の過誤の危険性が増すと批判している．Barr は，この問題を回避するために，項目要因のレベルを崩して被験者ごとの総数，また被験者要因のレベルを崩して項目ごとの総数からロジットを計算し別々の線形混合モデルを適用することを勧めている．しかし，この方法にはデータを集約することによって情報と検定力が失われるというデメリットもある．更に，先と同じデータで項目要因を崩したロジットに対する線形混合モデルにおいても，図 3(c)で示されている通り，その残差は正規分布には従わず(シャピロ・ウィルク検定においても有意)，線形モデルの適切さに問題が残る．データの集約を行わない計算方法では刺激絵上に注視のなかった(計測ロスまたはサッカードによる)データポイントをカウントしないことで，この問題を回避している(Barr, 2008 の方法では n はデータの総数であるためこれらもカウントしている)．しかし，この方法でもある時点で対象物 A を見ていたら，次のサンプルの時点においても同じ対象物 A を見ている確率は高くなるので完全に独立しているとは言えず，依存性の問題は解決されない．このように線形混合モデルのあてはまりが適切であるとは言い切れないケースでは，分散分析など他のアプローチによる結果と比較し，場合によっては両方の結果を報告することが有益だと考える(Roland, 2009)．

一つの有効な手段として，もしオッズの分布が 0 か 1 に集中している場合には，時間枠内のデータをまとめて二項変数化して分析することもできる(Kamide, 2012)．つまりその時間枠でターゲットに対して一つ以上の注視が観測された場合には 1，観測されなかった場合には 0 と変換することによって，混合ロジスティック回帰を用いて分析する(Jaeger, 2008)．このモデルは様々な分布を持つデータを扱える一般化混合線形モデルの一つで，分布ファミリーを二項変数，そしてリンク関数にロジットを用いたモデルである．この方法のデメリットは，オッズにおいて最頻値が 0.5 にくるような分布をもつデータにおいては多くのデータが変換されるため，多くの情報が失われてしまう点である．

```
#混合ロジスティック回帰の R コード
m0<−glmer(binary∼X*Z+(1+X*Z|subject)+(1+X*Z|item), family="binomial", data=dat)
summary(m0)
```

2.2.3 モデルにおける固定効果とランダム効果の指定

線形混合モデル及び混合ロジスティックモデルに含める各固定効果は平均値 0，標準偏差 0.5 を取るよう中心化を行う(2 水準で水準間のデータ数が同数の場合それぞれの水準が -0.5 と $+0.5$ となる)．これによって，要因間の共線性(collinearity)を最小限に抑えることができる．さらに，すべての説明変数を中心化した場合，切片は全体平均に相当し，説明変数の係数に対する検定テストは分散分析における主効果に対する検定テストに対応するため，回帰係数の解釈が容易になる．

```
#中心化(X は実数データである必要がある)
scale(d$X, scale=F)　#scale=T とすると中心化＋標準化(平均値 0, 標準偏差 1)
#又は単純に以下のように書いても同じ(標準化する場合には結果を標準偏差で割る)
d$X−mean(d$X)
```

上のモデルではかっこ内に subject と item が含まれ，被験者間の個人差，つまり各被験者がどの程度刺激絵内の各対象物を見たかにおける被験者間の差，また刺激絵の認知的顕著性の差などによって起こる項目間における注視量の差をランダム効果として説明している．これに加えて，個々のランダム効果における説明変数の効果の差をランダムスロープとして指定している．これによって，各説明効果がそれぞれのランダム要因の各レベルで異なる値をとることができる．

このように線形混合モデルでは平均値を求めるためのデータ集約を必要とせず個々の試行の各データサンプルをそのまま扱えるので(Barr, 2008 の方法を除く)，各実験の流れの中で説明変数の効果がどのように変化したか，また隣接するトライアルが説明変数の効果にどう影響を与えたかなど，今まで見過ごされていたかもしれない共変数(covariate)の影響を調査できるメリットがある．実際に実験前には想定していなかったが影響の強かった効果を共変数を加えることで(たとえば年齢)，元々興味のある効果の有意差がなくなるという可能性もある(Baayen, 2008 の語彙判断課題の反応時間の例を参照)．このようにして，線形混合モデルを用いることで，観測された効果が実験操作の影響ではなく，連動する他の要因によって観測された効果であるという擬似相関の可能性をテストすることができる．これは，従来の被験者ごと，あるいは項目ごとにデータを集約する方法では確かめることができないため，データ集約を行わないことのメリットは大きい．

先に述べた仮想実験の例では，語彙的制限(非制限 vs. 制限)を説明効果に，そして被験者，アイテムをランダム要因，そして語彙的制限のスロープを各ランダム変数に含めたモデルを作り分析を行う．説明変数のスロープがモデルに含まれない場合，効果は全被験者に同じ大きさで起きていることを仮定することを意味する．そのため，もし現実には被験者グループ内の一部の人だけに大きな効果が観測され他の被験者では全く効果がなかった場合でも，全被験者の平均を取ることで有意差がみられるケースもあるので注意が必要である．

この VWP の仮想実験の例では，1 要因 2 水準を想定しているが，実際の実験では制限・非制限の操作とは関係なく，OL とビール，子供とジュースという意味的な結びつきによって条件間に差が現れる可能性があるため，「喫茶点でウェイトレスが OL をつめたく冷えたビールでもてなした」と「喫茶点でウェイトレスが子供をつめたく冷えたビールでもてなした」という条件を加えて 2×2 デザインを組むことが好ましい．この場合，「OL に/を」の条件間の差に比べて「子供に/を」の条件間には差の方が大きいという交互作用の予測が成り立つ．そのため分析モデルには，2 要因(意味的制限の有無と助詞の種類)とその交互作用が説明効果として加えられ，同時にランダム効果としてもこの 3 つの傾きが被験者，アイテムのランダム効果に指定される．

実際にデータを分析する際には，モデルを当てはめる前にこのような不規則なデータパターンが起きていないか調べることが大切である．これによって，データ整形する過程で何か間違いを犯さなかったか，もしくはなにかしらの理由で眼球データがうまく記録できなかった被験者がいなかったかなど事前にチェックすることができる．R 上で被験者ごとのデータを描写するには lattice パッケージの xyplot 関数を使うことで，カテゴリー変数でも連続変数でも被験者ごとのデータパターンを確認することができる．図 4 は 2×2 デザインのある実験の読み時間データを加工したもの(RT)を条件別(X, Z)，被験者別(subject)に R 上で xyplot 関数を用いて描画したものである．

```
xyplot(RT~X|subject, group=Z, data=dat, col=(c("black","darkgray")), type=c("p", "r"),
    xlab="Factor1", ylab="Reading Time (ms)")
```

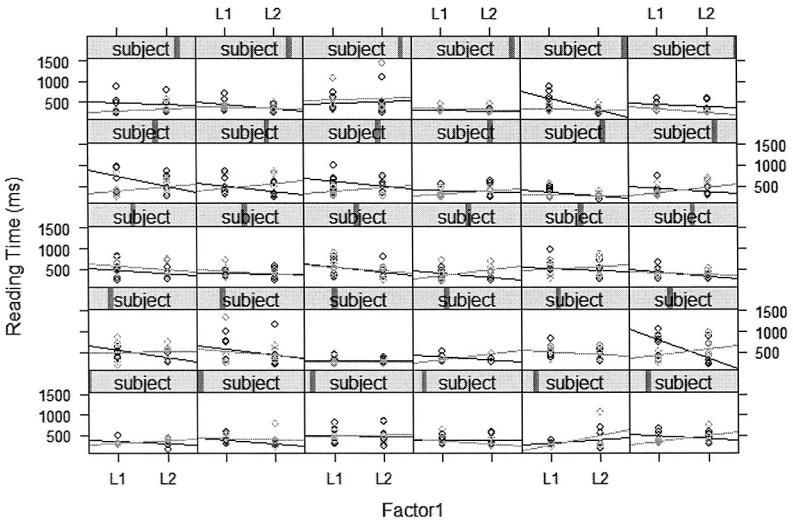


図 4．lattice パッケージの xyplot 関数による被験者ごとの読み時間における交互作用プロット．

2.2.4 経過時間情報のモデル化

上記の方法の一つの問題としては「時間枠」の選択(time-locking と呼ばれる)が実験者にゆだねられていることが挙げられる．場合によっては Altmann and Kamide (1999)のようにどの言語情報が(彼らの場合は動詞)特定の対象物への注視を引き起こすか明白な場合もあるが，そのような明白なキューが存在しない場合(もしくは複数考えられる場合)もある．そのような場合，実験者はグラフを見ていかにも条件の差がありそうな時間枠を恣意的に選択したくなるかもしれないがこの方法には客観性の問題があり，もし説明変数の効果が試行内のどの時間枠であっても平等に起こりうる可能性がある場合には第 1 種の過誤の確率を高める(無作為抽出の前提にも反する)．

また，時間枠を設定してそこに含まれる眼球データをまとめて扱った場合，たとえ分析結果では水準間の有意差が観測されず，平均値がほぼ同じであっても，それぞれの水準で異なるパターンが起きている可能性がある(たとえば一方の水準では時間枠内で注視量が上昇し，もう一方の水準では下降していて時間枠の中央で交差している場合)．そのため Barr (2008)は時間枠の開始時点における差を予測(anticipatory)効果，時間枠内の変動を進度(rate)効果として分けて考えるべきだと主張している．そのため上記の分析を行うには時間枠の開始点とその枠内でそのような変動がないことを少なくとも時系列に注視データをプロットしたグラフ上で確認した上で行わなければならない．このような効果がありそうな場合には選択された時間枠内を複数の区切りに分け，その区切りを時間軸に沿った連続変数(たとえば 500 ms からなる時間枠を 50 ms ごとに区切り 1 から 10 という連続数を当てる)として扱い，モデルに追加することができる．その際，時間経過における変化をモデル化するが，一定の進度で注視が増加，もしくは減少していてグラフ上のデータがほぼ直線になっている場合には線形(linear)として扱う．しかしデータが曲線になっていて，曲がりが一つある場合には 2 次項(quadratic term)，2 つある場合には 3 次項(cubic term)を用いることができる．このように時間経過上のデータ形状に合わせた多項式表現をモデルに加えることが可能である．以下の R コードでは線形に加えて 2 次項を追加している(*は乗算，^ は累乗を表す)．

lmer(logit∼(time+I(time^2))*X+(1+time+I(time^2)|subject), data=dat)

こうした分析は Growth Curve Analysis と呼ばれ，眼球運動のように，条件ごと，また被験者ごとに異なるカーブを描くデータをモデル化することができる（Mirman et al., 2008）．この分析では比較的小さな各時間区切り（time bin）におけるデータサンプルの数が少ない事と，サンプル間の依存性を回避するため，通常先ほどの Barr（2008）の方法と同じく項目要因のレベルは崩され被験者ごとの経験ロジットが通常計算される．この分析の難しさとして，経過時間に対する多項式表現はどこを中心（0）と設定するかによってカーブの形が大きく変わり説明変数の係数に大きな影響を及ぼす点が挙げられる．

また最近では，Permutation test と呼ばれる，時間枠の選択を完全に客観的に行う分析方法も提案されている（Maris, 2012）．この分析では 20 ms などの小さな時間の区切りをまず設定し，各区切り内で検定テスト（t 検定）を実行する．そして有意差の得られた連続する区切りをクラスターとして結合する．今度はそのように形成された各クラスター内で説明変数の水準の順列を無作為にシャッフルし検定テストを実行することを一定回（〜10000 回）繰り返す．そこから得られた t 値の分布を確率分布として利用して，実際のデータから得られた検定値（t 値）が，その分布上で 5%を切る確率で起こるのであれば，そのクラスターにおいて説明要因の効果が有意であると判断する．この手法のメリットは，同じデータに対して複数の時間枠を繰り返し分析する多重比較の問題を回避できることである．

2.3 眼球運動測定による読み時間データ（連続変数）分析

眼球運動測定を用いた実験手法として上で説明した VWP による方法とは別に，文そのものをモニター上に提示し，それを読んでいる間の眼球運動を測定する方法がある．読みにおける調査方法として，後述する自己ペース読み課題と比較して，眼球運動測定の大きなメリットは，実験参加者は提示された文をただ読むだけでよく，その間指示に従って特別な反応をする必要がなく，最も自然に近い形で文を読む際のデータが得られることである．これによって特定の実験手法に対する被験者の戦術的な反応の影響を最小限に抑えられると考えられる．

得られたデータは，刺激文を興味対象にしたがって単語や句などの複数のリージョンに分割し，特定のリージョンにおける停留時間や前のリージョンへの読み返し率など，様々な指標を計算し従属変数として分析される（各メジャーの詳しい説明は Rayner, 1998 を参照）．そのため，一つの試行から（つまり一つの文を一人の被験者が読んだデータから）実に多くの分析対象となる値が計算されるわけであるが，個々の分析においては，各試行に対して一つの値が計算されるため，図 1 と同じ二つのランダム効果（被験者と項目）に対して各試行が入れ子構造になっている 2 段階の階層的データとして扱われる[4]．

たとえば，first pass 読み時間と呼ばれる読み時間は，特定のリージョン内で初めて記録された停留からそのリージョンから抜け出るまでの停留時間の合計を計算した指標であり，語彙の情報の処理において即座に起こる処理を反映していると考えられている．図 5 はある眼球運動測定実験から得られた特定のリージョンにおける first pass 読み時間の分布を表している．見て取れるように分布は典型的に右に裾が長い左右非対称な分布を持つ．これによって後述する通り変換処理の是非が議論されているが，大多数の報告ではそのままの読み時間を分析対象としている．図 5 からもわかる通り，このデータには 1400 ms 近い外れ値（実際は 1360 ms）が存在する（このサンプルでは被験者平均が 283 ms，標準偏差が 68 なので，この値は標準偏差 15 倍以上離れていることになる）．

問題は，この値が分布を反映していない外れ値として始めから分析から除くべきであるのか，あるいは裾が右に長い特定の分布形状を構成する一要員として考えて除くべきではないのか判

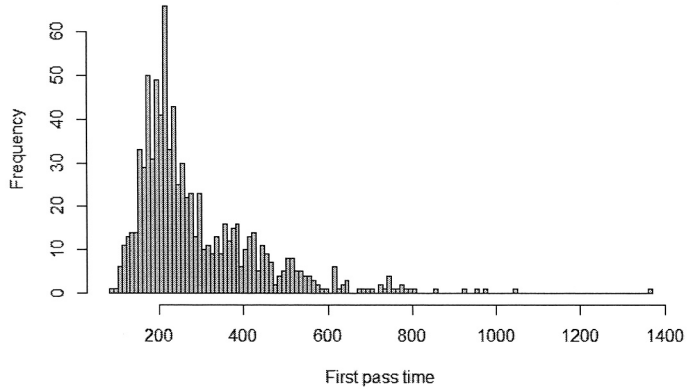


図 5. First pass 読み時間の分布の例.

別が難しい点である．前者だと想定した場合には，モデルを当てはめる前にそのデータの被験者平均から標準偏差の 3(または 2.5)倍を超える値を先に除外する(e.g., Sturt, 2007)．そして，外れ値を除外した残りのデータを従属変数として線形混合モデルを当てはめ分析を行う．しかし後者を想定した場合，事前にデータの除外を行わず先にモデルを当てはめ，その後でそのモデルに基づく推定値と実際の値との差(残差)の分布から標準偏差 3(または 2.5)を超える値を排除した上でもう一度同じモデルを当てはめ，その結果を報告する(Baayen, 2008; Baayen and Milin, 2010)．

```
#モデル(m0)の残差の絶対値が 3 を下回るデータのみを選択する
newdat<−data[abs(scale(resid(m0)))<3,]
```

後者の方法では，前者の方法と違って，外れ値をモデルを適用する時点で無かった事にしていない，つまり，それら外れ値も特定の分布からランダムにサンプルされたデータの一員であると見なしモデルの当てはめを行っていることを意味する．しかし，それら極端な値は説明要因の効果を計る上では過度な影響力を持つことが考えられるので，最終的なモデルにおいては除外して検定を行っている．実際にそれらの外れ値が全く別のことを考えていたなどのエラーによって起因していると確信できる場合を除いては(これは実際には難しく，実験中にアラームが鳴ってしまったような明白に問題のある試行についてはその試行から得られたデータ全てを除外すべきである)，提示されている言語情報に関係した認知的処理を反映していると考えられるため，後者の方がより望ましいと考えられる．実際に，このサンプルデータにおける最大値である 1360 ms を除いたモデルと，全てのデータポイントを含んだモデルそれぞれに対して，モデルのあてはまりの良さ(goodness of fit)を調べる R^2 を計算すると，除かなかったモデルの方があてはまりがよかったことが確認できた(つまり，この最大値は分布の一部を構成していると見なすべきである)．R^2 は以下のようにモデルからの推定値と実際の測定値との相関の 2 乗で求めることができ，値が高いほどあてはまりがよいことを示す(Baayen and Milin, 2010)．

```
cor(fitted(m0), dat$RT)^2
```

すでにみたように，読み時間における説明要因の影響を分析する上で，特定のリージョン内の言語情報特有の影響は，項目をランダム効果として含めることで説明が可能である．しかし，それとは別に，単純にそのリージョンにおける文字数(アルファベットないし漢字・かな等)の

影響が考えられる．過去の日本語における研究では，文字数よりもモーラ数の影響が強く，さらに単語親密度などの影響も存在することが知られている（Mazuka et al., 2002）．このことから，読み時間の分析にはこれらを共変数としてモデルに含めることが考えられるが，眼球運動データの場合，この問題はそれほど単純ではない．というのも，認識領域（perceptual span）と呼ばれる言語情報を認識することのできる領域というのはアルファベットで，個人差はあれど左側 3–4 文字，右側 14–15 文字程度（左から右へ読む文字の場合）だといわれているので，リージョンの区切りを超えている場合が多々ある．このような場合，文字数の影響は分析対象としているリージョンの前後に現れる単語の文字数（ないしモーラ数等）も考慮すべきだと考えられる．さらには，読み返しを含む Regression path 読み時間の場合，そのリージョン以前のすべてのリージョンの文字数も影響しうると考えられる．このような理由から，眼球運動データの分析では一般的にリージョン内の文字数をモデルに含めることは行われず（同じ理由で残差読み時間の計算も行われない），その代わりに条件間で文字数ないしモーラ数等をできる限り統制して実験文を用意する必要がある．もし特定のリージョンにおいて条件間で文字数等の違いがあり，さらにそこで固定効果に有意な差が検定された場合には，その効果が実際には文字数等の違いによって引き起こされた可能性があり注意が必要である．その場合，モデルに文字数等を共変数として追加しても説明変数の効果が有意なままであるかどうか確認することが必要となる．

読みの眼球運動データにおけるもう一つの問題は，図 5 で見た通り各指標のデータの分布が通常正規分布に従っていない点である．この場合，正規分布を仮定した線形混合モデルをそのままの読み時間データに適用することは厳密には適切ではない．この問題を回避するためさまざまな方法が試みられているが，この問題は次に扱う自己ペース読み課題による読み時間データにおいても共通しているので，次のセクションでまとめて扱うことにする．

3. 自己ペース読み課題による読み時間（連続変数）分析

自己ペース読み課題（Self-paced reading task）とは文を単語または句ごとに区切り，被験者によるキーボード等への反応と共に，各区切りごと順番に提示する実験手法である（Just et al., 1982）．そして，各区切りにおける反応時間がそこで提示されている言語情報の処理に要する時間を反映すると想定される．通常被験者のキー入力と共にその区切りは画面上から消え，次の区切りが提示されるため（moving window 法と呼ばれる），一度次の区切りへ移動してしまうと読み返しが行えない点が前述の眼球運動測定との決定的な違いである．図 6 は Nakamura and Arai（2016, 実験 2）における自己ペース読み課題による実験の一つの区切りにおける読み時間の分布を示している．

データの分布は典型的に眼球運動の読み時間と似て，典型的に右に裾が長い左右非対称である．自己ペース読み時間ではキー入力を必要とするため，間違ってキーを連続で叩いてしまって極端に短い値が起きたり，気が散ってキーを入力するのを忘れてしまい極端に長い値が起きたりすることが稀にある．このような極端な値は明らかにエラーであるので分析から取り除く必要がある．ここで重要なのは，これらの値は「間違いによる値」であり，これは先に出た「外れ値」とは区別されるべきである点である．「外れ値」は大部分のデータの傾向からは逸脱しているけれども，実験に関係ある認知的処理を反映している可能性が排除できない値であり，その扱いには注意が必要である．これらの間違いによる値は大抵かなり極端な値を取るため，平均値への影響は大きく，このような値が 1 つでも含まれたままデータ分析が行われるだけで実際には影響のない説明要因の効果が間違って有意に判定されてしまうこともある（第 1 種の過誤）（Ratcliff, 1993）．このような事態を避けるためにエラー値は最初に除外する必要があり，一般的に 100 ms 程度を下限としてこれを下回るデータポイントは除外されている．上限につい

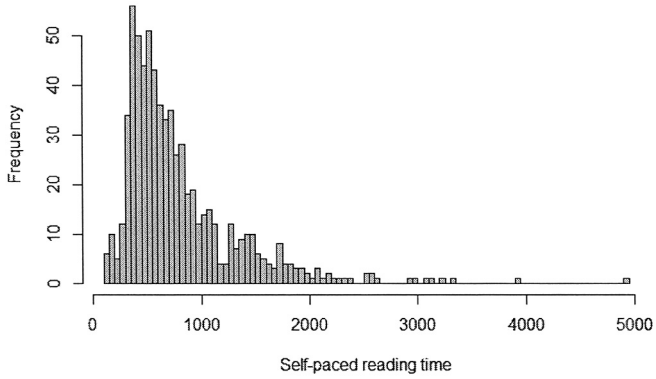


図 6. 自己ペース読み時間の分布例(Nakamura and Arai, 2016, 実験 2 より).

てははっきりとしたコンセンサスはないが(橋本, 2010 に各研究の基準がまとめられている), 2000 ms を超えるデータポイントが除外されているケースが多い(Heider et al., 2014). しかし実際には, データの分布を確認せずに特定の基準を適用するのは問題がある. なぜなら, 現実には極端だと思われるデータが「間違いによる値」か「外れ値」のどちらなのか判断が難しいというケースが多くあり, あくまで自分に都合のよい主観的で恣意的な判断を避け(特に実験条件を考慮に入れてこの判断を行ってはいけない), 客観的でデータに基づいた判断が必要である. そのため, Baayen and Milin (2010)は, 事前の「間違いによる値」のスクリーニングには非常に慎重になるべきであり, 常に最低限のデータスクリーニングとモデル批判を組み合わせるべきだと主張している. 具体的には上の例では, 5000 ms 付近にあるデータ 1 つを間違いによる値だと考え, これを除いたモデルで, 最適なモデルを構築する(モデル選択については後述), そしてこのデータポイントを含めたモデルと, あてはまりの良さを計る R^2 の値を比べ, 前者の方がより R^2 の値が高ければそのデータポイントは想定する分布に含まれるべきではない(つまりエラーによるもの)ので分析から除外することを正当化できる.

間違いによる値を除いた後, 残りのデータを従属変数として分析するのだが, 自己ペース読み課題による読み時間には課題特有の問題がある. それは, この課題ではキーを押すという不自然で身体的な運動を必要とするために, そこで計測される読み時間は, 言語情報の理解に要した認知的な処理時間と, キーを押すのに必要とした身体的な反応時間の両方によって構成されていると考えられる. この二つは必ずしも相関関係にはなく, 人によって理解の処理は遅くとも, キーの反応は非常に速いというケースも考えられる. そのため, いくつかの研究者は後者の影響を読み時間データから取り除くために, 事前にフィラーを含めたすべての文(練習文は除く)のすべてのリージョン(文頭・文末は除く)の読み時間を用いて(これによって課題全体における差, つまり刺激文に限定されない課題を行う上での反応の個人差を説明する)残差読み時間の計算を行っている(Ferreira and Clifton, 1986). 通常その計算には単純な線形回帰モデルが用いられ, 文字数を説明変数として含めることで, 1 文字あたりの読み時間の残差が計算できる. 残差読み時間の計算には, 線形回帰モデルの代わりに線形混合モデルを用いることも当然可能であり, これによって元々設定した説明変数以外の影響がありそうな他要因の効果を事前に取り除くことも可能である(e.g., Fine et al., 2013).

このように算出された残差読み時間データには未だ「外れ値」が含まれている. このため, 眼球運動の読み時間データの分析と同じように, 説明変数を含めた線形混合モデルを適用し, そ

の残差の標準偏差を元に外れ値の除外を行う必要がある（Baayen et al., 2008）．

#データ（dat）の読み時間（RT）に対して文字数（Wordlength），リージョン（Region），実験タイプ（ExpType）を固定効果に指定し（文字数はランダムスロープとしても指定），線形混合モデルで残差読み時間を計算するサンプルコード

```
m0<−lmer(RT∼Wordlength+Region+ExpType+(1+Wordlength|subject), data=dat)
#モデルから残差を算出する
data$RTresid<−resid(m0)
#実験文の特定のリージョン（この例では 3）の残差読み時間に対して LME モデルを適用
m1<−lmer(RTresid∼X*Z+(1+X*Z|subject)+(1+X*Z|item), data=dat[dat$ExpType ==
    "Exp"&dat$Region == 3,])
cor(fitted(m1), dat$RTresid)^2
#残差の絶対値が標準偏差 3 を下回るデータのみを選択し，もう一度モデルに当てはめる
newdat<−dat[abs(scale(resid(m1)))<3,]
m2<−lmer(RTresid∼X*Z+(1+X*Z|subject)+(1+X*Z|item), data=newdat)
cor(fitted(m2), newdat$RTresid)^2
```

経験上，元の読み時間の標準偏差を基準に除外を行った場合と，モデルの残差の標準偏差を基準に除外を行った場合では検定の結果に対する影響は最小限に留まることが多い．対照的に，外れ値を境界値で置き換えを行う場合と，除外する場合では，前者はデータの総数が変わらず，外れ値の影響も残るので検定の結果に大きな違いが出ることがあるので注意が必要である．また，被験者ごとに文字数に対して残差読み時間を先に求めてから，説明要因を含めて線形混合モデルで分析する手法と，残差を計算せず，始めから線形混合モデルを用い，文字数を共変数としてモデルに含める方法は同じではない点にも注意が必要である．前者の場合，上で述べた通りすべての文の全リージョンのデータを元にしており，後者では分析対象としている刺激文の特定のリージョンの読み時間のみにおける文字数の影響が考慮されている．現在線形混合モデルの普及と共に，残差読み時間を報告する例は減少してきているが，元となるデータサイズの差から各被験者のこの課題におけるベースパフォーマンスの違いと文字数の影響を説明する上では，前者の方が正確であると考えられる．

もう一つ，文字数に対して残差読み時間を計算する上で注意しなければならないのは，文字数が非常に少ないケースである．たとえば主語関係節文と目的語関係節文の読み時間を比べた Roland et al.（2012）の研究では一方で一人称主格代名詞（‘I’），もう一方で一人称目的格代名詞（‘me’）の読み時間を比較していて，前者の方が読み時間が早いことが示されているが（Roland et al., 2012, p.485），ここで文字数によって読み時間を割ってしまうと ‘me’ の読み時間が単純に半分となり，その傾向は逆転する．実際 2 文字の ‘me’ の方が一文字の ‘I’ よりも読むのに 2 倍の時間がかかると考えるのは現実的ではないため，このような場合には残差読み時間の計算は不適切だと考えられる（この場合，線形混合モデルに文字数を共変数として含めても同じ問題が起こる）．また言語の表記特有の問題もあり，日本語のように仮名や漢字のように異なる種類の文字の影響を無視して単純に文字数で読み時間を割るというアプローチがどれほど妥当であるのか不明である．そのため，日本語の読み時間の分析において残差読み時間の計算は適切だとは考えにくい．このような問題を始めから避けるためにも自己ペース読み課題実験においても同一リージョンで条件間での文字数はできる限り揃えることが望ましい．

4. 残された問題

4.1 データ変換

線形混合モデルを用いた分析において，読み時間データに何らかの変換を加えるべきかという議論がある．変換を加えるべきだと考える理由は大きく 2 つあり，一つには，先に見たとおり，眼球運動の読み時間データも自己ペース読み課題による読み時間データもどちらも右に裾が長い左右非対称の分布形状（歪度 > 0）を持ち正規分布に従わないため，平均値の推定に影響を及ぼす．つまり分布の右側の長い裾に位置するデータが主要なデータパターンを歪ませてしまうことである．もう一つには，このようなデータに線形混合モデルを適用した場合，残差の分布も正規分布には従わず，不等分散性を示す．これによって平均値と標準誤差の推定が影響を受けるため，正しい固定効果の有意判定が行えない．この問題は読み時間を含めた反応時間全般に共通していて，そのため過去の研究においてデータ変換を行っている例が報告されている．データ変換の方法は多数あるが，実際のデータに対して最もあてはまりが良くなる方法というのは一概には言えず，それぞれの実験データごとに適した方法が異なるため，複数の方法を確認してみる必要がある．しかし現実には，逆変換（back-transforming, つまり元の変換前の値に戻すこと）の容易さから対数正規（Log Normal）分布か逆ガウス（Inverse Gaussian）分布[5)]を仮定した変換のどちらかが用いられることが多いようである（Baayen and Milin, 2010; Juffs, 1998; Juffs, 2005; Frank et al., 2013）．

```
#逆ガウス分布を仮定した変換
d$inv_rt<−−1000/d$rt　#解釈のし易さから 1/RT の代わりに −1000/RT を採用
#対数正規分布を仮定した変換
d$log_rt<−log(d$rt)　#デフォルトで自然対数が計算される
```

図 7 は元々の読み時間（a）と，対数変換した読み時間（b），逆ガウス変換した読み時間（c）に対する線形混合モデルを当てはめ，モデルから予測された値と残差の関係を残差対推定値グラフ（Residuals vs. fitted plot）で示している．もし残差の分布が従属変数の値と独立して均等である場合，グラフ上で y 軸が 0 の値を中心に x 軸上すべての範囲において残差が散らばる．この図からわかる通り元々の読み時間ではモデルの推定値が高くなるほど，残差が大きくなっていて，左右非対称となっていることがわかる．一方対数変換した読み時間（b）と逆ガウス変換した読み時間（c）ではモデルの推定値とは関係なくほぼ均等になっているため，よりモデルのあてはまりがよいと言える．

```
*ここでは m2 を最適モデルと仮定する
plot(fitted(m2), resid(m2), xlab="Fitted Values", ylab="Residuals")
abline(h=0, lty=2)
```

データ変換に関しては多くの議論があるが，中には結果に基づき解釈を後付け（posthoc）で考えることに繋がるとしてデータ変換は行うべきではないという意見がある．これは，データ変換を行うことで，読み時間と説明変数との間の線形関係が失われ，説明変数の読み時間への影響を非線形関係によって説明する必要が出てくることに基づいている．いわゆる ‘mental chronometry’（心的時間測定法）と呼ばれる認知処理の時間を測定する研究全般においては，諸々の説明要因の影響は変換なしのそのままの読み時間に対して仮定されているのが一般的である．つまりある要因によって読み時間が遅くなった場合，その遅くなった分の時間の長さそのものがある要因によって引き起こされた認知処理の負荷増加を直接反映していると考えられている（Townsend, 1992）．そのため，上記のようなデータ変換はそういった仮定に反するため，そのままの読み時

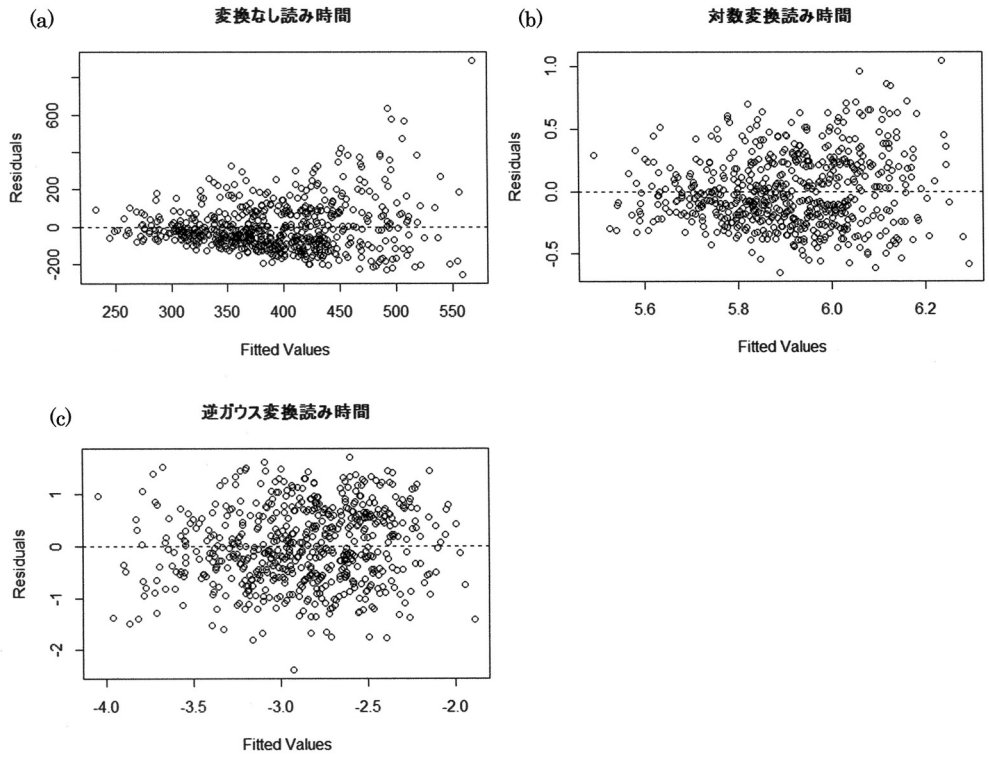


図 7. First pass 読み時間の変換なしの読み時間(a)，対数変換した値(b)，逆ガウス変換した値(c)に対する線形混合モデルの残差対推定値グラフ．

間を分析対象とするべきだと考えることもできる．

この問題のひとつの解決方法として，従属変数としてはそのままの読み時間を使うが，そのデータサンプルが属する元の確率分布を推定し，理論的仮説に基づく適切なリンク関数を指定することができる．つまり，そのままの読み時間に対する説明変数の影響を仮定しながら，同時に統計モデルによって課せられる数学的な制限(正規分布に従わないデータ分析)を満たすことのできる一般化線形モデルを使うという方法である(Lo and Andrews, 2015)．Lo and Andrews の研究では，Balota et al. (2013)の反応時間データに対して，複数のデータ変換およびリンク関数を試し，最もあてはまりのよいモデルを比較検証している．その結果，恒等(identity, つまりそのままの値を用いる)リンク関数を指定し(元々の反応時間に対して説明変数との線形関係を仮定)，そして確率分布として逆ガウス分布を用いた一般化線型モデルが理論的にも数学的にも最も適していると結論づけている．

```
#逆ガウス分布と恒等リンク関数による一般化線形混合モデルの実行(Lo and Andrews, 2015 参考)
glmer(RT~X*Z+(1+X*Z|subject)+(1+X*Z|item), family=inverse.gaussian(link="identity"),
    data=dat)
#ガンマ分布を仮定する場合は"inverse.gaussian"を"Gamma"に置き換える
```

これに従うと，反応時間特有の分布を持つデータは変換をしない，つまりそのままの読み時間を従属変数として(恒等リンク関数を指定して)モデル化することが理論的に正しいと考えることができる．残りの問題は，個々のデータに適した確率関数を見つけ出すことであり，このためにはモデルの分布と実際のデータの分布を Q-Q Plot 等で視覚化して比較し，さらに AIC/BIC などのモデル選択基準を用いて判別することが必要になる．

また，データ変換における問題は，どのような実験デザインのデータにはどのデータ変換が適しているのか，または複数の変換方法を試してどの程度特定の分布と近似したらその変換が適していると言えるのか，などについて客観的な判断基準がない事が挙げられる．シャピロ・ウィルク検定や，Q-Q Plot を用い，正規分布に従うか否かチェックし，著しく外れている場合には分布形状に合ったデータ変換を試みて，結果がどう変わるのか試し実証的に判断を行わなければならないが，データ変換によって検定結果の解釈自体が大きく変わることは実際少ないと考えられる．もし変わった場合にはその理由を突き止めることが重要である．一般的に読み時間において，データ変換を行うことで最も影響を受けるのは分布の右の裾に位置する著しく長い読み時間である．対数変換などのデータ変換は通常これら目立った読み時間を少し目立たなくしてくれる(平均値の推定への影響力を軽減する)．そのため，そういった少数の目立って長い値が特定の条件に偏って起きていたことで交互作用が有意になっている場合，データ変換(または外れ値の除外)によって交互作用の有意差が失われるということが起こりうる．このような場合，各説明要因の読み時間への影響に関する仮説に応じて(仮説が元々の読み時間に関するものであればデータ変換しない結果を支持する)判断する必要がある．データ分析を行う上でしてはならないのは様々な手法を有意差判定が出るまで試し，有意差が観測された結果のみを報告することである．当然ながらこれは，同データに対して多重比較を行っているため有意水準(α レベル)の問題が起こり，第一種の過誤の確率が高まる[6]．

4.2 モデル選択

George Box による有名な “All models are wrong” という言葉からも明らかなように統計モデルはあくまで実際のデータ(または現実)に対する「近似」(approximation)であり，その近似が(完全に正しくなくとも)現実に役に立つには，データの特性を最大限に説明できながら同時に可能な限り経済的でシンプルなモデルの構築が必要である(Box, 1976)．そのためにも実際に得られたデータに最もあてはまりのよいモデルを選択する，つまりモデル選択が大切であるが，どのようにして最適モデルを見つけるかについては様々な意見や方法が存在する．既に述べたように線形混合モデルは実験を行う前に設定した操作の影響と，実験を行う前には知り得ないランダムな影響を同時に理論上はいくつでも式に含めることができる．そのため，探索的に影響があるかも知れない要因を説明変数としてすべてモデルに含め，そこから AIC などのモデル選択基準を用いて，不必要な要因を削っていき最適モデルを探す「探索的データ分析」(exploratory data analysis)を行うことができる(久保, 2012)．この点も ANOVA を代表とする従来の分析方法とは決定的に異なる．

これとは別に，実験を行う前に結果に対する明確な仮説を立て，その仮説を実証するために必要な要因を設定し，その要因効果を検定するという「確証的仮説検定」(confirmatory hypothesis testing)を行う場合があり，言語心理学や実験心理学ではこちらの方がより一般的だと言える．後者の場合，2×2 などの決まった説明変数の影響の有無が興味対象であるので，モデルにおける独立変数は固定しておく場合が多い(つまりたとえ有意な差がなかったとしても実験操作としては存在していたのだからモデルから削らない)[7]．このような理由から，最適モデルを選ぶ上で問題となるのは最適なランダム効果構造の選び方である．この際，AIC のように，パラメータの数にペナルティを課すことで，モデルの実際のデータへのあてはまりの良さではなく，必

要最小限のパラメータによる予測の良さを重視する選択基準を使って，ランダム効果構造のみが異なるモデル間の比較を行った場合，大抵単純にパラメータの数が少ないモデルほどAICの値が小さくなる．これは，ランダム効果構造の違いがモデルの最大対数尤度に与える影響は小さいことを意味する．

線形混合モデルにおいて，説明変数の効果はランダム効果の構造を元にして評価されるため，このランダム効果構造をどう構築するかによって説明変数の効果は変わってくる．今までの研究により，固定効果とランダム効果との交互作用(以下「ランダムスロープ」)を含まないいわゆる「切片のみモデル」(Appendix 39行目のコードのm3に該当)では第一種の過誤が起こる可能性が高く，始めから切片のみモデルのみを使ってデータ分析を行うことは不適切であることがわかっている(Barr et al., 2013; Roland, 2009)．実際に，Roland (2009)はコーパスデータを使って，ある一つの項目のみで突出した大きな効果が観測されたデータにおいて(たとえそれ以外の項目では全く効果がなかったとしても)切片のみモデルを用いると，その効果が有意に判定されることがあることを明らかにしている．このような場合に今までのように被験者と項目で別の分析を行うと，被験者分析では有意な差が見られるが，項目分析では見られない(すると保守的な $minF'$ でも有意差は見られない)．そのためRoland (2009)は，このような第1種の過誤の危険を避けるためにも線形混合モデルの分析結果と合わせて分散分析の結果(F1とF2両方)を並記することを勧めている．

切片のみモデルを始めから採用してはいけない，ということは合意が得られていそうだが，ではどのように最も妥当なモデルを選ぶべきかについては意見がまとまっていない．Barr et al. (2013)は，確証的仮説検定においては第1種の過誤の危険性をできる限り最小にすることが必要で，そのためには常にデータが収束し得る最大のランダム効果構造を持ったモデルを採用すべきだと主張している．彼らは仮想データのシミュレーションに基づき，被験者内要因を含む実験デザインにおいて切片のみを含めた線形混合モデルでは第1種の過誤の確率が壊滅的に上昇する可能性があることを示した．彼らは，すべての説明変数とランダム効果との間に，すべての組み合わせの交互作用を含んだ最も複雑なランダム効果構造をもつ，いわゆる「最大モデル」(Appendix 25行目のコードのm0に該当)と呼ばれるモデルと，モデル選択をして得られた最適モデルとの間の分析力の違いはおおよそ無視できる程度であると主張している．さらには，たとえ実験デザイン上存在しないランダム効果の構成要素を含むモデル(いわゆる「パラメータ過多(Overparameterized)モデル」)においても分析力はほぼ変わらなかったことを報告している．Barr et al.はこの結果によって，モデルのアンダーフィッティングの悪影響は甚大だが，オーバーフィッティングによる影響は無視できる程度だと考えられ，常にすべての説明変数をランダム効果のスロープとして含める最大モデルを採用するべきだと結論づけている．

Barr et al.の主張は現在多くの論文で引用され大きな影響を持っているが，疑問視されている点もある．主な点として，彼らがシミュレートしたデータは1要因2水準のみ含む非常にシンプルな実験デザインを想定していることである．実際の心理言語実験などでは，通常複数の固定効果とその要因間の交互作用を含み，さらには探索的に試行回数や文字数の影響などを共変数としてモデルに追加したりするため，ここで扱われているデザインよりはるかに複雑なモデルを構築する必要がある．そのため，彼らのシンプルな実験デザインにもとづく結果がどの程度より一般的な実験デザインに当てはまるかについては疑問が残る．その一つの証拠として複雑なランダム変数構造を持つ線形混合モデルではしばしば収束しないという問題が起こる．これは，モデル評価のアルゴリズムにおける欠陥ではなく，単純に実際のデータによってサポートできない過度に複雑なモデル，つまりパラメータ過多モデルであることに起因する(Bates et al., 2015)．実際最大ランダム効果構造を持つモデルでのパラメータの数は「相関パラメータ」[8]も含めると，一般的な 2×2 デザインでは20，$2 \times 2 \times 2$ デザインでは72と指数関数的に増加す

るため，30 程度の被験者数とアイテム数の組み合わせからなる各試行のデータからそれらすべてのパラメータを正しく推定できると考えるのはやや楽観的過ぎるように思われる．それゆえ Bates et al. は，パラメータ過多は，たとえ収束したとしても，解釈不能なモデルを構築することにつながるとして Barr et al. の主張を批判している．彼らは，実際に複雑な実験デザインから得られるデータに対して最大モデルを用いた場合，分散の推定はパラメータ過多によって信頼性が下がることを実証している．

Bates et al. はこの問題を解決するために，探索的データ解析によく利用される主成分分析 (Principal Components Analysis, 以下 PCA) とよばれる統計手法を用いて，分散を説明する主成分の個数を割り出し，そこからモデルを単純化していくことを提案している．彼らは R の RePsychLing パッケージを公開して，それに含まれる rePCA 関数を使って，まず (1) ランダム効果構造内の相関パラメータを含めた最大モデルと，それらを含めない最大モデルの両方で PCA を行い，必要な主成分の個数を算出しモデルがパラメータ過多になっていないか確認する．相関パラメータを含めないモデルはパラメータの数が少ない分，その分散の推定値は信頼性が高いので，そのモデルにおいても分散がゼロに近い構成要素が含まれている場合かなりの確率でモデルがパラメータ過多に陥っていると判断することができる．そして，(2) その相関パラメータを含まないモデルから，分散の値の小さい構成要素から順に尤度比検定 (Likelihood ratio test) を用いて有意でないランダム効果の構成要素を取り除き，モデルを簡略化していく．これをモデルのあてはまりが有意に低くなるまで繰り返し続ける (backward selection または iterative reduction approach と呼ばれる)．こうして相関パラメータを含まない最適モデルまでたどり着いたら，最後に (3) 残されたランダム効果の構成要素間の相関パラメータを加えてモデルのあてはまりが有意に高まるか再び尤度比検定で確かめる．有意に高まる場合には相関パラメータを含めたモデルを，高まらない場合には含めないモデルを最適モデルとして採用する．相関パラメータを最後まで加えないのは，意味を持たないランダム効果の構成要素に対して他の要素との相関関係を想定するのは非論理的で現実的ではないからだと考えられる．

Barr et al. が警告しているように，モデルを簡略化することはモデルのアンダーフィッティング，および第一種の過誤の確率を高める可能性があるので，慎重に行う必要がある．そのためいくつかの研究者は尤度比検定において，保守的な有意水準として 0.10 を採用している (つまり，$p < .10$ である場合にはそこでモデル選択を止め，複雑な方のモデルを採用する，Clifton, 2013)．ここで追記すべき重要な点として，Bates et al. の論文では上記の過程を通して選ばれた最適モデルの結果と階層ベイズモデルによる分析の結果を比較している．その結果，この二者間では固定効果の推定はほぼ同一であり，さらに PCA を用いて特定した主要な分散成分のパラメータは階層ベイズモデルを用いた分析において支配的だったパラメータと正確に合致したと報告している．これは線形混合モデルを用いた分析手法と，近年広がりをみせている階層ベイズモデルを用いた手法を比較する上でも非常に重要な報告と言える．これを踏まえると，非常に単純な実験デザインを用いていない限り，最大モデルを全てのデータ分析において採用するのはパラメータ過多のリスクが伴い現実とは言えない．そのため，PCA と尤度比検定を併用し，実際のデータによってサポートされる最適モデルを慎重に探索する Bates et al. のアプローチが，少なくとも現時点では，最良であるように思える．参考として Appendix にこの分析方法を行う手順の R コードを掲載しておく (詳しくは RePsychLing パッケージ内の各ドキュメントを参照してほしい)．

4.3 *p* 値の算出

線形混合モデルにおいては，自由度の決定が難しく，そのため係数ごとに計算される t 値には自由度が考慮されていない．そのため，線形混合モデル分析を行う代表的な R パッケージで

ある lme4 では p 値の計算が行われない（Bates, 2005）．これに対して様々な方法が提案されていて，それぞれの方法のメリット・デメリットが現在も議論されている（この点に関しては以下のホームページに詳細な情報が載っているので参照して欲しい．http://glmm.wikidot.com/faq）．その中でも，最もシンプルで，特定のパッケージ・アルゴリズムに依存しない方法は，p 値を報告せず，正規分布を仮定し，各効果に対して産出された t 値の絶対値が 2 と等しいかそれ以上の（つまり 0 から標準誤差 2 つ分以上離れている）場合に有意な差があると判別することである（Gelman and Hill, 2007）[9]．しかし，当然ながらこの方法のデメリットはその効果の再現確率がどの程度なのか明確に提示することができない点である．そのため，p 値を算出する一つの方法は，最適モデルから，p 値を産出したい効果のみを除いたモデルを用意し，尤度比検定を用いて二つの分布を直接比較し，その効果が除かれることでどれだけ分布に違いが生まれるかを見ることである．分布の形状を問わず，二つの分布の違いはカイ二乗分布に従うことが知られているので，この比較から得られた p 値を報告することができる．過去の研究から，尤度比検定によって求められる p 値は十分な信用性があることがわかっている（Barr et al., 2013）．具体例として 2 要因 (X, Z) 交互作用の p 値は，以下のように R コードを指定することで算出できる（ここではランダムスロープの構造として被験者ランダム効果に対する説明変数 X のスロープのみ含んだモデルが最適モデルだと仮定している）．

```
m4<−lmer(RT∼X+Z+X:Z+(1+X | subject)+(1 | item), REML=F, data=dat)
m4i<−lmer(RT∼X+Z+(1+X | subject)+(1 | item), REML=F, data=dat)
anova(m0,m0i)
```

ちなみに lme4 パッケージの lmer 関数は最尤法としてデフォルトで最尤法（Maximal Likelihood: ML）ではなく比較的小さなデータでもバイアスの少ない制限付き最尤法（Restricted maximum likelihood: REML）を採用しているが，後者による推定結果は尤度比検定で比較できないため，上のモデルでは REML=F を指定し最尤法を適用している．個々のデータによるが，データサイズが一定以上大きければ両者の計算結果に大きな差が生まれることは少ないと予測される．この他にも，比較的容易に p 値を算出する方法として R の lmerTest パッケージの lmer 関数を用いることができる．このパッケージでは自由度を Satterthwaite 近似法を用いて算出し p 値を算出している[10]．この lmer 関数は lme4 に依存しているため，lme4 パッケージを用いた分析結果と一致する．またこのパッケージでは，モデル選択を行う step 関数が用意されていて一度にモデル選択を行うこともできるが，前述した Barr et al. らの手法による結果とどの程度一致するかは不明である．

4.4 下位検定

言語研究においてよく用いられる要因デザイン（factorial design）において有意な交互作用がみられた時，どのようにして単純主効果を検定すべきかという問題がある．一つの方法として片方の説明変数における一つの水準のデータ（サブセットデータ）のみを抜き取り，単純主効果を調べたい説明変数のみを含めた新しいモデルで有意差検定を行っている例がみられる（e.g., Nakamura et al., 2012）．しかし，この方法は同じデータに対して複数のモデルを当てはめることになり，多重比較の問題が生じ第 1 種の過誤の確率が上がる．前に出てきた R の lmerTest パッケージに含まれる difflsmeans 関数などを使うと，自動的に主効果に加えすべての下位レベルの組み合わせの検定結果を p 値と共に出力するが，この方法も同じ理由で問題がある．また，2×2 デザインで片方の要因が時間軸や試行順序のように連続変数であり，もう一方の要因と交互作用が見られた場合に，各水準でどのような変化が起こったのか知りたい場合，以前は前半と後半というように連続変数からカテゴリー変数へ変換し，データのサブセットに対して下位

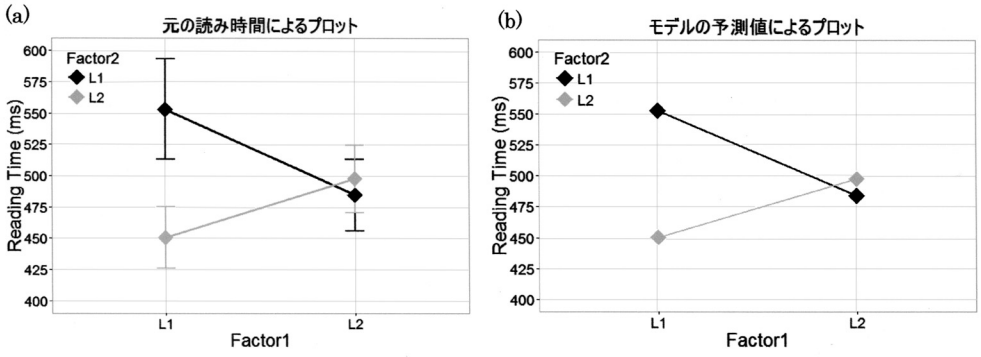


図 8．実測値による交互作用プロット(エラーバーは標準誤差)と最適モデルの切片，説明変数の係数から計算された値による交互作用プロット．

検定を行うということが行われてきた．しかし，この方法には多重比較の問題の他に連続変数のデータが失われるというデメリットがある．

まず交互作用が得られた時に最も重要なのは結果をグラフ化することである．その際エラーバーとして 95%信頼区間，もしくは標準誤差を表示することが非常に大事になる(標準偏差はサンプルデータのばらつきを示す記述的な値なので避ける)．特に 95%信頼区間は有意差との関係が直接的でありわかりやすいため非常に有益である．当然サンプルサイズに依存するが，エラーバーが 95%信頼区間を表す場合，二つのバーを互いに近づけていって，バーの全長の約 25%ほどオーバーラップした辺りが，有意水準($\alpha = 0.05$)にあたる．つまり，そこを境界としてそれよりお互いのバーが重なり合っている(つまり平均値が近い)場合には有意差はなく，逆にそれよりさらに離れている場合には有意差があると判定できる(標準誤差の場合にはお互いのバーがオーバーラップしておらずかつバーの長さの半分以上離れている場合有意な差があると考えられる)．実際，現在でも多くの研究者がエラーバーを提示していない，していても何を表しているか記述がない，またはエラーバーの解釈自体よくわかっていないことが多いという指摘もあり注意が必要である(Belia et al., 2005)．

このように視覚的に確認することで，どのようなパターンの交互作用が起こっているかおおよそ見当をつけることができる．もう一つ役に立つ方法は，線形混合モデルの推定値によって交互作用のパターンをグラフ化することである．モデルの実際のデータへのあてはまりが良ければ，モデルからの推定値から計算される平均値と，元々のデータから計算される平均値は非常に近似するはずであり，あてはまりの良さを視覚的に確認できる(図 8 参照)．

これは試行順序や時間などの連続変数と説明変数との交互作用が見られた時には特に有益である．実際の値はばらつきが大きく，連続変数の効果のパターンが読み取りにくい場合がある．そのためモデルによって算出された説明変数の係数は，ランダム効果による分散を取り除いた上での言わば純粋な説明変数の効果を示しており，その理論値に基づいてグラフを作ることで，交互作用のパターンが把握しやすくなる．たとえば説明変数を中心化して平均値が 0, 標準偏差が 0.5 である場合(つまり各水準は $-0.5, +0.5$ にコーディング)，モデルから求められた切片と各説明変数及び交互作用の係数から，以下のように各セルの平均値を手計算で求めることができる[11]．ダミーコーディング $(0, 1)$ の場合，平均値が 0.5 になるので(標準偏差は同じ)以下の各 $\pm$ 標準偏差の値に平均値 0.5 を先に加算する必要がある[12]．連続変数を中心化した場合平均値は 0, 標準偏差は水準の数に依存し，計算した標準偏差の値を以下の式内の ± 0.5 と差し替える

ことで算出できる．

```
#代数を設定(直接式を書いてもよい)
b0=319.12    #切片(数値は任意)
b1=−18.00    #説明変数 X の係数
b2=−8.53     #説明変数 Z の係数
b1b2=46.76   #X と Z の交互作用の係数

(−0.5*b1)+(−0.5*b2)+((−0.5*−0.5)*b1b2)+b0    #b1=−0.5, b2=−0.5 のセル平均
(−0.5*b1)+(0.5*b2)+((−0.5*0.5)*b1b2)+b0      #b1=−0.5, b2=0.5 のセル平均
(0.5*b1)+(−0.5*b2)+((0.5*−0.5)*b1b2)+b0     #b1=0.5, b2=−0.5 のセル平均
(0.5*b1)+(0.5*b2)+((0.5*0.5)*b1b2)+b0        #b1=0.5, b2=0.5 のセル平均
```

特に試行順序の学習効果などを調査する時には，中心化された連続変数のコーディングの最大値で ±0.5 を差し替えることで，最初と最後の試行におけるモデルからの推定値を求めることができ全体のデータパターンを把握するのに非常に便利である．当然，推定値のグラフにもエラーバーを提示するのが望ましく，そのためには各水準の推定値に対して 95%信頼区間(もしくは標準誤差)を計算する必要がある．これには理論的な分散・共分散の推定値を求める必要があり，複雑な計算が必要となる．おそらく最も容易な方法は，上で触れた lmerTest パッケージの lsmeans 関数を用いる方法である．最適モデルに対してこの関数を用いることで各水準の推定値と 95%信頼区間を算出してくれる．他には，先に紹介したウェブサイト(http://glmm.wikidot.com/faq)において，R コードが公開されているのでこちらを利用しても算出できる(ggplot2 パッケージを使ったグラフ化のコードも掲載されている)．手元のデータで試した限り二つの方法による推定値と 95%信頼区間の値はほぼ完全に一致した(推定値に関しては上の手計算による値とも一致)．

モデルを再計算することなく，下位検定を行う最もシンプルなのは，単純主効果を調べたい要因のモデルにおける主効果に対する 95%信頼区間(上は各要因の平均値の 95%信頼区間であり，ここでは水準間の差の 95%信頼区間を意味する)を計算し，単純主効果の平均値(被験者平均に基づく)における差がその 95%信頼区間を上回っていたら有意であると判定する方法である．最適モデルから 95%信頼区間を推定するには，最適モデルによって推定された主効果の標準誤差(SE)を 2 倍した値(2 SE)をそのまま 95%信頼区間として採用することができる(Sturt et al., 2010)．これはサンプル数が一定数以上大きい場合には，95%信頼区間は標準誤差の 2 倍として推定できることに基づいている．95%信頼区間の推定方法としては他にもブートストラップ法などの方法があり，どのように推定したかについては明示する必要がある．通常，上記の標準誤差を使う方法はブートストラップ法などの方法よりも保守的な値(つまり大きい値)が得られるため，第 1 種の過誤の確率は低いと考えられる．

```
#ブートストラップ法による 95%信頼区間の算出
confint(m2, method="boot")
```

たとえば，モデルによる要因 1 の標準誤差の推定値が 20 だとした場合には，帰無仮説の平均値 =0 を中心として ±40 の間に 95%の確率で母集団における要因 1 の水準間の平均差が来るはずである．しかし，実際のデータにおいて要因 2 の片方の水準における要因 1 の水準間の平均(周辺平均)の差が絶対値で 40 を超えた場合，その差は 5%以下の確率でしか偶然には起きないのでその単純主効果は有意であると結論づけられる．

下位検定のもう一つの方法として，分けて分析したい要因の水準を 0 値としてコーディング

することができる(‘computer code’ と呼ばれる，Aiken and West, 1991; Dawson, 2014)．固定効果を二つ(X, Z)とその交互作用を含む回帰式$(Y = b_0 + b_1X + b_2Z + b_3XZ + \varepsilon)$において，$b_1$ は Z が 0 の時の X と Y の関係性を表す．つまり，Z の説明変数をダミーコーディングで，$(0, 1)$ とコーディングすると，b_1 は Z が 0 である場合の X の単純主効果に相当する．これを今度は逆に $(1, 0)$ とコーディングすることで今度は Z のもう一方の水準の X の単純主効果の検定結果を調べることができる．重要なのは，2 つのモデルは元々の中心化したモデルと同じ固定効果を持つ，すべてのデータを含めた実質的に同一モデルであるため(実際交互作用の係数は変わらない)，同じデータのサブセットに限定した下位検定のような多重検定の問題がない点である．また，この方法は試行順序や時間などの連続変数との交互作用の下位検定にも用いることができるメリットもある．

5. まとめ

本稿では言語理解研究における眼球運動測定実験及び自己ペース読み課題によって得られるデータの分析方法をまとめ，データ形式・構造に合った適切な分析アプローチを考察した．これらすべての分析において線形混合モデル及び一般化線形混合モデルを用い，分散分析に代表される今まで広く用いられてきた分析に対する優位性を説明してきた．しかし，その具体的かつ詳細な適用方法においてはまだ課題の残る点も多く，現在も研究者によって異なるアプローチが採用されている．本稿のひとつの目的は，これら異なるアプローチを比較し，それらが分析結果にどのように影響するか検討することである．すでに説明した通り，全てのデータタイプに対して唯一の理想的な分析方法は存在せず，ここで紹介した方法は全て多かれ少なかれメリット・デメリットがあり，個別のデータに最も適した方法を各研究者が検討し，採用しなければならない．その際の基本的な考え方として以下の 4 点を本稿のまとめとして示しておく．

1) 不要なデータの集約は避け，各試行で得られたそのままのデータを分析対象とし，統計モデルを適用する前に様々な角度からデータの特性を調べる．その際グラフを描画し，視覚的にデータのパターンを見ることが重要である．これによってうまく視線が記録できなかった被験者やエラー値がないか，また外れ値がどのように分布しているかチェックする．
2) データ変換，またはモデル選択などデータ分析における選択肢が複数ある場合には，第 1 種の過誤を避けるためにもまず最も保守的な方法(データの変換なし，最大モデル)から検討する．
3) そして実際のデータ構造に基づく合理的な判断によって，より適切な方法，実データに最もあてはまりが良いモデルを探索する．
4) 採用した分析手法が適切であるか確証が得られない場合には，分散分析など他のアプローチによる結果と比較し，必要であれば結果を並記し報告する．

最後に，データ解析理論・手法の発展には情報の交換が最も重要であり，そのためにも，各研究者が論文，学会・ワークショップなどを通して研究を発表する際にはどのようなアプローチを用いて分析が行われたのか詳細に報告することが不可欠だと考える．また研究コミュニティー全体で統計分析手法をオープンに共有することによって個々の実験デザインから得られるデータに対してどのアプローチが適切であるか自然と明らかになってくるものと期待され，本稿がその一端を担えたら非常に幸いである．

Appendix

RePsychLing パッケージの rePCA 関数を用いた Principal Components Analysis(PCA)の手順

```
#初めて RePsychLing パッケージをインストールする場合のみ以下の 2 行を実行する.
install.packages("devtools")
devtools::install_github("dmbates/RePsychLing")

#必要なパッケージの読み込み
library(RePsychLing)  #rePCA 関数に必要
library(lme4)         #lmer 関数に必要
library(MASS)         #truehist 関数に必要

#データ(d)読み込みとデータ構造のチェック
setwd("任意のファイルの場所をフルパスで指定")
dat<−read.csv("readingtimedata.csv", header=T)
head(dat)
summary(dat)
str(dat)

#説明効果 (X, Z) の中心化
dat$cX<−scale(dat$X, scale=F)
dat$cZ<−scale(dat$Z, scale=F)

#読み時間(RT)の分布のチェック
truehist(dat$RT, 100, prob=F, col="gray", xlab="読み時間", ylab="頻度")

#1 要因(X)の水準ごとの確率密度プロット
plot(density(dat[dat$X == 1,]$RT), xlim=c(0,3500), ylim=c(0,0.0015), lty=1,
main="要因 X の各水準における確率密度")
lines(density(dat[dat$X == 2,]$RT), lty=2)

#水準が多い場合(ここでは X*Z)には箱ひげ図が便利
boxplot(RT∼X*Z, data=dat, ylim=c(0,800), col=(c("white","darkgray")),
names=c("a","b","c","d"))

#まず相関パラメータを含む最大ランダム効果構造モデル (m0) がパラメータ過多となっ
ているかチェック
m0<−lmer(RT∼1+cX+cZ+cX:cZ+(1+cX+cZ+cX:cZ | subject)
+(1+cX+cZ+cX:cZ | item), REML=F, data=dat)
summary(m0, corr=F)

#subject と item それぞれのランダム効果において意味のある分散構成成分がいくつある
かチェック
summary(rePCA(m0))   #rePCA 関数を用いて主成分分析を行う

#次に相関パラメータを含まないモデル (m1) が未だパラメータ過多となっているかチェ
ック
#二重縦線( || )は相関パラメータを含まないことを意味する
m1<−lmer(RT∼cX+cZ+cX:cZ+(1+cX+cZ+cX:cZ||subject)+(1+cX+cZ+cX:cZ||item),
REML=F, data=dat)
summary(m1, corr=F)
```

```
anova(m1, m0)　#通常 m0 の方があてはまりがよいがひとまず無視する
summary(rePCA(m1))　#再び rePCA 関数を用いてより信頼性の高い主成分分析を行う

#モデルの簡素化：分散の値の少ない構成要素から順に尤度比検定を使って有意にフィットが下がるまで削っていく．ここでは cX のランダムスロープ以外のパラメーターはすべて有意差がなかったと仮定
m2<−lmer(RT∼cX+cZ+cX:cZ+(1+cX||subject)+(1|item), REML=F, data=dat)
anova(m1, m2)

m3<−lmer(RT∼cX+cZ+cX:cZ+(1|subject)+(1|item), REML=F, data=dat)
anova(m2, m3)
#ここで有意差が確認され m2 の方が m3 よりもあてはまりが良かったと想定

#相関パラメータを m2 に加えて有意にあてはまりがよくなるかチェック
m4<−lmer(RT∼cX+cZ+cX:cZ+(1+cX|subject)+(1|item), REML=F, data=dat)
summary(m4, corr=F)
anova(m2, m4)
#もし有意な差が確認できたら m4 を最適モデルとして選択し，残差の標準偏差を基準に（この場合 3 倍以上）外れ値の除外を行う場合は以下のように指定してモデルをアップデートする．
summary(update(m4,subset=abs(scale(resid(m4)))<3), corr=F)

#各水準におけるモデルの推定値および 95%信頼区間を lmerTest パッケージを用いて求める
library(lmerTest)
#最適モデルを再度走らせてから lsmeans を適用
m4.2<−lmer(RT∼cX+cZ+cX:cZ+(1+cX|subject)+(1|item), REML=F, data=dat,
subset=abs(scale(resid(m4)))<3)
lsmeans(m4.2)
```

注．

[1] 読みにおいては常に左から右，上から下へ進むのではなく，約 2 秒に一回程度既に通り過ぎた箇所へ後戻りする移動運動（regressive saccade）が起こっていて，認知的に困難な情報に直面した際に頻繁に起こるので，この発生頻度を指標にした分析も行われるが本稿では割愛する．

[2] このようにデータを二項変数に変換せずに，二つ以上の対象物への注視をそのまま多値（polytomous）変数として分析できる多項ロジットモデル（Multinomial logit model）も存在するが，ここでは扱わない（Barr and Frank, 2009）．

[3] 割合とロジットを両方計算して検定を行った場合，割合の平均値が 0.3 から 0.7 に収まる場合には検定結果にあまり違いが見られないと報告されている（Barr, 2008）．しかし VWP における眼球運動データにおいては複数の対象物が存在するため多くの場合その下限を下回るので割合の計算には問題が生じる．

[4] Arai and Nakamura (2016) は読みにおける眼球運動データ（Right-bounded 読み時間）をオンラインで公開している．以下のアドレスから Appendix の S2 Appendix として csv ファイルのデータをダウンロードすることができる．この論文の実験は交互作用を含む 2×2

の一般的なデザインなので，本稿最後に記述されている R のサンプルコード（Appendix）を適用するサンプルデータとして利用できる．
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0156482

[5] 逆ガウス分布はワルド分布とも呼ばれ正の値しかとらず正に歪んでいる．そのため右に裾が長い分布を持つデータのモデル化に使用される．

[6] Second pass と total time という一度右に抜けた後に再び戻った注視を読み時間を含める比較的遅い処理を見ている場合，0 の値（つまりスキップ）を分析に含めるのが慣例である．0 の対数は定義されていないため，これらのデータには変換が行えないという問題がある（ゆえに眼球運動読みデータは通常変換されない）．しかし 0 データの頻度が多い場合そもそも分布形状が特殊になるため（0 の値は右に裾の長い分布の一部とは考えにくい）データ変換とは関係なく，正規分布を仮定した線形モデルの採用に問題が生じる．

[7] しかし実際はこのような場合においても，練習効果や，疲労効果の影響など個々の試行のレベルで実験基準とは独立しているがデータの分散に影響を及ぼしている要因はいくつも考えられ，これらの効果の影響を調べ，これらの要因による分散を取り除き，実験操作の純粋な効果を計る事ができることが線形混合モデルの大きなメリットである．

[8] 相関パラメータとは，ランダム効果の構成要素間の相関関係を説明するパラメータである．たとえば読み時間計測実験において，被験者ランダム効果内の，ある説明要因のランダムスロープと被験者ランダム効果の間に正の相関が存在する場合，前者が説明する説明要因の効果の大きさにおけるランダムな個人差と，後者が説明する被験者の読み時間におけるランダムな個人差とは別に，読み時間の遅い被験者においては，説明要因の効果が大きかったという 2 者間のランダムな（あくまで実験をする前には予測不可能であるという意味で）相関関係を意味する．

[9] 仮説に方向性がある場合には t 値に対して片側検定の基準を用いることも理論上可能だが，基本的に，片側に起こる差のみ意味があると確信できる場合を除いて片側検定は避けるべきであり，通常両側検定を採用する．また一般化線形混合モデルにおいては最尤法により z 値が算出されるが，この場合には，正規分布に基づいて z 値が 1.96 かそれ以上で有意と判定できる（lme4 パッケージの一般化線形混合モデルでは正規分布に基づいて p 値が算出される）．

[10] 線形混合モデルおよび一般化線形モデルに対して自由度を近似する方法にはいくつかあるが，これらの近似法全てには，うまく機能しない反例が存在するようだと警告されていることを述べておく（前述したウェブサイト（http://glmm.wikidot.com/faq）を参照）．

[11] Dawson は以下のホームページで 2 要因及び 3 要因の交互作用のパターンを線形モデルの切片と係数から直接計算し図示できるエクセルのワークシートを公開している．コーディングが標準化されていない場合，そのコーディングの平均値，標準偏差を合わせて入力することで各セルの平均値を計算できる．
http://www.jeremydawson.co.uk/slopes.htm

[12] R 上で factor 関数を使って説明変数を要因型に変換すると 0 と 1 のダミーコーディングが適用される．

参 考 文 献

Agresti, A. (2002). *Categorical Data Analysis*, 2nd ed., John Wiley & Sons, New York.

Aiken, L. S. and West, S. G. (1991). *Multiple Regression*: *Testing and Interpreting Interactions*, Sage, Newbury Park, London.

Altmann, G. T. and Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference, *Cognition*, **73**, 247–264.

Arai, M. and Nakamura, C. (2016). It's harder to break a relationship when you commit long, *PLoS ONE*, **11**, e0156482, doi:10.1371/journal.pone.0156482.

Arai, M., van Gompel, R. P. G. and Scheepers, C. (2007). Priming ditransitive structures in comprehension, *Cognitive Psychology*, **54**, 218–250.

Arai, M., Nakamura, C. and Mazuka, R. (2015). Predicting the unbeaten path through syntactic priming, *Journal of Experimental Psychology*: *Learning, Memory and Cognition*, **41**, 482–500.

Baayen, R. H. (2008). *Analyzing Linguistic Data*: *A Practical Introduction to Statistics Using R*, Cambridge University Press, Cambridge.

Baayen, R. H. and Milin, P. (2010). Analyzing reaction times, *International Journal of Psychological Research*, **3**, 12–28.

Baayen, R. H., Davidson, D. J. and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items, *Journal of Memory and Language*, **59**, 390–412.

Balota, D. A., Aschenbrenner, A. J. and Yap, M. J. (2013). Additive effects of word frequency and stimulus quality: The influence of trial history and data transformations, *Journal of Experimental Psychology*: *Learning, Memory and Cognition*, **39**, 1563–1571.

Barr, D. J. (2008). Analyzing "visual world" eyetracking data using multilevel logistic regression, *Journal of Memory and Language*, **59**, 457–474.

Barr, D. J. and Frank, A. F. (2009). Analyzing multinomial and time-series data, Workshop on Ordinary and Multilevel Modeling at 2009 CUNY Conference on Sentence Processing, University of California, Davis. Slides available at https://www.hlp.rochester.edu/resources/WOMM/BarrFrank.pdf.

Barr, D. J., Levy, R., Scheepers, C. and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal, *Journal of Memory and Language*, **68**, 255–278.

Bates, D. M. (2005). Fitting linear mixed models in R: Using the lme4 package, *R News*: *The Newsletter of the R Project*, **5**, 27–30.

Bates, D. M., Kliegl, R., Vasishth, S. and Baayen, H. (2015). Parsimonious mixed models, arXiv:1506.04967, 1–27.

Belia, S., Fidler, F., Williams, J. and Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars, *Psychological Methods*, **10**, 389–396.

Box, G. E. P. (1976). Science and Statistics, *Journal of the American Statistical Association*, **71**, 791–799.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research, *Journal of Verbal Learning and Verbal Behavior*, **12**, 335–359.

Clifton C., Jr. (2013). Situational context affects definiteness preferences: Accommodation of presuppositions, *Journal of Experimental Psychology*: *Learning, Memory and Cognition*, **39**, 487–501.

Dawson, J. F. (2014). Moderation in management research: What, why, when, and how, *Journal of Business and Psychology*, **29**, 1–19.

Ferreira, F. and Clifton, C. J. (1986). The independence of syntactic processing, *Journal of Memory and Language*, **25**, 348–368.

Fine, A., Jaeger, F., Farmer, T. and Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension, *PloS One*, **8**, doi:10.1371/journal.pone.0077661.

Frank, S. L., Monsalve, I. F., Thompson, R. L. and Vigliocco, G. (2013). Reading-time data for evaluating broad-coverage models of English sentence processing, *Behavior Research Methods*, **45**, 1182–1190.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*,

Cambridge University Press, Cambridge.
橋本健一 (2010). 反応時間計測実験における外れ値の取扱い—L2 心理言語実験の場合—, より良い外国語教育研究のための方法, 外国語教育メディア学会(LET)関西支部メソドロジー研究部会 2010 年度報告論集, 133–145.
Heider, P. M., Dery, J. E. and Roland, D. (2014). The processing of it object relative clauses: Evidence against a fine-grained frequency account, *Journal of Memory and Language*, **75**, 58–76.
Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models, *Journal of Memory and Language*, **59**, 434–446.
Juffs, A. (1998). Main verb versus reduced relative clause ambiguity resolution in L2 sentence processing, *Language Learning*, **48**, 107–147.
Juffs, A. (2005). The influence of first language on the processing of wh-movement in English as a second language, *Second Language Research*, **21**, 121–151.
Just, M. A., Carpenter, P. A. and Woolley, J. D. (1982). Paradigms and processes and in reading comprehension, *Journal of Experimental Psychology*: *General*, **3**, 228–238.
Kamide, Y. (2012). Learning individual talkers' structural preferences, *Cognition*, **124**, 66–71.
Kamide, Y., Altmann, G. T. M. and Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements, *Journal of Memory and Language*, **49**, 133–156.
久保拓弥 (2012).『データ解析のための統計モデリング入門：一般化線形モデル・階層ベイズモデル・MCMC』, 岩波書店, 東京.
Lo, S. and Andrews, S. (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data, *Frontiers in Psychology*, **6**, 1–16.
Maris, E. (2012). Statistical testing in electrophysiological studies, *Psychophysiology*, **49**, 549–565.
Matin, E., Shao, K. C. and Boff, K. R. (1993). Saccadic overhead: Information processing time with and without saccades, *Perception & Psychophysics*, **53**, 372–380.
Mazuka, R., Ito, K. and Kondo, T. (2002). Costs of scrambling in Japanese sentence processing, *Sentence Processing in East Asian Languages* (ed. M. Nakayama), 131–166, CSLI Publications, Stanford, California.
McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*, Chapman and Hall, London.
Mirman, D., Dixon, J. A. and Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences, *Journal of Memory and Language*, **59**, 475–494.
Nakamura, C. and Arai, M. (2016). Persistence of initial misanalysis with no referential ambiguity, *Cognitive Science*, **40**, 909–940.
Nakamura, C., Arai, M. and Mazuka, R. (2012). Immediate use of prosody and context in predicting a syntactic structure, *Cognition*, **125**, 317–323.
Raaijmakers, J. G. W., Schrijnemakers, J. M. C. and Gremmen, F. (1999). How to deal with "The Language-as-Fixed-Effect Fallacy": Common misconceptions and alternative solutions, *Journal of Memory and Language*, **426**, 416–426.
Ratcliff, R. (1993). Methods for dealing with reaction time outliers, *Psychological Bulletin*, **114**, 510–532.
Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research, *Psychological Bulletin*, **124**, 372–422.
Rayner, K. and Pollatsek, A. (1989). *The Psychology of Reading*, Lawrence Erlbaum Associates, Hillsdale, New Jersey.
Roland, D. (2009). Relative clauses remodeled: The problem with mixed effect models, Poster presentation at the 2009 CUNY sentence processing conference, University of California, Davis.

Roland, D., Mauner, G., O'Meara, C. and Yun, H. (2012). Discourse expectations and relative clause processing, *Journal of Memory and Language*, **66**, 479–508.

清水裕士 (2014).『個人と集団のマルチレベル分析』, ナカニシヤ出版, 京都.

Sturt, P. (2007). Semantic re-interpretation and garden path recovery, *Cognition*, **105**, 477–488.

Sturt, P., Keller, F. and Dubey, A. (2010). Syntactic priming in comprehension: Parallelism effects with and without coordination, *Journal of Memory and Language*, **62**, 333–351.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M. and Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension, *Science*, **268**, 1632–1634.

Townsend, J. T. (1992). On the proper scale for reaction time, *Cognition, Information Processing and Psychophysics*: *Basic Issues* (eds. H. Geissler, S. Link and J. T. Townsend), Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Proceedings of the Institute of Statistical Mathematics Vol. 64, No. 2, 201–231 (2016) 231

Statistical Analysis of Eye-movement Data and Reading Time Data in Language Comprehension Research

Manabu Arai[1] and Douglas Roland[2]

[1]Faculty of Economics, Seijo University
[2]Graduate School of Arts and Sciences, The University of Tokyo

Research on language comprehension has made significant advances over the last 30 years or so largely owing to technological advances that have enabled researchers to conduct chronometric studies with little required cost. Furthermore, eye-tracking devices, which have played an important role in advancing the research on language comprehension, were once only available to well-funded laboratories, but are now within many researchers' reach. Although the collection of time-encoded data is easier than ever, appropriate handling of such data often requires not-so-straightforward statistical modeling. In this paper, we discuss statistical methods for analyzing eye-movement data from visual world and reading studies as well as reading times from the self-paced reading task. We argue that careful and reasonable application of Linear Mixed-Effects models as well as Generalized Mixed-Effects models can offer great advantages in many ways over traditional analyses such as ANOVA that require data aggregation over participants or items.

Key words: Linear mixed effects models, generalized linear mixed effects models, eye-movements, visual world paradigm, self-paced reading, reading times.

統計数理(2016)
第 64 巻 第 2 号 233–246

特集「統計的言語研究の現在」
［研究ノート］

ツイート数と現実の統計量との差異に関する検討

荒牧 英治†・若宮 翔子†

(受付 2015 年 12 月 31 日；改訂 2016 年 11 月 10 日；採択 11 月 21 日)

要　　旨

ソーシャルメディアサービスの普及により，人々や社会の状況を調査する新たなアプローチが開拓された．この結果，インフルエンザや地震などを対象とした多くのサーベイランスや監視システムが提案され，現在も稼働している．しかし，ソーシャルメディア上のユーザ発信データ(発言内容，時間や場所)が必ずしも現実を正確に反映しているとは限らない．例えば，デマや流言などが出現することもあり，新聞などの既存のメディアと比べて，内容の信頼性は十分ではなく，時間的または空間的な正確性にも限界がある．本稿では，ソーシャルメディアを代表する Twitter を用いて構築したインフルエンザ・サーベイランス・システムを例に，ツイート数と現実の統計量の時間的なずれと空間的なずれについて検討し，背後にあるバイアスについて議論する．

キーワード：ソーシャルメディア，Twitter，自然言語処理，ソーシャル・コンピューティング，インフルエンザ．

1. はじめに

近年の情報処理の発展は World Wide Web(以降，Web)の存在なしに語ることはできない．Web はかつてないほどの巨大なデータを含み，かつ，誰もが発信できるメディアである．この特性を活かし，Web ならではの新たな研究分野も形成されている．評判情報抽出 Pang et al. (2002) がその代表例である．さらに，近年では，Twitter や Facebook などのソーシャルメディアが爆発的に普及し，ここから，評判だけでなく，より詳細な情報を抽出する試みにも関心が集まっている．ソーシャルメディアのデータは，大規模かつ即時的であり，一部には位置情報も付与されているなど，これまでの Web データには見られなかった特徴がある．この特徴を利用し，疾病サーベイランス(Aramaki et al., 2011; Paul and Dredze, 2011; 谷田 他, 2011)，地震検知(Sakaki et al., 2010)，選挙結果予測(Tumasjan et al., 2010)や株価予測(Bollen et al., 2011)など，その応用領域は多岐にわたっている．

これらの研究は，暗に Web テキストが現実に対応しているという仮定がベースとなっている．しかし，Web テキストは必ずしも現実と正確に対応しているわけではなく，様々なバイアスから両者の間にギャップが生じる場合がある．最も大きなバイアスの一つは，SNS ユーザの偏りによるものである．例えば，人口 1 万人あたりの Twitter ユーザは，東京都では 369.82 人であるのに対し，最も少ない佐賀県では 63.67 人であり，約 6 倍もの差がある(odomon.net, 2013)．さらに，18 歳から 24 歳のユーザが多い．これらを考えると，都市部の若者を中心にデータを採取していることになる．これ以外にも，問題となりうるバイアスは多く存在し，現

† 奈良先端科学技術大学院大学：〒630–0192 奈良県生駒市高山町 8916–5

実とソーシャルメディア・データとの定量的な差異を生み出している．

本研究では，ソーシャルデータを実世界の現象の「センサ」として用いる先行研究に対し，現実とソーシャルデータに存在する時間的，空間的なギャップを補正し，より高い精度で実世界の現象をモデル化することを目指す．本稿では，感染症サーベイランスと呼ばれる感染症流行の把握のために，Twitter のようなユーザ投稿発言データの利活用が進みつつある（国立研究開発法人日本医療研究開発機構（AMED），2015）という背景を受け，主な感染症の一つであるインフルエンザを対象に，代表的なソーシャルメディアである Twitter を用いて構築したサーベイランスシステムを実例として，時間的なギャップ（2 章）と空間的なギャップ（3 章）について議論し，これらのギャップを補正した方法をインフルエンザの患者推定という事例（4 章）を挙げて考察する．5 章で関連研究を紹介し，6 章でまとめを述べる．

2. 時間的ギャップ

本章では，ソーシャルメディアと現実がどのような時間的なギャップを持ちうるのか，時系列データであるインフルエンザの流行を題材に議論を行う．

2.1 材料: インフル・コーパス

インフルエンザに関する Twitter 上での発言を集めたコーパス（以下，インフル・コーパス）を用いて，時間的なずれの調査を行った．インフル・コーパスはインフルエンザ・サーベイランスサイト（奈良先端科学技術大学院大学ソーシャル・コンピューティング研究室, 2016）を稼働し，収集されたデータをもとにしたコーパスであり，以下の手順で構築されている．

まず，2008 年 11 月から 2010 年 7 月にかけて Twitter API を用いて 30 億発言を収集し，そこから「インフル」を含む発言（インフル関連発言）を 10,443 件無作為に抽出した．これに対して作業者が，発言者がインフルエンザ罹患者（正例）であるか否か（負例）という事実性を判定し，ラベル付けした．アノテーションについては，以下のような基準に照らし，一つでも該当するものがあれば，負例とみなした．より詳細な基準に関しては，アノテーション・ガイドライン（Aramaki and Wakamiya, 2016）を参照していただきたい．

(1) 発言者または発言者と距離的に近い人物（同一都道府県近郊の人間）の疾患でない場合
(2) 現在または近い過去（24 時間以内）の疾患でない場合
(3) 否定，疑問や「かもしれない」といった事実でない場合

このトレーニング・データをもとに，2011 年 8 月からインフルエンザ・サーベイランスサイトを運用し，現在までに，8,129,571 件のインフル関連発言を抽出し，インフルエンザ罹患者（正例）による発言であるか否か（負例）という事実性に関するラベルを自動推定した．なお，全ての単語の表層系を素性とし，SVM を用いて構築した分類器を用いた．

この結果，テストデータの 58% が正例と判定された．F 値を求めたところ，0.76 という高い精度を示した（Aramaki et al., 2011）．

2.2 結果

2012 年から以降 3 年間のシーズン時の発言データから，自動判別で得られた正例（以降，単に正例と表す）の発言データを抽出し，インフルエンザの患者数を下記の式により推定した．

$$I_0(a,t) = \bar{I_0} \cdot \frac{M(a,t)}{N(a)} \tag{2.1}$$

ここで，$M(a,t)$ は対象地域 a における特定の日 t のインフルエンザの正例数，$N(a)$ は対象地

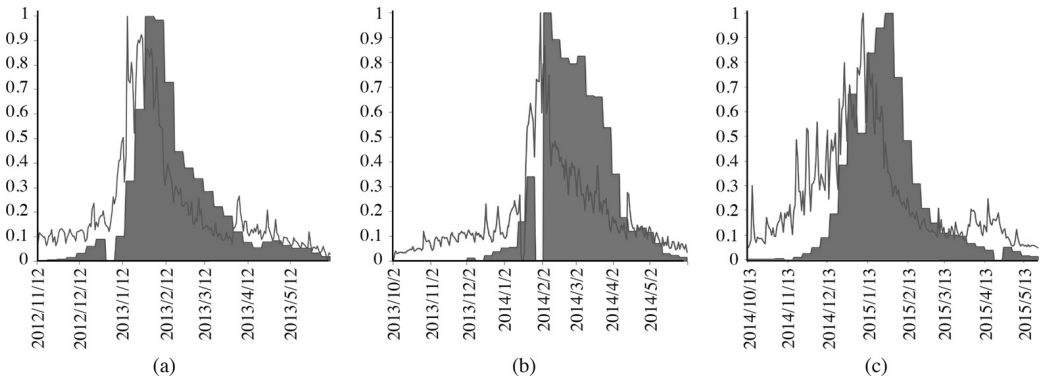


図 1. 2012–2014 年度のインフルエンザの患者数と正例の発言を用いて推定した患者数．横軸は日付 d，縦軸は患者数(塗り潰し)と Twitter 発言数(折れ線)とする．それぞれの値は，年度の最大値を 1 として，0 から 1 の値になるように正規化して表示している．

域 a におけるソーシャルセンサ数(ユーザ数)，$\bar{I}$ はスケールパラメータである．なお，本研究では，発言数はユーザ数に比例するものとみなしている．そのため，ソーシャルセンサ数は，各地域における対象期間中の平均発言数に基づき算出される．結果を図 1 に示す．2011 年については，Twitter API の仕様変更に伴い，一時期クロールが停止していたため，本研究では対象外とした．

正解データのインフルエンザの患者数として，国立感染症研究所 感染症情報センターが報告している患者数を用いた．これは，国立感染症研究所のホームページにおける患者発生状況の「都道府県別報告数・定点当たり報告数」より取得可能であり，秋から春のインフルエンザシーズンにかけて，都道府県ごとのデータが毎週公開される．発言数より推定された患者数と実際の患者数との差分を誤差として調査を行った．誤差を算出する際，年度ごとの最大値が 1 となるようにそれぞれの値を正規化している．誤差は正規化した推定患者数の値から正規化した実際の患者数の値を引いた値であり，そのスケールは −1 から 1 である．正の誤差は，推定患者数が実際の患者数を上回っていることを意味し，反対に負の誤差は，推定患者数が実際の患者数を下回っていることを意味する．なお，この誤差の時間ごとの和が小さくなるように，発言数から患者数を推定するときのパラメータ $\bar{I}_0$ を調整した．

一般的に，感染症の把握／予測は，推定した値と患者数との相関係数が評価基準であり，ここでいう誤差の最小化は，間接的にこの評価をよくすることができる．

いずれの年においても，実際の流行のピーク前に，Web 上でのインフルエンザ発言が増加し，逆に，ピーク後は実際の流行度合いよりも発言が減少する，という傾向が確認された．対象とした 3 年間の平均をとった結果を図 2(a)に示す．

全体を平均すると，3 年間の平均誤差は −0.0178(標準偏差 0.188)となり，実際の患者数がソーシャルメディアの発言数を上回っているが，ピーク前，ピーク直後，平常時と 3 つの異なる状態があることが分かる．

まず，ピーク前は平均誤差が 0.136(標準偏差 0.062)となり，ソーシャルメディア上での発言数が実際の患者数を上回り続ける．標準偏差が小さいことから，この誤差は安定しており，ソーシャルメディア上では現実より常に加熱した状態となっているといえる(図 2(b))．

次にピーク後は，平均誤差が −0.256(標準偏差 0.122)となり，今度は逆に，実際の患者数がソーシャルメディアの発言数を上回っている．標準偏差が大きいことから分かるように，この

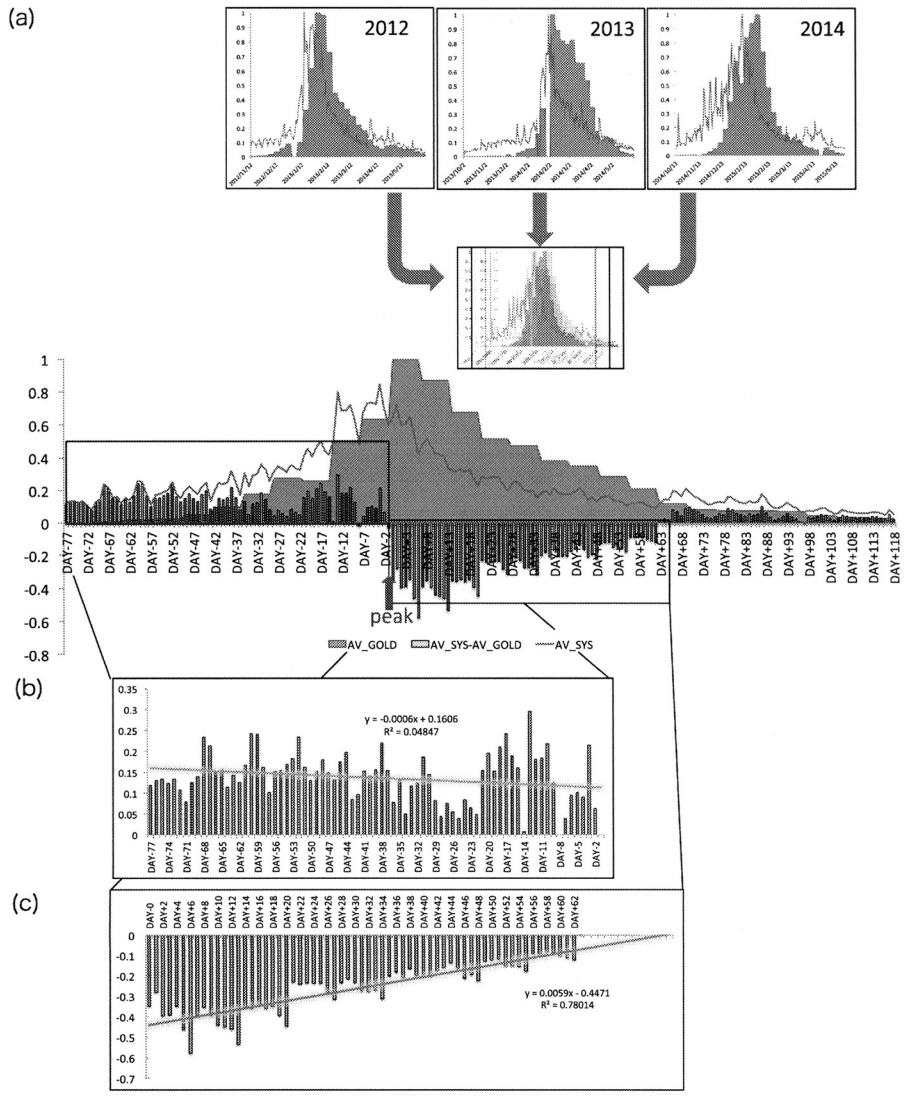


図 2. 各年度の患者数がピークとなる日時を合わせた平均．横軸は患者数のピーク日時からの相対的な日付，縦軸は，以下で定義される患者数の平均(塗り潰し)，Twitter 発言数の平均(折れ線)，および両者の平均誤差(棒)を示す．

上回り方は時期によって異なる．ピーク直後は最も誤差が大きく，0.40 に近い．この誤差は次第に減少し，最後には 0.10 付近となる(図 2(c))．

最後に，平常時は，ピーク前と同様に，ソーシャルメディア上での発言数が患者数を上回る状態が安定して続く．

2.3 考察

この結果から，次の 2 つの知見が得られた．まず，誤差は常に存在しており，(1)ピーク前，(2)ピーク直後，(3)平常時の 3 つの段階がある．(1)ピーク前と(3)平常時における誤差は似て

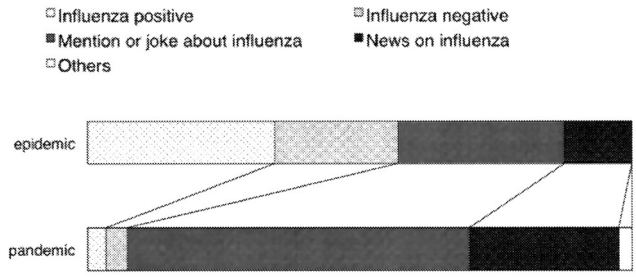


図 3. インフルエンザに関連する発言の分類．“Influenza positive” はインフルエンザ患者の発言．“Influenza negative” は「インフルエンザでなくてよかった」など否定を含む発言．epidemic（2011 年 11 月）および pandemic（2009 年 1 月）より，それぞれ 200 発言を無作為抽出し，人手で分類した．

おり，発言数が実際の患者数を上回る．一方，(2)ピーク直後では，発言数よりも患者数が多くなり，徐々に誤差が減少し，(3)平常時へと収束する．この性質を利用すれば，ピークを知ることにより，流行推定の精度を向上させることができる．

次に，ソーシャルメディア上での発言数のピークがいつであるかを事前に知ることは困難である．(1)ピーク前はほぼ単調な誤差が続くばかりであり（図 2(b)），いつピークを迎えるのかを伺わせる手がかりはない．つまり，予測を行うことは難しい．

この 2 つの知見，すなわち (1)誤差はピーク前には実際よりも発言が多く，ピーク直後には発言が少なくなること，(2)ピークがいつであるかを知ることは困難であること，はインフルエンザ調査の実用的応用を困難にしている．

このような発言量の減少が生じる原因について考察する．図 3 は，その一端を示す例であり，平常時（epidemic, 2011 年 11 月）と WHO が新型インフルエンザへの懸念を表明した時期（pandemic, 2009 年 1 月）における発言の内訳を示している．epidemic 期には，「インフル」を話題にした発言やニュースに対する発言が増加する．高まる不安や対応法への懸念として「インフル」という単語が使用されるといえる．このような epidemic 期を過ぎると，すでに多くのユーザは，大量のニュースで「インフル」を目にしていることになる．このようにインフルエンザ流行がすでに周知され一般に既知となった状態では，自分（または周囲）にインフルエンザが発生したとしても，話題としての価値が低下しているため，ソーシャルメディア上で発言することを躊躇すると考えられる．これは，図 1 の Twitter 発言数の推移にも表れており，患者数がピークを迎えると発言数は大きく減少しており，情報伝搬のタイミングがギャップを生み出している可能性がある．

3. 空間的ギャップ

本章では，ソーシャルメディアと現実がどのような空間的なギャップを持ちうるのか，位置情報が付与されており，かつ，ランドマーク表現が含まれる発言を用いて調査する．なお，ランドマーク表現が含まれる場合には，次の可能性がある．

(1) ランドマークが地名を示す
- ランドマーク上にいる
- ランドマーク上にいない

(2) 地名を示さない

先行研究においては，(1)と(2)を自動分類するアプローチ(Awamura et al., 2015) もあるが，本研究では(Antoine et al., 2015)と同様に，これらを区別せずに，どのような場所が言及されるか調査し，議論する．

3.1 材料: 京都 GPS コーパス

空間的なずれを調査するために，京都を対象地域として，京都市近郊で発信された GPS 情報付き発言(以下，**京都 GPS コーパス**)を用いた．京都 GPS コーパスは，Twitter API を用いて約 1 年間(2011 年 7 月 15 日から 2012 年 7 月 31 日)にわたり収集した約 3.7 万件発言から構成されている．

次に京都市近郊の場所として，以下の 4 つのタイプの 6 つのランドマークを選択した．(1)広域領域内に複数の施設を持つ広域複合ランドマーク((a)同志社大学，(b)京都大学)，(2)局所的な特定の施設からなる狭域単一ランドマーク((c)河原町駅，(d)四条駅)，(3)広域領域だが単独の施設からなる広域単一ランドマーク((e)京都府立植物園)，(4)境界を持たないランドマーク((f)吉田)．図 4 に，これら 4 つのタイプを代表する 6 つのランドマークとそれらを含んだ発言をそれぞれの位置情報に基づき地図上にマッピングした結果を示す．

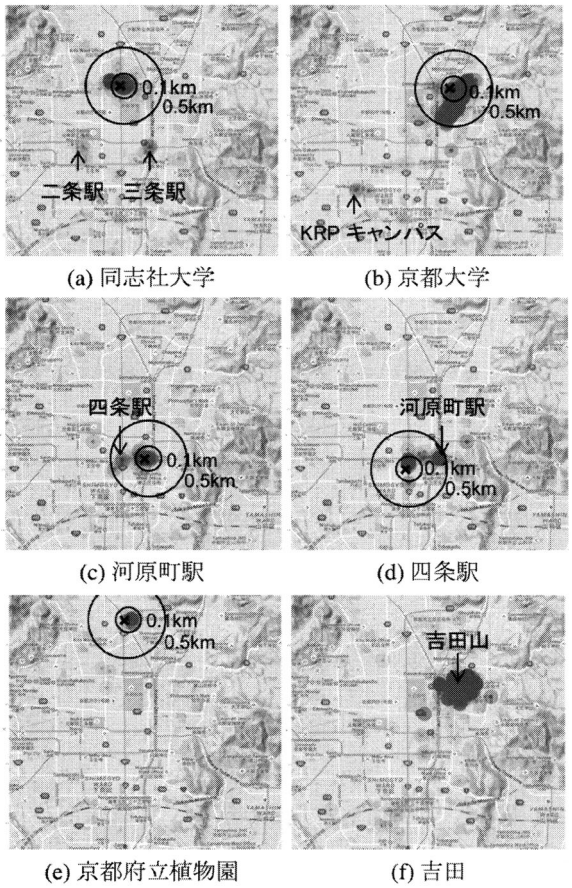


図 4. 6 つのランドマークに関する発言の地理的分布．× 印はランドマークの場所であり，2 つの円はそれぞれランドマークの場所から半径 0.1 km, 0.5 km の領域を示す．

3.2 結果

図 4 より，多くの場合，特定のランドマークに関する発言はそのランドマークの場所から 0.1 km 以内に集中しており，自分が滞在している場所について，言及していることを伺わせる．図 5 は，京都大学，四条駅，京都府立植物園に関する発言の距離ごとの密度である．多少異なった場所においても小さなピークが出現していることが分かる．

タイプ別に見れば，大学については，(a)「同志社大学」は，広域領域内に複数の施設を持つが，そのほとんどの発言は，「同志社」周辺に集中し，顕著なずれは見られない(図 4(a))．ただし，二条駅や三条駅といった周辺の駅にもピークが見られる．一方，同じく広域領域内に複数の施設を持つ(b)「京都大学」に関しては，図 4(b)のように京都大学の別キャンパスである KRP(京都リサーチパーク)キャンパスや最寄り駅である出町柳駅周辺での発言が多く見られる．このように，これからその場所へ向かう，または，離れる際に発言が行われ，実際の位置よりずれる現象が見られる．

次に駅を見てみる．(c)「河原町駅」は，この場所周辺での発言数が相対的に多いものの，四条駅付近についても発言が多い(図 4(c))．同様に，(d)「四条駅」に関しても，河原町駅周辺でも多くの発言がなされている．これは，河原町駅にいる人々が京都駅に向かう場合，四条駅を経由する行き方が一般的であるため，これから移動する場所，または，直前までいた場所について言及するという傾向が反映された結果であると考えられる．

最も集中して発言が見られたのは，上記の中でも京都大学に次ぐ大きな面積を持つ(e)「京都府立植物園」である(図 4(e))．広域であっても，発言箇所が入園時の入り口付近に集中している．

最後に (4)境界を持たないランドマーク(f)「吉田」の結果を示す．このランドマーク名は，「吉田(神社)」「吉田(山)」「吉田(寮)」など周辺の複数の地名を表し多義的であり，かつ，明確な境界を持たない語である．このため，「吉田山」を中心として発言がなされているものの，広域に発言が分散している(図 4(f))．

3.3 考察

特定の場所に関する場所参照発言がどこからなされているのかというパターンは場所ごとに依存するが，最寄り駅(同志社大学や京都大学)や隣接駅(河原町駅や四条駅)，出入り口付近(植物園)など，そのずれには一定の傾向があることが多い．これを，あらかじめ知ることができれば，GPS ベースの位置情報が付加されていない発言であっても，発言内容をもとに位置情報を推定することが可能になる．例えば，広域であっても施設が単一あるいは少数であれば，京都府立植物園のように出入り口付近にいると推定可能な場合がある．逆に，遠くの場所から言及される場合は，これから移動する予定，または，移動経路について言及しているなどのパ

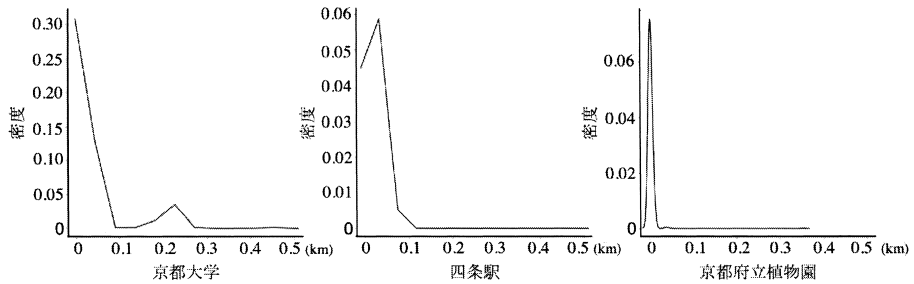


図 5. 3 つのランドマークに関する発言の距離ごとの密度.

ターンが見られる場合がある．この場合，どのような理由で遠くから言及するかを推察することで，より高い推定が可能になると考えられる．このようなずれは，発言を前後の文脈を考慮して自然言語処理により解析することにより，分離できる可能性がある．

4. 時間的・空間的なギャップの補正：インフルエンザ患者数推定を事例として

ここまで時間的なギャップ(2 章)と空間的なギャップ(3 章)について述べてきた．2 章では，ピークを過ぎると発言量が低下すること，その原因としては，ピーク後の話題としての価値低下が考えられることについて述べた．3 章では，空間的なギャップの原因にはある種の定型性(例えば，移動経路や最寄り駅)があることを示した．本章では，これらを考慮しつつ，インフルエンザ推定精度の向上を試みる．

図 6 に典型的なインフルエンザ流行の推移(2013 年北海道)を示す．感染症情報センターの値とソーシャルメディアから得られた 2 つの値が描かれている．ソーシャルメディアに基づく値は，北海道内でのインフルエンザに関する発言と，空間的なギャップを含む発言(以降，**遠距離言及発言**)により算出される．ここでいう遠距離言及発言とは，北海道以外の地域から北海道のインフルエンザについて言及している(ここでは，単語「北海道」を含む)発言とする．

前者は，ピーク後に実際の患者数に相対して大幅に減少し，2 章で述べた通りの時間的なギャップを示している．後者(遠距離言及発言)は，数は少ないものの，数回程度のバースト(急峻な盛り上がり)が存在し，特にピーク時に大きな盛り上がりを見せている．また，ピーク後には大きなバーストはない．

ピーク前にバーストが存在する理由について考察する．遠距離言及発言は，自分がいない離れた場所でのインフルエンザに関して言及しており，実際は，(1)ユーザ本人が移動した結果，遠隔地への発言となる，または，(2)ニュース・メディアなどの二次情報を通して遠隔地のインフルエンザ流行を知り，それについて発言する場合が考えられる．(1)の例として，「新型インフルエンザかぁ…北海道のライヴに遠征にいって家に帰ったら発症したんだよなぁ…絶対あのライヴ会場に発症者居たよなｗ」(千葉県より発信)，「北海道にてインフルエンザになった。

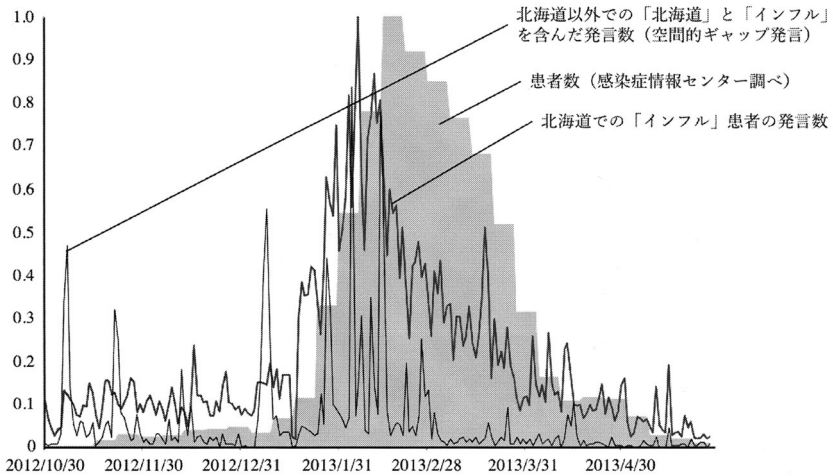


図 6. X 軸は日付．Y 軸は，2013 年北海道におけるインフルエンザ関連の発言量，遠距離言及発言，感染症情報センターの報告を示す．なお，シーズン内の最大値を 1.0 として各値を正規化している．

まじ、ありえないわ」(東京より発信)といった発言があった．しかし，大勢の人間が同時期に移動することは考えにくいため，(1)は通常はバーストの原因となりにくい．

すなわち，インフルエンザ流行のピーク前，またはピーク中に，ニュース・メディアで取り上げられたことをきっかけに，(2)のニュースの引用(例えば，「【■■ニュース】北海道で例年より早くインフルエンザ大流行」)や，流行場所を懸念する発言(例えば，「@■■ いわく北海道でインフルエンザ蔓延中」)などが大量に発生すると考えられ，バーストを形成することになる．なお，北海道以外から北海道のインフルエンザに関する発言を 412 件取得し，ニュース・メディアなどの二次情報を通して知り，発言していると思われる発言を人手で調査した．その結果，(1)ユーザ本人が移動した結果，遠隔地への発言は約 10%，(2)ニュース・メディア，あるいは，北海道に住む家族や知人などからの情報に対して発言は約 23% であった．残りの発言の大半は，北海道とインフルエンザについてそれぞれ別の文脈で述べているものや，北海道のインフルエンザについて述べているが負例のものであった．さらに，北海道におけるインフルエンザ流行のピーク中(2014 年 2 月 15 日)に，北海道以外から北海道のインフルエンザに関して発信された 100 件の発言を調査した．その結果，約 9 割の発言が(2)ニュース・メディア，あるいは，北海道に住む家族や知人などからの情報に対しての発言であった．

このように，遠距離言及発言がネット上でのある種の注目度を示しているとすると，注目を浴びるにつれ，話題としての価値が下がり，関連する発言が減少するというモデルを考えることができる．

4.1 ソーシャルセンサの劣化モデル

遠距離言及発言によって，インフルエンザ関連発言が減少していく過程を以下のようにモデル化する．

ピーク前：インフルエンザ流行前にインフルエンザに罹患したユーザは，インフルエンザ関連発言を行う．

ここで，N 人の Twitter ユーザによる発言が T 個あるとき，発言数 T はユーザ数 N にそのまま比例するものとみなすと，T 個のソーシャルセンサが機能しているといえる．なお，発言に付与されているユーザ ID を参照すればユーザ数を求めることも可能ではあるが，リアルタイム性を重視したシステムへの実装には向いていない．そのため，発言数をそのままユーザ数と比例するものとみなしている．

ピーク中：インフルエンザが流行し始めると，ニュースなど，遠く離れた場所からでも対象地域のインフルエンザについての言及(遠距離言及発言)が増える(この数を IRT_{gap} とみなす)．この遠距離言及発言の量に応じて，話題としての価値が低減し，インフルエンザ関連発言を行わないユーザが増えるとみなす．このようなユーザはソーシャルセンサとして機能しないため，本稿では，**劣化ソーシャルセンサ**と呼ぶ．劣化ソーシャルセンサの数は，スケールパラメータ W を用いて，遠距離言及発言数を log で鈍らせた $W \cdot \log(IRT_{gap}+1)$ とする．なお，発言数(話題の度合い)は，対象地域によっては極端に大きな発言数となることもある非常に偏った分布(べき分布に近い)になる可能性があるため，べき分布を扱う際に一般的な log による対数をとっている．

このように考えると，対象地域 a の特定の日 t における患者数($I_1(a,t)$)(なお，単位は数ではなく対数とする)は以下のようにモデル化できる．

$$I_1(a,t) = \bar{I_1} \cdot \frac{M(a,t)}{N(a) - \bar{I} \cdot \log(G(a,t)+1)} \tag{4.1}$$

ここで，$M(a,t)$ は対象地域 a における任意の日 t におけるインフルエンザ関連発言数，

$G(a,t)$ は対象地域外からの対象地域名(県名)を含むインフルエンザ関連発言数，$N(a)$ は対象地域 a のソーシャルセンサ数(式(2.1)と同様に，対象地域における対象期間中の平均発言数)，$\log(G(a,t)+1)$ は話題の度合いを表す．なお，$G(a,t)$ は任意の日 t における対象地域 a に関するニュースや RT の発言量であり，インフルエンザ患者が発信するツイート数よりも指数的に増加すると考えられる．そのため，対数を用いてこのような発言による影響を抑えている．また，$\bar{I}_1$ と $\bar{I}$ はそれぞれスケールパラメータである．実験では，評価データにより患者数との誤差が最小となるようにフィッティングを行い，$\bar{I}_1$ は 1，$\bar{I}$ は 20 とした．なお，今回はツイート数と現実の統計量との差異をなるべく小さくするようなアプローチに焦点を当てており，パラメータやパフォーマンスの最適化についての検討は，今後の課題である．

このモデルは，ピーク前とピーク後などピークのタイミングを必要とせず，遠距離言及発言の数により，ピーク前後 2 つの状態を再現できる．すなわち，遠距離言及発言の数により，徐々にソーシャルセンサの数が低下し，この結果，インフルエンザ関連発言数の価値が相対的に高まる．

4.2 結果

提案手法(PROPOSED)の精度を測るために，感染症情報センターの報告をゴールドスタンダードとして，相関係数を求めた．比較のために，先行研究(EMNLP2011)(Aramaki et al., 2011)との相関係数を用いた．なお，先行研究(EMNLP2011)における患者数(I_0)推定式は式(2.1)の通りである．

2012–2014 年の結果を表 1 に示す．この結果に示されるように，全シーズンを通して PROPOSED の相関係数が EMNLP2011 の相関係数を上回っていることが分かる．都道府県別での相関係数を図 7 に示す．ほぼすべての地域で提案モデルの方が EMNLP2011 よりも相関係数が高く，精度が向上している．

提案モデルは，単純なものであるが，「話題としての価値が低くなると，発言量が減る」という人間の性質を取り込むだけで，2 章で述べたピーク後の発言量が実際よりも小さくなるという誤差を補正できる．このことは，今後，Twitter などのソーシャルメディアデータをより深く活用する際の重要な知見であると考える．

表 1．年度別の感染症情報センターの報告との相関係数の比較．

Method	2012	2013	2014	TOTAL
PROPOSED（劣化ソーシャルセンサモデル）	0.79	0.73	0.73	0.74
EMNLP2011	0.74	0.68	0.67	0.69

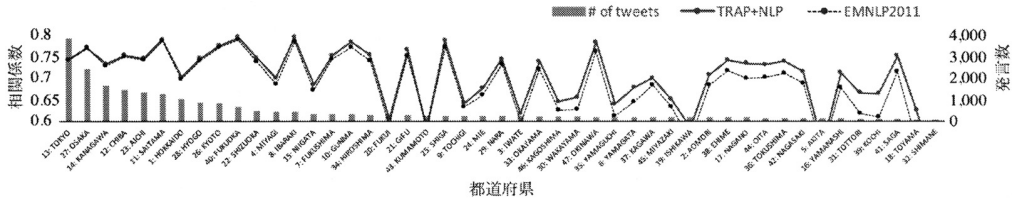


図 7．地域別の感染症情報センターの報告との相関係数の比較．X 軸は都道府県(並び順は各地域の Twitter 発言数(棒グラフ)に基づく)．Y 軸は，感染症情報センターの報告との相関係数．実線は提案手法(PROPOSED)，破線は先行研究(EMNLP2011)を示す．

5. 関連研究

Twitter をはじめとしたソーシャルメディアの普及により，新しくいくつかの研究が始まった，その代表的な例として 2 つの研究を示す．一つは，ソーシャルメディアの非文法的でノイジーな文章に対して処理の頑健性を高める研究である．この結果，Web 上にあるような非文法的な文章に対して，固有表現の抽出や単語の正規化を行う研究が行われてきた（Chrupała, 2014; Han and Baldwin, 2011; Plank et al., 2014）．

もう一つは，ソーシャルメディアから現実の世の中の知識を抽出する研究である．先の研究が自然言語処理に対する新たな課題であるとすると，こちらは新たな応用であるといえる．この結果，ソーシャルメディアからの意見抽出（O'Connor et al., 2010），イベント抽出（Li et al., 2014a; Marchetti-Bowick and Chambers, 2012; Sakaki et al., 2010; Shen et al., 2013; Thelwall et al., 2011），ユーザ行動分析（Bergsma et al., 2013; Han et al., 2013; Li et al., 2014b; Zhou et al., 2014），災害対応（Varga et al., 2013），世界知識抽出（Williams and Katz, 2012）など，様々なアプリケーションが提案されてきた．これらの応用例の中でも，疾患情報（特に，即時的な把握が必要とされる感染症）の流行検出に関しては，主要な Twitter 利用法の一つとして多くの研究がある（Aramaki et al., 2011; Paul and Dredze, 2011; 谷田 他, 2011）．Paul and Dredze (2011) は，病名ラベルが付与された発言なしで，より幅広い病気に関する発言を抽出する手法を提案している．そのために，事前知識として病気について書かれた記事を利用して Ailment Topic Aspect モデル拡張を行っている．これに対し，本研究ではインフルエンザに対象を絞り，発言のみを用いてインフルエンザ罹患者の判定を行い，患者数を推定している点が異なる．谷田他 (2011) は，風邪の流行度合いを推測するために，発言を用いて風邪の流行と関連した単語の出現頻度を回帰分析している．そのための変数選択において，選択する単語同士の相関をもとに，風邪の流行の特徴を捉えた推測を可能としている．一方，本研究はインフルエンザを対象にして，インフルエンザ罹患者による発言であるか否かを判定しているため，単純にあらかじめ決めた単語（今回は「インフル」）を用いるだけでも，都道府県単位の患者数を高い精度で推定できることを示している．

このような研究が盛んに行われているのは，感染症は未だ百万人を越える患者を出しており，恒常的な対策が必要であること（国立感染症研究所, 2006），および，新型インフルエンザといった危機事象についても，危惧されているという現状があるからである（Ferguson et al., 2005）．このため，感染症流行の把握は**感染症サーベイランス**と呼ばれ，各国で膨大なコストをかけて調査・集計が行われている．本邦でも，2016 年度から，国立研究開発法人日本医療研究開発機構（AMED）にて研究班が立ち上がり，Twitter のようなユーザ投稿発言データの利活用が進みつつある（国立研究開発法人日本医療研究開発機構（AMED）, 2015）．これにともない，本稿にて Twitter を中心に述べたソーシャル・メディアと現実の差異に関する議論が今後より進むものと思われる．

6. おわりに

本研究では時間的，および空間的ずれについて示した．いずれにおいても，人間の記述する欲求の偏りにより，不正確さを生んでいると考えられる．本研究で強調したいのは，ソーシャルメディアからの情報抽出に関する研究は，人間をセンサとみなしている（Sakaki et al., 2010）ものの，それはムラの多いセンサであることである．これを使いこなすためには，センサとしての人間の性質を十分に理解し，解析する必要がある．時間的には人々のピーク前の関心の加熱により，空間的には個々の地名の特徴（広さ，施設の個数）により，複雑な現象が生じ，ずれを生んでいる．これらを説明するためには，今後，Web において発言する人間の心理を真の対

象として解析することが必要となる可能性がある．同時に，人間の心理については，これまで心理学や社会学の分野で多くの知見が集積されているが，今後ソーシャルメディアの解析において，これら人間の心理を解析した知見との融合が必要になると考えられる．

参 考 文 献

Antoine, Émilien, Jatowt, Adam, Wakamiya, Shoko, Kawai, Yukiko and Akiyama, Toyokazu (2015). Portraying collective spatial attention in Twitter, *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (*KDD*), 39–48.

Aramaki, Eiji and Wakamiya, Shoko (2016). NAIST-ARS Guideline Ver. 1, https://dx.doi.org/10.6084/m9.figshare.3123160.v1 (in Japanese).

Aramaki, Eiji, Maskawa, Sachiko and Morita, Mizuki (2011). Twitter catches the flu: Detecting influenza epidemics using Twitter, *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 1568–1576.

Awamura, Takashi, Kawahara, Daisuke, Aramaki, Eiji, Shibata, Tomohide and Kurohashi, Sadao (2015). Location name disambiguation exploiting spatial proximity and temporal consistency, *Proceedings of the International Workshop on Natural Language Processing for Social Media* (*SocialNLP*), 1–9.

Bergsma, Shane, Dredze, Mark, Van Durme, Benjamin, Wilson, Theresa and Yarowsky, David (2013). Broadly improving user classification via communication-based name and location clustering on Twitter, *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, 1010–1019.

Bollen, Johan, Mao, Huina and Zeng, Xiaojun (2011). Twitter mood predicts the stock market, *Journal of Computational Science*, **2**, 1–8.

Chrupała, Grzegorz (2014). Normalizing tweets with edit scripts and recurrent neural embeddings, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (*Volume 2*: *Short Papers*), 680–686.

Ferguson, Neil M., Cummings, Derek A. T., Cauchemez, Simon, Fraser, Christophe, Riley, Steven, Meeyai, Aronrag, Iamsirithaworn, Sopon and Burke, Donald S. (2005). Strategies for containing an emerging influenza pandemic in Southeast Asia, *Nature*, **437**(7056), 209–214, http://www.ncbi.nlm.nih.gov/pubmed/16079797.

Han, Bo and Baldwin, Timothy (2011). Lexical normalisation of short text messages: Makn Sens a #twitter, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*: *Human Language Technologies*, 368–378.

Han, Bo, Cook, Paul and Baldwin, Timothy (2013). A stacking-based approach to Twitter user geolocation prediction, *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*: *System Demonstrations*, 7–12.

国立感染症研究所 (2006).『インフルエンザ・パンデミックに関する Q&A（2006.12 改訂版）』, 国立感染症研究所 感染症情報センター, 東京.

国立研究開発法人日本医療研究開発機構（AMED）(2015). 平成 28 年度「新興・再興感染症に対する革新的医薬品等開発推進研究事業」に係る公募について, http://www.amed.go.jp/koubo/010620151113-01.html.

Li, Jiwei, Ritter, Alan, Cardie, Claire and Hovy, Eduard (2014a). Major life event extraction from Twitter based on congratulations/condolences speech acts, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), 1997–2007.

Li, Jiwei, Ritter, Alan and Hovy, Eduard (2014b). Weakly supervised user profile extraction from Twitter, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), 165–174.

Marchetti-Bowick, Micol and Chambers, Nathanael (2012). Learning for microblogs with distant su-

pervision: Political forecasting with Twitter, *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 603–612.

奈良先端科学技術大学院大学ソーシャル・コンピューティング研究室 (2016). INFLU-KUN: Twitter-based Influenza Surveillance System, http://mednlp.jp/influ/.

O'Connor, Brendan, Balasubramanyan, Ramnath, Routledge, Bryan R. and Smith, Noah A. (2010). From Tweets to polls: Linking text sentiment to public opinion time series, *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media* (*ICWSM*), 122–129.

odomon.net (2013). Twitter ユーザー数[2013 年第一位 東京都], http://todo-ran.com/t/kiji/13528.

Pang, Bo, Lee, Lillian and Vaithyanathan, Shivakumar (2002). Thumbs up?: Sentiment classification using machine learning techniques, *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, 79–86.

Paul, M. J. and Dredze, M. (2011). You are what you tweet: Analysing Twitter for public health, *Processing of the Fifth International AAAI Conference on Weblogs and Social Media* (*ICWSM*).

Plank, Barbara, Hovy, Dirk, McDonald, Ryan and Søgaard, Anders (2014). Adapting taggers to Twitter with not-so-distant supervision, *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics*: *Technical Papers*, 1783–1792.

Sakaki, Takeshi, Okazaki, Makoto and Matsuo, Yutaka (2010). Earthquake shakes Twitter users: Real-time event detection by social sensors, *Proceedings of the 19th international conference on World Wide Web* (*WWW*), 851–860.

Shen, Chao, Liu, Fei, Weng, Fuliang and Li, Tao (2013). A participant-based approach for event summarization using Twitter streams, *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, 1152–1162.

谷田和章, 荒牧英治, 佐藤一誠, 吉田稔, 中川裕志 (2011). Twitter による風邪流行の推測, TETDM&情報編纂研究会(第 6 回), 42–47.

Thelwall, Mike, Buckley, Kevan and Paltoglou, Georgios (2011). Sentiment in Twitter events, *Journal of the American Society for Information Science and Technology*, **62**(2), 406–418.

Tumasjan, Andranik, Sprenger, Timm O., Sandner, Philipp G. and Welpe, Isabell M. (2010). Predicting elections with Twitter: What 140 characters reveal about political sentiment, *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media* (*ICWSM*), 178–185.

Varga, István, Sano, Motoki, Torisawa, Kentaro, Hashimoto, Chikara, Ohtake, Kiyonori, Kawai, Takao, Oh, Jong-Hoon and De Saeger, Stijn (2013). Aid is out there: Looking for help from tweets during a large scale disaster, *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), 1619–1629.

Williams, Jennifer and Katz, Graham (2012). Extracting and modeling durations for habits and events from Twitter, *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics* (*Volume 2*: *Short Papers*), 223–227.

Zhou, Deyu, Chen, Liangyu and He, Yulan (2014). A simple Bayesian modelling approach to event extraction from Twitter, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (*Volume 2*: *Short Papers*), 700–705.

246 Proceedings of the Institute of Statistical Mathematics Vol. 64, No. 2, 233–246 (2016)

Difference between Number of Tweets and Real World Statistics

Eiji Aramaki and Shoko Wakamiya

Nara Institute of Science and Technology (NAIST)

The prevalence of social media services has brought a new approach for surveying people and social conditions. So far, various systems, such as an influenza surveillance system, an earthquake detection system and so on, have been proposed. However, information shared on social media doesn't always correspond to the real one. For example, social media services often suffer from rumors, causing lower reliability than existing media. In addition, several studies have been pointed out a limitation of both temporal and spatial accuracy in social media services. In this paper we examine the differences in terms of temporal and spatial perspectives based on Twitter data collected using our influenza surveillance system. Furthermore, we discuss a bias behind the differences.

Key words: Social media, Twitter, natural language processing, social computing, influenza.