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

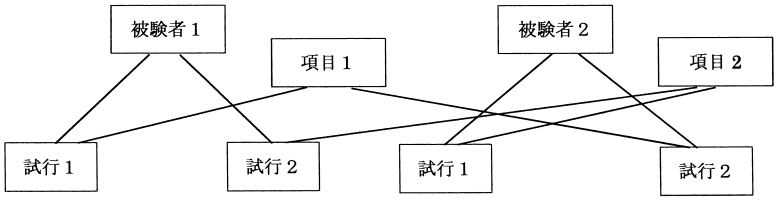

図 1. 複数の被験者と項目による階層的データの構造.

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

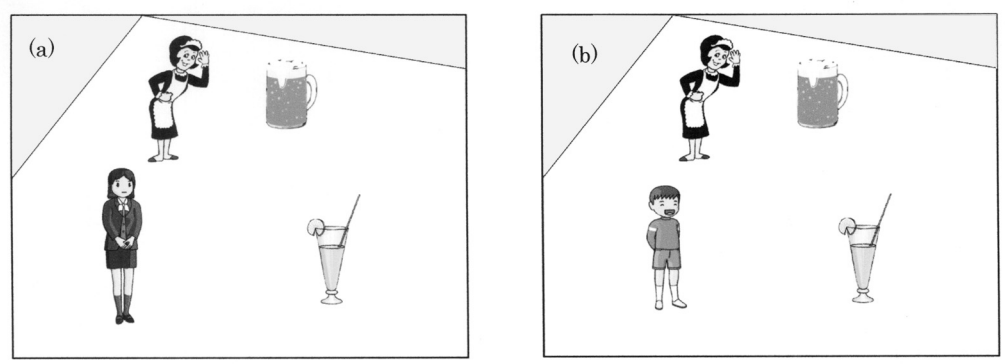

図 2. VWP 実験の刺激絵例.

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

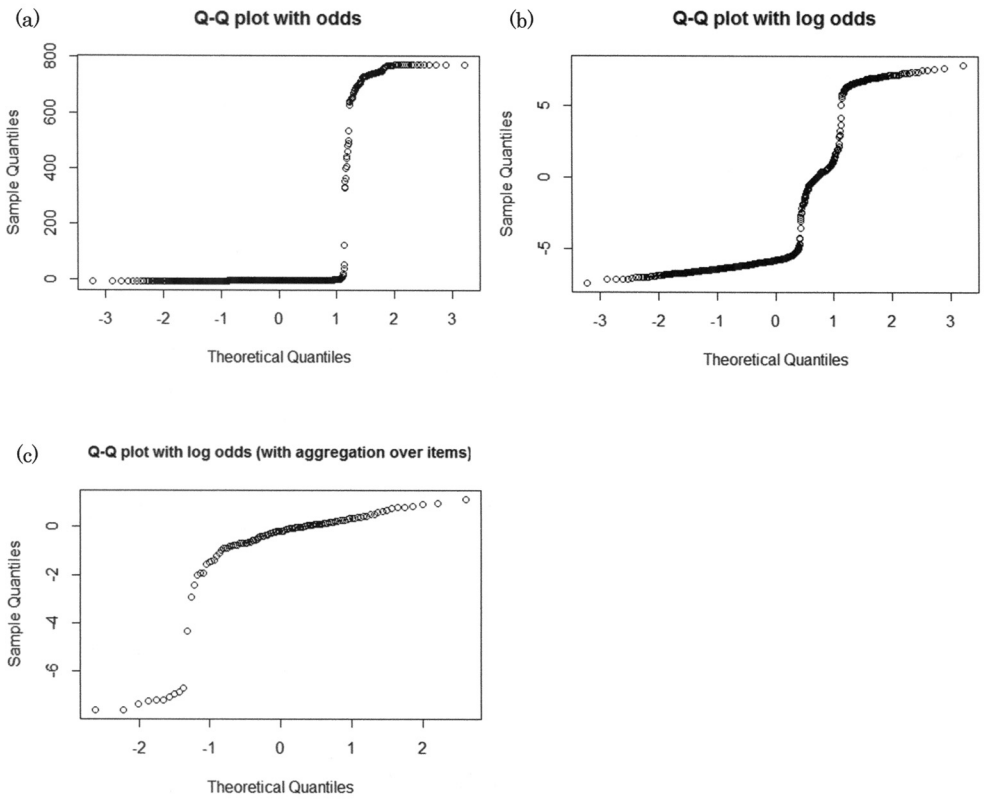

図 3. VWP 実験の眼球運動データから計算されたオッズ(a)，ロジット(b)，そして被験者ごとにデータ集約して計算したロジット(c)に対する線形混合モデルの残差の Normal Q-Q Plot.

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)")
```

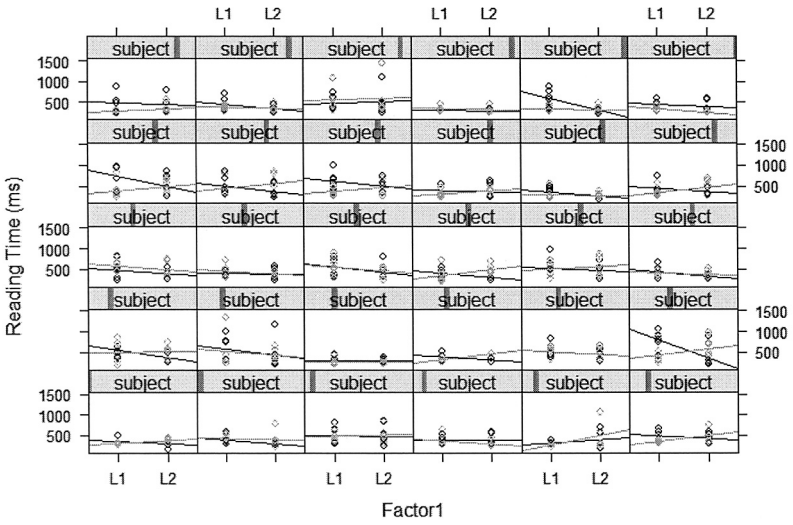

図 4．lattice パッケージの xyplot 関数による被験者ごとの読み時間における交互作用プロット．

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

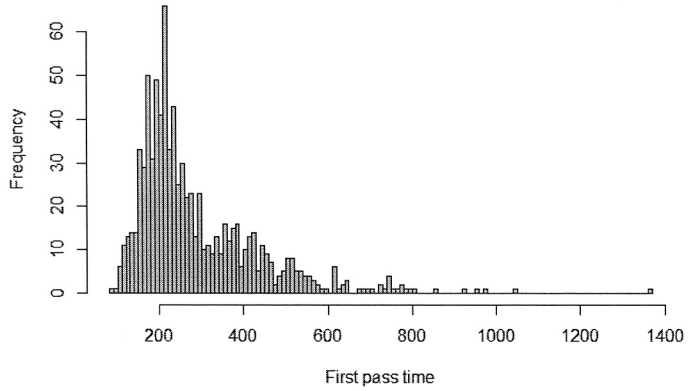

図 5. First pass 読み時間の分布の例.

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

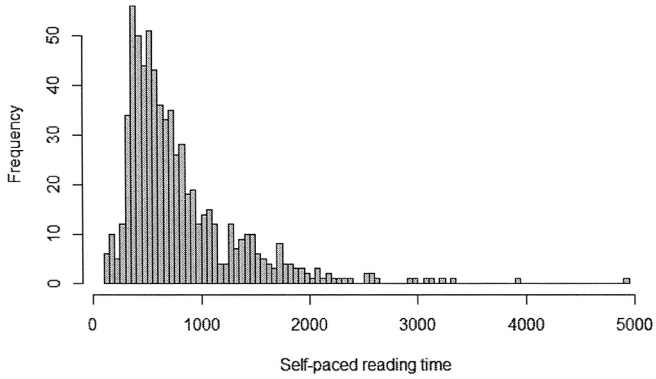

図 6. 自己ペース読み時間の分布例(Nakamura and Arai, 2016, 実験 2 より).

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

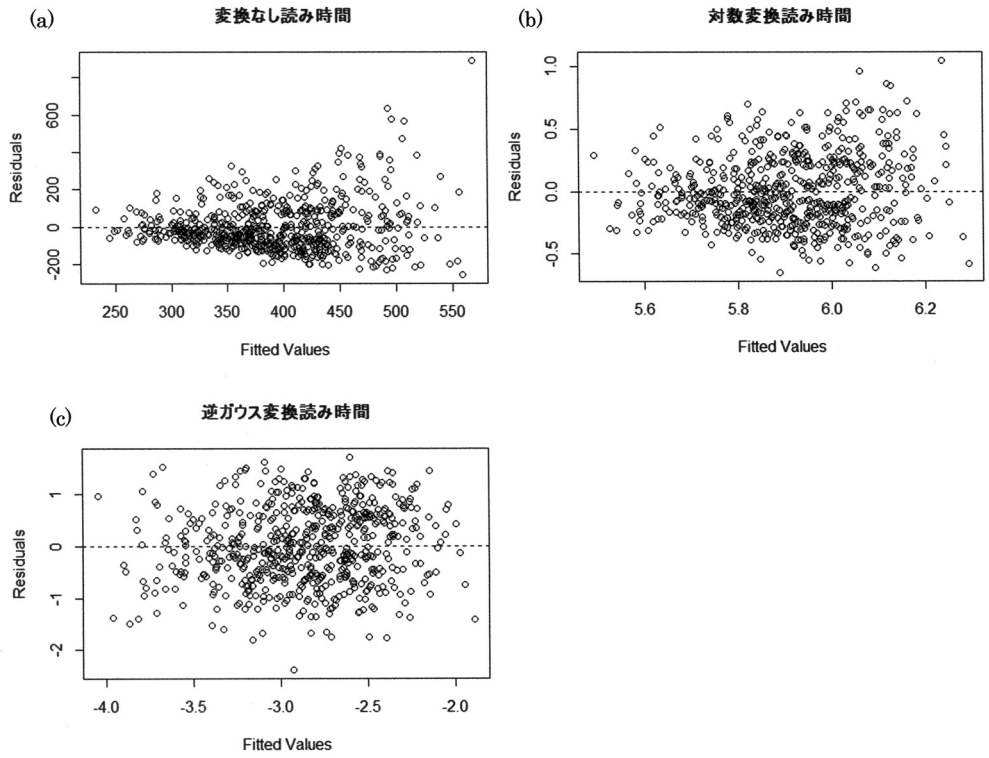

図 7. First pass 読み時間の変換なしの読み時間(a)，対数変換した値(b)，逆ガウス変換した値(c)に対する線形混合モデルの残差対推定値グラフ．

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

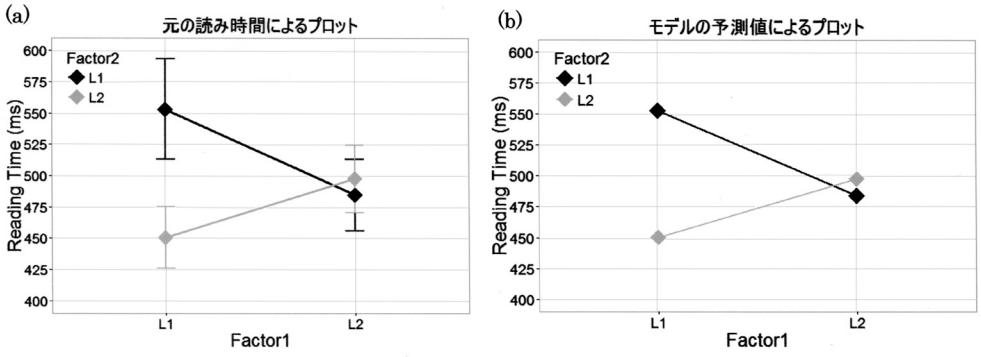

図 8．実測値による交互作用プロット(エラーバーは標準誤差)と最適モデルの切片，説明変数の係数から計算された値による交互作用プロット．

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.