英語教師に求められる統計分析技法

—— 混合効果モデルの初歩 ——*

法政大学文学部教授　石川　潔

1 はじめに

英語教師自らが、教育法の効果を調べる実験を行ったりすることはあるだろうが、言語を用いた実験の場合には、通常の統計学の入門書で紹介されている分析法には不十分な面がある。我々は実験結果を、実験参加者を越えて一般化したいだけでなく、用いた具体的な言語刺激をも越えて一般化したい。これは統計学的には、実験参加者と言語刺激の両方が変量要因となるということであるが、伝統的な検定法（*t* 検定や分散分析、古典的な回帰分析など）はこのような状況を扱うのが困難である。

本稿の目的は、このような状況に対処するための手段としての混合効果線形モデルを紹介することである。

第 2 節において、複数の変量要因に対応するために伝統的に採用されてきた手法を簡単に振り返る。第 3 節において、混合効果モデルの紹介を行うための例題として、Ishikawa *et al.*（2017）の実験およびデータを紹介する。第 4 節において、混合効果モデルの紹介を行い、第 5 節において、Ishikawa *et al.* のデータに対して具体的に混合効果モデルでの分析を行う手順を紹介する。第 6 節において、まとめとして、本稿で扱いきれなかった事柄をいくつか述べる。

2 複数の変量要因に対応するための伝統的な手法

上記のように、*t* 検定や分散分析などの伝統的な手法は、複数の変量要因がある状況に対処できないが、言語を用いた実験の場合は通常、実験参加者と同時に言語刺激も変量要因となってしまう。このことを最初に明確に指摘したのは Clark（1973）だが、皮肉なことに、Clark 自身は勧めなかった方法が、その指摘以降しばらくの間、広く採用されることになった。すなわち、一方では、実験参加者ごとの平均を求めた結果をデータとみなした検定（participant analysis）により実験参加者についての一般化を行い、他方では、刺激ごとの平均を求めた結果をデータとみなした検定（item analysis）によって刺激についての一般化を行う、という方法である（郡司・坂本, 1999 など）。

しかし、このような加工により、本来データにあった構造は破壊されてしまう。特に、多くの実験で行われるデータのスクリーニングの結果として欠損値が生じた場合には、上記のような加工での構造の破壊は深刻であろう。本当なら、そのような加工は行わない方が望ましい。混合効果モデルは、そのような加工を行わない手法である。

3 読み時間に対する長さ効果

混合効果モデルを紹介するにあたり、具体的な例題として、Ishikawa *et al.*（2017）が取り上げた問題を紹介する。

読解研究においては、実験者側が注目する条件を操作して、様々な領域の読み時間（self-paced reading の場合）や注視時間（眼球運動測定の場合）をデータとすることが多い（以下、両者をまとめて「読解時間」と呼ぶ）。しかし、条件操作とは別に、それぞれの領域の長さが異なることが一般的である。長い領域なら読解時間は長くなるので、条件効果を検出するためには、長さ効果を統制する必要がある。では、領域の「長さ」とは具体的に何だろうか？

伝統的には、それぞれの領域の文字数を「長さ」とみなし、実験参加者ごとに「文字数」による読解時間の線形回帰の予測式を推定し、それぞれの領域における実測値と予測値の間の残差をデータとみなす、という方法が採用されてきた（Ferreira and Clifton, 1986; Trueswell *et al.*, 1994）。「文字数」は（等幅フォントでの視覚呈示の場合）視線移動の量と対応するので、これはいわば、領域の「視覚的長さ」に着目した方法である。実際、筆者ら自身、Ishii and Ishikawa（2016）においては、このような方法を採用した。

しかし、このような方法の妥当性には疑問がある。というのは、黙読時にも「内言」が頭の中で鳴っているという The Implicit Prosody Hypothesis（Fodor

* 本稿の原稿にコメントをいただいた石井創氏に感謝する。勿論、なお残っている間違いなどの責任はすべて筆者にある。

表 1 刺激文（例）の、文字数およびモーラ数

呈示刺激	王は	冠を	外した
文字数	2	2	3
モーラ数	3	5	4

1998; Fodor and Hirose, 2003; Hirose, 2003 など）からしても、長さを視覚的なもの（文字数）だけに限定し、音韻的な「長さ」（音節数やモーラ数など）を先験的に排除することには、問題がある。確かに、英語などの表音文字を用いる言語なら、視覚的な長さと音韻的な長さとの間には十分強い相関があり、視覚的な長さによる統制と音韻的な長さによる統制とでは結果に大した違いはないかもしれない。しかし日本語においては、漢字が訓読みで用いられた場合、視覚的な長さと音韻的な長さとの間には大きな乖離が生じる場合がある。例えば

王は　冠を　外した。

というような文を考えてみよう。表 1 に見られるように、文字数としては同じ「長さ」だとしてもモーラ数としては違う「長さ」という場合があり得る（「王は」と「冠を」）。このような場合には、文字数による長さ効果の統制と、モーラ数による長さ効果の統制とでは、結果が極端に異なるであろう。では、長さ効果の統制にはモーラ数と文字数のいずれを用いるべきであろうか？ Hara（2010）、Jincho and Mazuka（2011）、Mazuka *et al.*（1997）などは、長さ効果の統制にモーラ数を用いているが、そうすべきだという統計学的な正当化は与えられていない。よって、Ishikawa *et al.*（2017）は、長さ効果の統制にモーラ数を用いるべきかどうかを検討した。採用された実験デザインおよび分析方針はきわめて単純である。まず、文字数とモーラ数とがなるべく乖離するような刺激文を作成し、self-paced reading でその読み時間を測定した。ついで、観察された読み時間データに対して、文字数およびモーラ数による（多重）線形回帰を行い、文字数およびモーラ数それぞれの予測への寄与分が有意かどうかを見ることにより、（両者の間で共通に予測される部分とは別に）それぞれの予測変数が独自に予測に寄与するかを検討する、という方法である。

では、このように得られたデータに対して、どのように具体的に多重線形回帰を実行したらよいだろうか？

実験では、刺激文の数が 24、実験参加者の数が 22 だったので、得られた生データは 24 × 22 = 528 個の読み時間であった。事後質問への答えや実験参加者ごとの標準偏差に基づくスクリーニングの結果、採用されたデータは 511 に減った。しかし、このデータをすべてそのまま用いて多重回帰を行って、文字数やモーラ数の予測への寄与分に関する検定を行うのは、統計学的には不当である。というのは、一人の実験参加者から得られた複数の読み時間データは、お互いに独立に得られたわけではないし、一個の刺激文から得られた複数の読み時間データも、お互いに独立しているわけではないからである（回帰に基づく検定は、それぞれのデータの間の独立性が前提となる）。

4　混合効果モデルの基本

上記のデータに対して混合効果モデルでの分析を行う前に、混合効果モデルの基本的な考え方をきわめて簡単に紹介しておく。混合効果モデルは、Baayen（2008）の出版により言語研究者にも広く知られるようになった。しかし、Baayen（2008）は統計分析環境 R において混合効果モデルでの分析を行うためのいわば how-to 本という色彩が濃厚であり、「考え方」を十分紹介しているわけではない。よってここでは、混合効果モデルの考え方を非常に簡単に述べる（但し本稿では、話をより単純化するため、線形回帰の場合を例にとる）。

4.1　重視すべき箇所についての発想の転換

最初に、古典的な（線形）回帰分析を例にして、伝統的な統計分析の仕組みを簡単に復習しておこう。まず、独立変数（予測変数）x の測定データがあり、従属変数（目的変数）y の測定データがあるとする。そこで我々は、

$$y = a + bx \tag{1}$$

という式でこの一連のデータをモデル化することを試みる。ここで、a は切片の係数、b は傾きの係数である。一般に、a および b の値をどのように決めても、y の予測値がすべて実測値と一致するというわけにはいかないが、何らかの方法により、予測値と実測値のズレが最小になるように a および b の値を決定する（具体的な方法としては、たいていの入門書では最小二乗法が紹介されている）。一般に、b の推定値はゼロにはならない。

このようにしてモデルが決まったら、次に、b についての検定を行うことになる。例えば、$b = 0$ と仮定した「空のモデル」と、b の推定値がゼロでないモデルとの間で、データとの適合度が有意に異なるかを調べる（相関係数に関する Fを用いた検定は、まさにこれにあたる）。両者が有意に異なるのなら、ゼロでな

い傾きを想定することに、すなわち、x という独立変数を想定することに、統計学的な意味がある、ということになるから、x と y との間に有意味な相関があるという結論が得られる。

ここで重要なのは、我々は通常、「検定」という段階にのみ注目しがちであり、「モデル化」という段階にあまり注意が行かないことである。実際、伝統的な participant analysis と item analysis の組み合わせという手法は、伝統的なモデル化の手法を用いたまま、データ加工により複数の変量要因に対する検定を即座にやってしまおう、という手法であった。しかし混合効果モデルは、（名前の通り）複数の変量要因の効果をモデル化の段階で組み込んでしまうという手法である。つまり、混合効果モデル分析を行うためには、まずは「モデル化」という段階に注目する必要がある。

4.2 混合効果モデル（予備編）：切片および傾きにおける固定効果および変量効果

よって、まずは最初のモデル化の段階から考え直してみよう。実験参加者 A～D のそれぞれについて、独立（予測）変数 x および従属（目的）変数 y の値を測定したとする。実験参加者ごとに回帰直線を求めると、例えば表 2 および図 1 のような状況になるだろう。つまり、実験参加者ごとに、切片の係数 a も傾きの係数 b も異なる、というのが普通であろう。このような状況でデータ全体をモデル化するとなると、

- データ全体としての切片の値は、「全員に共通の切片の成分」と「個人ごとに異なる切片のランダムな成分」との両方から成り立っている。
- データ全体としての傾きの値は、「全員に共通の傾きの成分」と「個人ごとに異なる傾きのランダムな成分」との両方から成り立っている。

と考えられる。

混合効果モデルの基本は、このように、それぞれの係数が「全体に共通の成分」と「（実験参加者などの）変量因子ごとに異なるランダムな成分」とから成り立っているという想定でモデル化を行うことである。変量因子が実験参加者のみの場合は、モデル式は

$$y = (a + u_i) + (b + v_i)x \qquad (2)$$

となる。ここで、a および b がそれぞれ、全員に共通

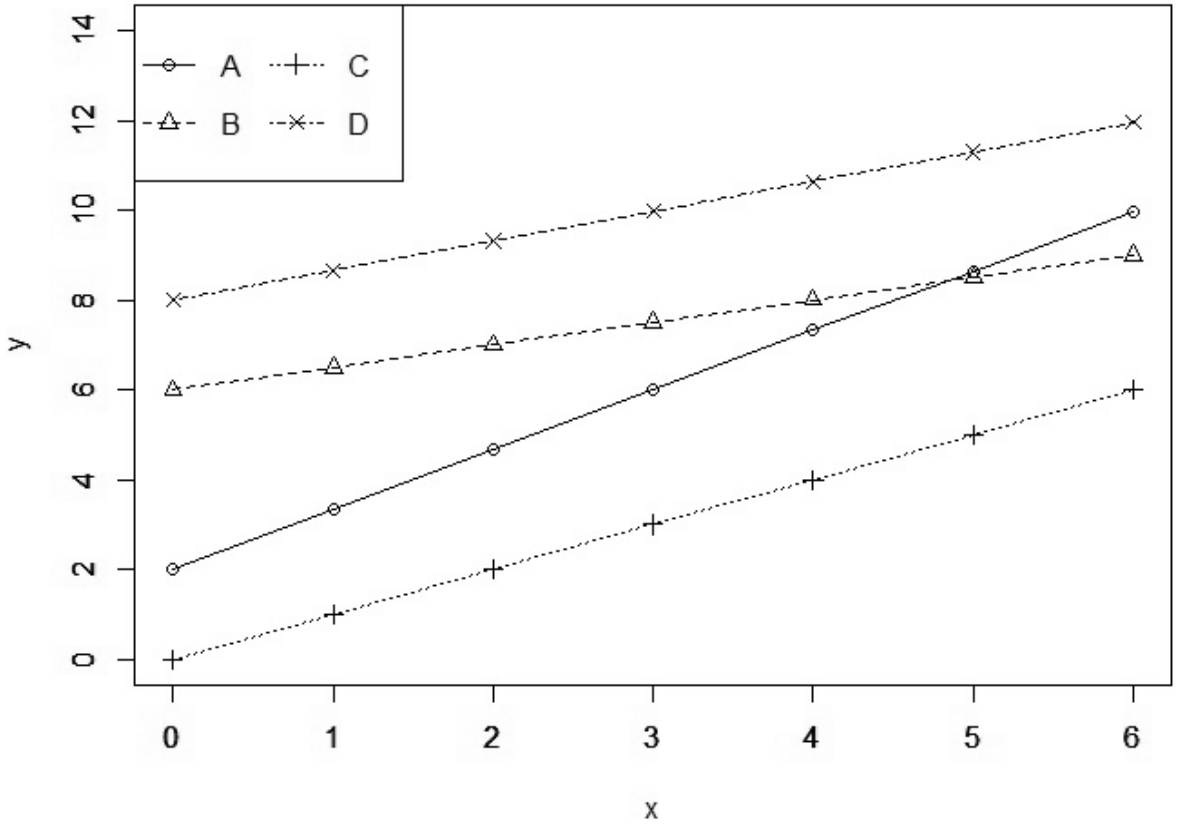

図 1　実験参加者ごとの回帰直線の例

表 2　実験参加者ごとの回帰式の例

実験参加者	回帰式
A	$y = 2 + 1.33x$
B	$y = 6 + 0.50x$
C	$y = 0 + 1.00x$
D	$y = 8 + 0.66x$

な切片および傾きの成分であり、u_iおよび v_iがそれぞれ、実験参加者ごとに異なる切片および傾きのランダムな成分である。

4.3　混合効果モデル（非・予備編）: 係数の推定

上記のモデル式には、実験参加者それぞれについての切片および傾きの係数が入っているが、現実の混合効果モデル分析（例えば統計分析環境 R 上で実装されている lme4; Bates et al., 2015） では、そういった係数の具体的な値が推定されるわけではなく、変量効果（間）の（共）分散や固定効果間の相関のみが推定される。

しかし、それでも数学的に十分複雑であり、伝統的な回帰分析のように演繹的に推定値を求めることは出来ない。よって、（係数の推定法としては、最小二乗法でなく最尤法を用いた上で）一種の試行錯誤によって推定値を求めることになる。しかし、推定がいつも成功するとは限らない。上記の lme4 を用いて混合効果モデル分析を実際に行ってみると、推定が converge しなかったというメッセージが出力されることがしばしばあるが、これは、想定されたモデルの係数の推定がうまくいかなかった、ということである。このような場合にはしばしば、モデル中のランダムな成分の項の数を減らす（例えば刺激による傾きのランダムな成分をモデル式から外すなど）ことにより、モデルを単純化した上で改めて推定を行う、ということになる（但し、このようなモデル選択が正しい方法かについては論争がある。第 6 節を参照）。

5　混合効果モデルでの分析：実践例

無事にモデル推定が出来たら、次は検定を行いたいわけだが、ここでは上記の Ishikawa *et al.*（2017）の実験データを例として、具体的な手順とともに検定の方法（の一例）を見ることにしよう。

最初のステップは、データをいわゆる long format に整形することである。すなわち、「文字数」や「モーラ数」といった、実験者側が注目している独立（予測）変数のみでなく、「実験参加者」や「刺激」といった変量因子も、一種の独立（予測）変数とみなして、例えば表 3 のようにデータを整形する（ここでは一人の実験参加者のデータの箇所の一部のみ示した）。

表 3　Ishikawa *et al.* (2017) の long format でのデータ：sbj は実験参加者 ID、char はそれぞれの刺激文の文字数、mora はそれぞれの刺激文のモーラ数、time は読み時間、item は刺激文 ID

sbj	char	mora	time	item
A	11	17	1783	a
A	10	18	4512	b

Ishikawa *et al.*（2017）は、このような long format でのデータを、統計分析環境 R（ver 3.2.2; R Core Team, 2015）に読み込み、lme4（ver. 1.1-10; Bates *et al.*, 2015）を用いて分析した。具体的には、プロンプトに対して

```
lmer(time ~ 1 + char + mora +
(1 + char + mora | sbj)
+ (1 | item), data = …)
```

とタイプする。冒頭の lmer は（ モデル推定を行う）関数名であり、最後の data = … という部分はデータ・オブジェクトの指定なので、モデルの指定の部分そのものは

```
time ~ 1 + char + mora +
(1 + char + mora | sbj) + (1 | item)
```

となる。ここで、チルダの左側の time が従属（目的）変数、1 は「切片あり」という意味、[1] カッコに入っていない char および mora は、「文字数およびモーラ数による傾きを固定効果として入れる」という指定となる。カッコの中に入った部分が変量効果の指定となる。

```
(1 + char + mora | sbj)
```

は、「実験参加者ごとに、切片、文字数の傾き、モーラ数の傾きが異なる」という指定、

```
(1 | item)
```

[1] 傾きが指定されている場合の 1 は省略できる（上記の場合、カッコの外の 1、および（1 + char + mora | sbj）における 1）。省略しても、切片の成分は自動的に補われる。

表 4　Ishikawa *et al*. (2017) のデータに対する混合効果モデル分析の結果（固定効果）

	Coefficient	SE	t
(Intercept)	505.85	439.34	1.151
char	-74.79	94.93	-0.788
mora	164.39	57.08	2.880

は、「刺激ごとに、切片が異なる」という指定となる（文字数やモーラ数はそれぞれの刺激の属性なので、刺激ごとの傾きの違いはモデルに入れなかった）。lme4 は、この指示に基づき、モデル構築を行う。このモデルを例えば model2 というオブジェクトとして構築したなら、

```
summary(model2)
```

とタイプすることにより、得られたモデルの固定効果の推定結果（など）が表示される。表 4 は、そうやって得られた固定効果（全体に共通な切片および傾き）の推定結果である（このときのデータの場合には、上記のモデル指定のままで推定は converge した）。

ここで注意すべきは、係数の推定値に続いて、標準誤差（SE）および *t* 値が載っているが、*p* 値は載っていないことである。つまり、（切片または）傾きがゼロであるという帰無仮説に基づく検定が行われていないのである。実は、混合効果モデル分析での *p* 値の算出法についての研究者間の合意が得られていないのだ。しかし、ここでは、文字数（char）およびモーラ数（mora）それぞれが独立（予測）変数として統計学的に有意味だったかどうかを知りたいのであれば、どうしたらよいだろうか？

しばしば用いられる手っ取り早い方法は、「*t* 値の絶対値が 2 を越えている固定効果は有意とみなす」というものであるが、この方法によれば、モーラ数単独の効果は有意だということになる。Ishikawa *et al*.（2017）はさらに、

- 独立（予測）変数が文字数のみ（文字数のみモデル）
- 独立（予測）変数がモーラ数のみ（モーラ数のみモデル）

などの混合効果モデルを別途構築して、文字数・モーラ数の両方の独立（予測）変数が入っているモデルとの間での対数尤度比によるモデル間の比較を行った。

例えば、「文字数のみモデル」は、

```
time ~ 1 + char +
(1 + char + mora | sbj) + (1 | item)
```

というモデル指定で構築できる。この推定結果を model1 というオブジェクトとして構築したなら、

```
anova(model1, model2)
```

とタイプすることにより、対数尤度比による検定が行われて結果が表示されるが、その結果、モーラ数をも入れたモデルは文字数のみのモデルより有意にデータに適合していた［$\chi(1) = 8.1098$, $p = .004403$］。この結果に基づき筆者らは、（少なくとも self-paced reading における）読み時間を予測する変数としてのモーラ数の（文字数では捉えられない）効果は有意であると結論した。[2]

6　まとめ

ここでは、混合効果モデルのきわめて初歩的な紹介を行ったが、ここでの紹介は二重の意味で不十分であるのも確かである。

ここでの紹介が不十分である第一の理由は、本稿で紹介した実践例が「正解」とは限らないということである。例えば、モデルにどのような変量効果を入れるのが正当かについては論争がある（Barr *et al*., 2013; Bates *et al*., 2015; Matuschek *et al*., 2017）。本稿での実践例では、想定し得る変量効果をすべて入れたモデル推定が converge したので、それをそのままモデルとして採用したが、その選択が正しいとは限らない。

また、モデル比較の方法としては、対数尤度比による検定は唯一の方法でもない（AIC や BIC といった適合度の指標も広く知られている）。

また、本稿で紹介した実践例では、（一連の先行研究と同様）自然数である予測変数に対して線形回帰を適用しているが、厳密に言えば、自然数が正規分布を構成するわけがない。正規でない分布と考えられるデータ（例えば「正答」と「誤答」）について、様々な分布を用いた「一般化線形モデル」が開発されているが（例えばロジスティック回帰）、一般化線形モデルも混合効果モデルとして構成できる（久保, 2012）。本稿での実践例では、（一連の先行研究と同様）変数の分布の種類についての検討は行なわれていない。

[2] 他方で、「モーラ数と文字数の両方を用いたモデル」と「モーラ数のみモデル」との間での検定結果は、有意でなかった。つまり、文字数を予測変数として採用しても適合度は有意に向上しなかったということであり、これは表 4 における char の *t* 値の絶対値の小ささと整合する。

さらに、本稿での実践例は、いわゆる頻度論の統計学（Fisher、Peason、Neyman などの流派）に基づいているが、最近はベイズ流の統計学も広がりを見せつつある。ベイズ流でも混合効果モデルは可能であり、本稿の実践例のデータに対してのベイズ流の分析も、可能な選択肢の一つである。

ここでの紹介が不十分である第二の理由は、独立（予測）変数が名義尺度である場合（従来 *t* 検定や分散分析が用いられてきたような場合）が取り上げられていないことである。*t* 検定や分散分析は、予測変数をダミー変数でコード化することにより、線形回帰として表現しなおすことができるので、そのような実験にも混合効果モデルを用いた分析が可能だが、本稿ではそのような状況（やコード化の詳細）はカバーできなかった。また、名義尺度の独立（予測）変数が複数ある場合、（ 要因数または水準数が 3 以上の際の交互作用や）単純主効果の分析、そして多重比較なども、lme4 を用いるだけでは難しい。

このように、混合効果モデルを用いた分析は、従来の *t* 検定、分散分析、古典的な線形回帰などのようないわば「パッケージ化」された分析法に比べて難易度は確かに上がる。しかし、データが本来持っている構造をなるべく破壊せずに、複数の変量要因の効果に対処するためには、我々はその難易度と付き合わざるを得ない。

文献表

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

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. http://dx.doi.org/10.1016/j.jml.2012.11.001.

Bates, D., Kliegl, R., Vasishth, S., and Baayen, R. H. (2015). Parsimonious mixed models. Available from arXiv:1506.04967 (stat.ME).

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67(1)*, 1–48. Doi:10.18637/jss.v067.i01.

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.

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

Fodor, J. D. (1998). Learning to parse? *Journal of Psycholinguistic Research, 27(2)*, 285–319.

Fodor, J. D., and Hirose, Y. (2003). What Japanese parsing tells us about parsing. In McClure, W., ed., *Japanese/Korean Linguistics, Vol. 12*. Stanford: CSLI Publications, pp. 192–205.

郡司隆男・坂本勉. (1999).『言語学の方法』. 東京：岩波書店.

Hara, M. 2010. Second language gap processing of Japanese scrambling under a Simpler Syntax account. In VanPatten, B., and Jegerski, J. (eds.), *Research in Second Language Processing and Parsing*. Amsterdam and Philadelphia: John Benjamins, pp.177–205.

Hirose, Y. (2003). Recycling prosodic boundaries. *Journal of Psycholinguistic Research, 32(2)*, 167–195.

Ishii, S., and Ishikawa, K. (2016). The bi-directionality and the graded nature of aspectual coercion: An eye-tracking study. 電子情報通信学会技術研究報告 IEICE Technical Report Vol. 116 No.159, pp.43–48 (TL2016-20).

Ishikawa, K., Yamashita, R., and Ishii, S. (2017). Mora-based control for the length effect: A self-paced reading study in Japanese. 電子情報通信学会技術研究報告 IEICE Technical Report Vol. 117 No.149, pp. 63–66 (TL2017-25).

Jincho, N., and Mazuka, R. (2011). Individual differences in sentence processing: Effects of verbal working memory and cumulative linguistic knowledge. In Yamashita, H., Hirose, Y., and Packard, J. (eds.)., *Processing and Producing Head-final Structures*. Dordrecht: Springer, pp. 49–65.

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』. 東京：岩波書店.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., and Bates, D. (2017). Balancing type I error and power in linear mixed models. *Journal of Memory and Language, 94*, 305–315, http://dx.doi.org/10.1016/j.jml.2017.01.001.

Mazuka, R., Itoh, K., and Kondo, T. (1997). Processing down the garden path in Japanese: Processing of sentences with lexical homonyms. *Journal of Psycholinguistic Research, 26(2)*, 207–228.

R Core Team. (2015). A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R- project/org/.

Trueswell, J. C., Tanenhaus, M. K., and Garnsey, S.M. (1994). Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language, 33*, 285–318.