# Summer R Study Group Report What I Couldn't Have Done Otherwise

Josef Fruehwald

University of Pennsylvania

Splunch

### Outline

```
The R Study Group
Why use R?
Novel Results
    ING
        Grammatical Class
        Following Segment
       Style
       Post-Collapse
       First Model
       Interactions
        Conclusions
    TD
       Slope
        Category Effects
        Interaction Effects
        Refinement Across Subsets
       Final Refined Model
        TD Conclusions
```

### I know what I did this summer...

- ► This summer I led an R study group that was regularly attended by a group of committed individuals
- R Study Group Page

### What did we accomplish?

- ▶ We all learned quite a bit of R.
- ▶ Laid the groundwork for the R study group meeting this semester.

# Why use R?

R is becoming the statistics software paradigm of our day.

- ► Why?
  - Free
  - Open Source
  - Platform independent
  - High Power of Use : Difficulty of Use Ratio
- Consequences
  - ► Large Committed User Base
  - Ever Expanding Functionality
  - ▶ Broad Support

When we do our data analysis in R, we benefit from, and contribute to the use of statistics in the whole of the social sciences.

# One Stop Shopping

ae, 3 levels ae, 4 levels ae, 5 levels ae, 6 levels

# Why this is worth talking about

- ▶ I used ING and TD model data sets
- Applied a broader range of techniques than usual
- Some novel / More clearly stated results were achieved

### ING

### The Variable

ING is the name given to variable realization of unstressed  $/\text{-}\mathrm{n}\mathrm{j}/$  as  $[\text{-}\mathrm{n}\mathrm{j}]$  or  $[\text{-}\mathrm{m}].$  It has been shown to be affected by morphological identity of the affix, as well as by usual sociolinguistic influences, like style.

### The Data

The data for this presentation is 1139 tokens coded by hand from sociolinguistic interviews, mostly of South Philadelphians. The coding was done by members of the 2006-2007 Ling 560 class. The predictors of ING that I'll be looking at will be Grammatical Status, Following Segment, and Style.

#### **Data Location**

The data here is located at http://www.ling.upenn.edu/~joseff/rstudy/data/ing2.csv

### Grammatical Status: First Glance

Let's begin by looking at the distribution of variants across Grammatical Classes:

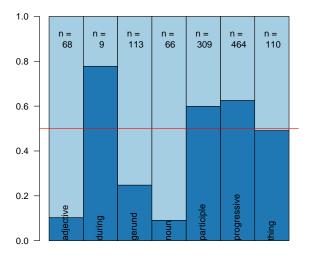
```
> ing.gram.table <- table(ing$DepVar, ing$GramStatus)</pre>
```

> prop.table(ing.gram.table, 2)

```
adjective during gerund noun In 0.1029412 0.7777778 0.2477876 0.0909091 Ing 0.8970588 0.2222222 0.7522124 0.9090909
```

```
participle progressive thing
In 0.5987055 0.6250000 0.4909091
Ing 0.4012945 0.3750000 0.5090909
```

### Grammatical Status: First Glance



# Proper Category Subdivisions?

We coded for many different categories within Grammatical Class. Our category definitions are probably too fine grained, but our knowledge of the morphological variation involved isn't well defined enough to have known that before hand.

### Motivation for Collapsing Categories

If we can fit predictive models just as well with fewer categories per factor (drop degrees of freedom), the power of our models go up.

#### Caution

If we collapse across categories that are meaningfully different, but accidentally have the same effect, the value of our predictive model decreases.

# Grammatical Status: Comparisons Across Groups

We can eyeball equivalent grammatical classes from the plot before, but it's probably best to do a pairwise test of proportions, controlling for the fact multiple tests are being done.

```
> pair.gram1 <- pairwise.prop.test(t(ing.gram.table),
+    p.adj = "bonf")</pre>
```

	adjective	during	gerund	noun	participle	progre
during	0.00016					
gerund	0.59	0.057				
noun	1	7.4e-05	0.36			
participle	7.6e-12	1	7.4e-09	4e-12		
progressive	3.5e-14	1	2.2e-11	1.8e-14	1	
thing	5.9e-06	1	0.0061	3.1e-06	1	

# Grammatical Status: Collapsing Across Groups

It looks like we can safely collapse participles and progressives into one category, and nouns and adjectives into another. The "thing" class has a very similar proportion of [-in] to the participles and progressives, but we have no theoretical reason to believe why this should be so.

### Grammatical Status: Second Look

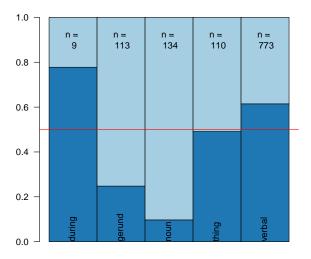
Let's recalculate the percentages across these new categories

- > ing.gram.table2 <- table(ing\$DepVar, ing\$Gram2)</pre>
- > prop.table(ing.gram.table2, 2)

during gerund noun thing
In 0.77777778 0.24778761 0.09701493 0.49090909
Ing 0.22222222 0.75221239 0.90298507 0.50909091

verbal In 0.61448900 Ing 0.38551100

### Grammatical Status: Second Look



### Grammatical Status: New Comparisons

And a pairwise test across new categories.

```
> pair.gram2 <- pairwise.prop.test(t(ing.gram.table2),
+ p.adj = "bonf")</pre>
```

	during	gerund	noun	thing
gerund	0.027			
noun	2e-06	0.027		
thing	1	0.0029	1.9e-10	
verbal	1	4.2e-12	3.9e-27	0.18

"during" appears to not be significantly different from "thing" or the verbal class, but with only 9 tokens this shouldn't be surprising.

# Following Segment: First Glance

Now let's look at the distribution across Following Segment

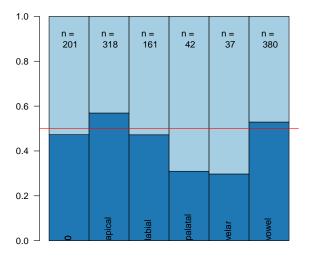
```
> ing.fol.table1 <- table(ing$DepVar, ing$Following.Seg)</pre>
```

> prop.table(ing.fol.table1, 2)

```
0 apical labial palatal
In 0.4726368 0.5691824 0.4720497 0.3095238
Ing 0.5273632 0.4308176 0.5279503 0.6904762
```

velar vowel
In 0.2972973 0.5289474
Ing 0.7027027 0.4710526

# Following Segment: First Glance



# Following Segment: First Glance

As per past descriptions of ING, the Following Segment has a much weaker effect than Grammatical Status. However, it doesn't appear to be non-existant. Both palatals and velars appear to have much higher rates if [-ing] than the others. This shouldn't be too surprising, since nasal-place assimilation, especially to a following velar, is such a common process.

# Following Segment: Comparisons Across Groups

Here's a pairwise proportion test, using the Bonferroni correction for multiple tests

```
> pair.fol1 <- pairwise.prop.test(t(ing.fol.table1),
+ p.adj = "bonf")</pre>
```

	0	apical	labial	palatal	velar
apical	0.6				
labial	1	0.83			
palatal	1	0.039	1		
velar	1	0.045	1	1	
vowel	1	1	1	0.17	0.18

# Following Segment: Collapse Across Groups

I'll collapse the palatal and velar categories into one posterior category, leave the apical category alone, and collapse all others into an "other" category. My hypothesis is that it is a lingual articulation that will affect ING.

# Following Segment: Second Look

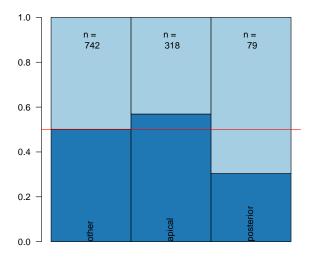
Here is the distribution across new categories of Following Segment

```
> ing.fol.table2 <- table(ing$DepVar, ing$Tongue)</pre>
```

> prop.table(ing.fol.table2, 2)

other apical posterior In 0.5013477 0.5691824 0.3037975 Ing 0.4986523 0.4308176 0.6962025

# Following Segment: Second Look



# Following Segment: New Comparisons

```
> pair.fol2 <- pairwise.prop.test(t(ing.fol.table2),
+ p.adj = "bonf")</pre>
```

	other	apical
apical	0.15	
posterior	0.0038	0.00012

## Style: First Glance

We have a lot of categories in the Style factor group.

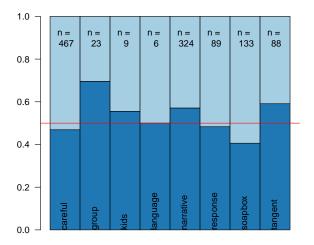
```
> ing.sty.table <- table(ing$DepVar, ing$Style)</pre>
```

> prop.table(ing.sty.table, 2)

```
careful group kids language
In 0.4689507 0.6956522 0.5555556 0.5000000
Ing 0.5310493 0.3043478 0.4444444 0.5000000
```

narrative response soapbox tangent
In 0.5709877 0.4831461 0.4060150 0.5909091
Ing 0.4290123 0.5168539 0.5939850 0.4090909

# Style: First Glance



### Style: First Glance

With style, it is unclear how to collapse across categories, as the nature of what makes one style different from another is a lot more fluid. Yet, I'd like to eliminate those marginal categories like "kids" and "language," and generally reduce the number of factors here.

## Style: Collapsing Across Categories

Based on their numbers and behavior, I'll be collapsing careful and response into one "careful" category, narrative and tangent into one "narrative" category, leave soapbox alone, and collapse all others into one "other" category.

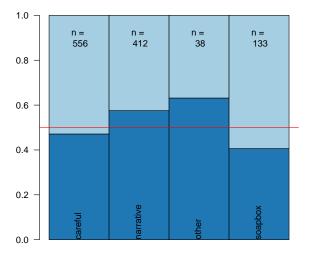
### Style: Second Look

Here is the distribution across the new categories.

- > ing.sty.table2 <- table(ing\$DepVar, ing\$Style2)</pre>
- > prop.table(ing.sty.table2, 2)

careful narrative other soapbox In 0.4712230 0.5752427 0.6315789 0.4060150 Ing 0.5287770 0.4247573 0.3684211 0.5939850

# Style: Second Look



# Style: New Comparisons

And here is the pairwise proportion test using the Bonferroni adjustment.

```
> pair.sty <- pairwise.prop.test(t(ing.sty.table2),
+ p.adj = "bonf")</pre>
```

	careful	narrative	other
narrative	0.01		
other	0.48	1	
soapbox	1	0.0058	0.14

# Was all that really ok?

Was it really ok to collapse all of those categories? What affect with it have on the models we eventually fit?

#### Good News

We can directly inspect the impact of all of these category collapses on the fitted models.

### Check Collapses

Step 1: Fit the models

Here, I fit 4 models. Each one has one more collapsed category in it than the one before.

```
> model1 <- glm(DepVar ~ GramStatus + Following.Seg +
+ Style, data = ing, family = binomial)
> model2 <- glm(DepVar ~ Gram2 + Following.Seg +
+ Style, data = ing, family = binomial)
> model3 <- glm(DepVar ~ Gram2 + Tongue +
+ Style, data = ing, family = binomial)
> model4 <- glm(DepVar ~ Gram2 + Tongue +
+ Style2, data = ing, family = binomial)</pre>
```

# Check Collapses

#### Step 2: Look at the Deviance

- > collapseTest <- anova(model1, model2,</pre>
- + model3, model4, test = "Chisq")

	Resid. Df	Resid. Dev	Df	Deviance	P(> Chi )
1	1120	1382.18			
2	1122	1382.54	-2	-0.35	0.8377
3	1125	1382.56	-3	-0.02	0.9991
4	1129	1385.12	-4	-2.56	0.6333

- 1. Original Categories
- 2. Collapsed Grammatical Class
- 3. Col Gram + Collapsed Following Segment
- 4. Col Gram + Col FolSeg + Collapsed Style

#### Result

No single stepwise collapse significantly worsened the fit of the model.

### Check Collapses

#### Step 2: Look at the Deviance

```
> totalCollapseTest <- anova(model1, model4,
+ test = "Chisq")</pre>
```

	Resid. Df	Resid. Dev	Df	Deviance	P(> Chi )
1	1120	1382.18			
2	1129	1385.12	-9	-2.94	0.9666

- 1. Original Categories
- 2. Col Gram + Col FolSeg + Collapsed Style

#### Result

The jump from the fully fine grained model to the fully collapsed model doesn't significantly worsen the fit of the model either.

### To Review...

- Grammatical Status
  - ► Adjective, Noun, Participle, Progressive, Gerund, During, Thing → (Adjective, Noun), (Participle, Progressive), Gerund, During, Thing → Noun, Verbal, Gerund, During, Thing
- Following Segment
  - ◆ 0, Labial, Vowel, Apical, Palatal, Velar → (0, Labial, Vowel),
     Apical, (Palatal, Velar) → Other, Apical, Posterior
- Style
  - Careful, Response, Narrative, Tangent, Soapbox, Group, Language, Kids → (Careful, Response), (Narrative, Tangent), Soapbox, (Group, Language, Kids) → Careful, Narrative, Soapbox, Other

### What does it buy us?

- ► More Power
- ► More Interpretable Coefficients

### First Model

Let's examine the significance of the factor groups in our fully collapsed model in an analysis of deviance table:

> factorGroups <- anova(model4, test = "Chisq")</pre>

	Df	Deviance	Resid. Df	Resid. Dev	P(> Chi )
NULL			1138	1578.79	
Gram2	4	174.20	1134	1404.59	1.312e-36
Tongue	2	7.76	1132	1396.84	0.0207
Style2	3	11.71	1129	1385.12	0.008438

#### Result

All three factor groups are significant.

#### Note

For Sociolinguists: This is what we try to do when calculating the range of GoldVarb factor weights.

#### Setting Up Contrasts

Before we examine the coefficients of the model fit, we should first be sure that we are using the correct contrasts for each factor group.

#### **Treatment**

If there is some factor level that we could consider "untreated," or the control, we should use treatment contrasts. The coefficients of the other levels will represent their difference from this untreated level.

#### Sum

If we cannot determine which factor level is the untreated one, then we should probably use sum contrasts. The coefficients for each level will represent the difference of each level from a sort of abstract mean effect

### Note for sociolinguists

We use sum contrasts we use when we select "centered" in GoldVarb. That is why all factor weights will average to 50%

I've written a script to make changing factor contrasts easier. You can find it at

 $http://www.ling.upenn.edu/{\sim}joseff/scripts/recontrast.R$ 

#### **Treatment**

I've hypothesized that only a lingual articulation will effect ING, so the "other" level in Tongue is untreated

```
> levels(ing$Tongue)
```

```
[1] "other" "apical" "posterior"
```

> ing\$Tongue <- recontrast(ing\$Tongue, "treatment")</pre>

#### Sum

We can't determine which grammatical class or style is untreated, so we will use sum contrasts for these

```
> ing$Gram2 <- recontrast(ing$Gram2, "sum")</pre>
```

> ing\$Style2 <- recontrast(ing\$Style2, "sum")</pre>

Style2-careful

Style2-other

Style2-narrative

#### Model Fit

```
> model4 <- glm(DepVar ~ Gram2 + Tongue +
```

Style2, data = ing, family = binomial)							
	Estimate	Std. Error	z value	Pr(> z			
(Intercept)	0.4255	0.2145	1.98	0.04			
Gram2-during	-1.3495	0.6541	-2.06	0.03			
Gram2-gerund	0.7022	0.2502	2.81	0.00			
Gram2-noun	1.8340	0.2914	6.29	0.00			
Gram2-thing	-0.3805	0.2366	-1.61	0.10			

(Intercept)	0.4255	0.2145	1.98	0.0474
Gram2-during	-1.3495	0.6541	-2.06	0.0391
Gram2-gerund	0.7022	0.0541	2.81	0.0050
Gram2-noun	1.8340	0.2914	6.29	0.0000
Gram2-thing	-0.3805	0.2366	-1.61	0.1077
Tongue-apical	-0.2281	0.1476	-1.54	0.1224
Tongue-posterior	0.5633	0.1470	2.04	0.1224

0.0215

-0.2661

-0.1959

0.1230

0.1287

0.2627

0.18

-2.07

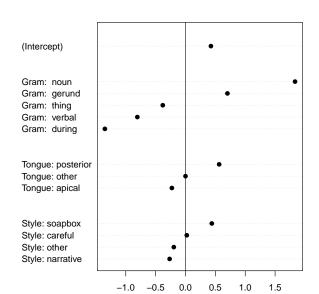
-0.75

0.8610

0.0387

0.4557

 $\leftarrow$  More -in ; More -ing  $\rightarrow$ 



# Refitting

The effect size for careful speech was very close to 0, and there may be reason to think this is an untreated style. So, I'd like to re-contrast and refit the model this way.

#### **Treatment**

- > ing\$Style2 <- relevel(ing\$Style2, "careful")</pre>
- > ing\$Style2 <- recontrast(ing\$Style2, "treatment")</pre>
- > ing\$Tongue <- recontrast(ing\$Tongue, "treatment")</pre>

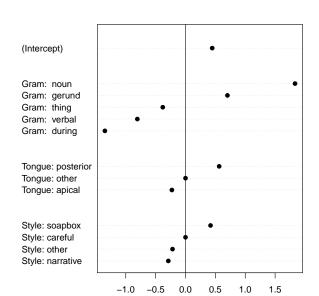
#### Sum

> ing\$Gram2 <- recontrast(ing\$Gram2)</pre>

## Refit

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	0.4470	0.2025	2.21	0.0273
Gram2-during	-1.3495	0.6541	-2.06	0.0391
Gram2-gerund	0.7022	0.2502	2.81	0.0050
Gram2-noun	1.8340	0.2914	6.29	0.0000
Gram2-thing	-0.3805	0.2366	-1.61	0.1077
Tongue-apical	-0.2281	0.1476	-1.54	0.1224
Tongue-posterior	0.5633	0.2768	2.04	0.0418
Style2-narrative	-0.2877	0.1423	-2.02	0.0432
Style2-other	-0.2175	0.3542	-0.61	0.5393
Style2-soapbox	0.4190	0.2090	2.00	0.0450

 $\leftarrow$  More -in ; More -ing  $\rightarrow$ 



### What does it all mean?

The effect of following segment is not usually found to be significant. This is probably due to the evaluation metric used in stepwise regressions like GoldVarb. The non-significance of the following segment effect is sometimes taken to show that the variation is the product of a modular, feed-forward grammar. Have we disproven that here?

#### No!

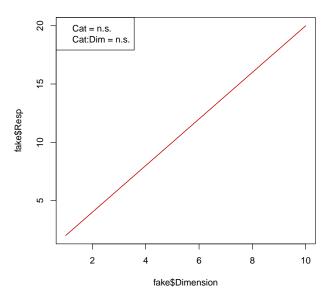
The feed-forward model is specifically a claim about possible interactions of factors. This model has no interactions.

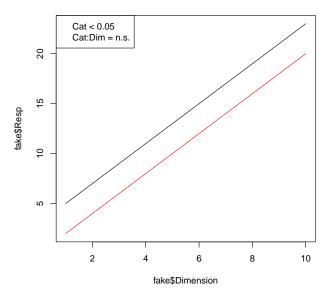
## Interaction Interpretation

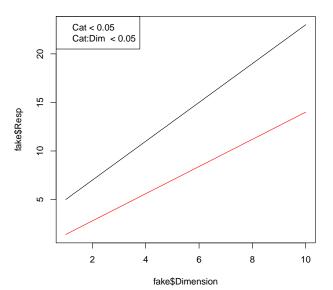
Here is the formula for Categories varying across some Dimension

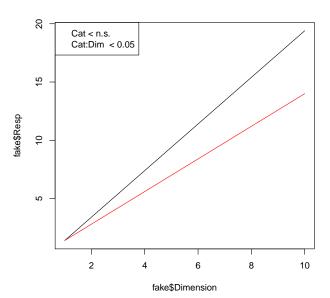
•  $y = Intercept + \beta$ ;  $Cat_X + (\delta * Dim) + (\gamma; Cat_X * Dim) + \epsilon$ Reinterpret it this way

- ightharpoonup y = (Intercept +  $\beta$ ; Cat<sub>X</sub>) + ( $\delta$  +  $\gamma$ ; Cat<sub>X</sub>) \* Dim +  $\epsilon$
- (Intercept  $+ \beta$ ;  $Cat_X$ ) = Category X Intercept
- $(\delta + \gamma; Cat_X)$  = Slope of Category X across Dimension
- if for all Categories, β; Catχ = 0,
   Category intercept for all categories = Intercept
   All Categories have same intercept
- if for all Categories,  $\gamma$ ;  $Cat_X = 0$ , Category slope for all categories =  $\delta$ All Categories have same slope









## Interactions Across Categories

### Interpretation

A significant interaction effect between two categories means that a member from one category's response is modulated by its membership in some other category.

## **Application**

In the simplest cases, we wouldn't expect the effect of a following segment to be modulated by the morphological context in a feed-forward grammar.

## Refitting with an interaction

```
> model5 <- glm(DepVar ~ Gram2 + Tongue +
+ Style2 + Gram2:Tongue, data = ing,
+ family = binomial)</pre>
```

	Df	Deviance	Resid. Df	Resid. Dev	P(> Chi )
NULL			1138	1578.79	
Gram2	4	174.20	1134	1404.59	1.312e-36
Tongue	2	7.76	1132	1396.84	0.0207
Style2	3	11.71	1129	1385.12	0.008438
Gram2:Tongue	6	2.02	1123	1383.11	0.918

#### Result

No significant interaction effect. Simple modularity cannot be ruled out!

## Other Interactions?

What other interesting interactions might be at play? What about with Style?

## Style: Grammatical Status

What if style has a significant interaction with Grammatical Status? This would mean that the /-in/ or /-ing/ morphemes themselves are marked specially for certain styles.

## Style:Tongue

What if style has a significant interaction with Tongue? This would mean that style is some kind of active phonological process.

#### None of the above

What if style has no significant interactions? We can't really conclude anything, but it seems likely to me that this would mean style is some kind of very late, or post-phonological process.

## Style Interactions

#### **Grammatical Status**

```
> model6 <- glm(DepVar ~ Gram2 + Tongue +
+ Style2 + Style2:Gram2, data = ing,
+ family = binomial)</pre>
```

	Df	Deviance	Resid. Df	Resid. Dev	P(> Chi )
NULL			1138	1578.79	
Gram2	4	174.20	1134	1404.59	1.312e-36
Tongue	2	7.76	1132	1396.84	0.0207
Style2	3	11.71	1129	1385.12	0.008438
Gram2:Style2	9	7.11	1120	1378.01	0.6255

#### Result

No significant interaction, and no NWAV paper.

## Style Interactions

#### Tongue

```
> model7 <- glm(DepVar ~ Gram2 + Tongue +
+ Style2 + Style2:Tongue, data = ing,
+ family = binomial)</pre>
```

	Df	Deviance	Resid. Df	Resid. Dev	P(> Chi )
NULL			1138	1578.79	
Gram2	4	174.20	1134	1404.59	1.312e-36
Tongue	2	7.76	1132	1396.84	0.0207
Style2	3	11.71	1129	1385.12	0.008438
Tongue:Style2	6	8.42	1123	1376.70	0.2088

#### Result

Another non-significant interaction, and I'm walking to the conference because travel will not be reimbursed.

## **ING Conclusions**

Results

## Simple Model

- 1. We determined some grammatical equivalencies.
  - (Noun, Adjective); (Participle, Progressive)
- 2. We classified and quantified the phonological effects.
  - Nasal place assimilation.
  - Strongest with following posterior segments.
- 3. "Careful" speech is relatively unmarked.

## **ING Conclusions**

Results

#### Interactions

- 1. Non-significant Gram: Tongue
  - The phonological process doesn't utilize morphological information
  - The morphological process doesn't utilize phonological information
- 2. Non-significant Style:Gram
  - Style doesn't interact with the morphology.
- 3. Non-significant Style:Tongue
  - Style doesn't interact with the phonology

## **ING Conclusions**

Philosophical Questions

#### Was it all worth it?

Yes. We determined a few equivalencies with regards to the variable, and fit cleaner models. Then, with the use of interactions, we asked a few very interesting questions.

#### Could I have done it without R?

Probably, in some other statistics package. *I* couldn't have done it as quickly and clearly, or in a replicable manner, but that's not to say someone else couldn't.

### Could I have done it with GoldVarb?

I could not have asked these questions in GoldVarb.

### TD

There is a *lot* more work that can be done with this variable than I can do here. This might be surprising given the huge size of the literature on the topic. However, most of the work currently still done on this variable is usually done by very intelligent researchers shackled to very inadequate statistical devices. This section merely replicates the discussion here.

### TD

#### The Variable

TD is the name given to consonant cluster simplification in English. The context could be schematized as

$$\left\{\begin{array}{c}t\\d\end{array}\right\}\to\emptyset\ /\ \mathsf{C}\_\ ]_{\sigma}$$

It has been demonstrated to be affected both by morphological and phonological contexts.

#### The Data

The data I use here was automatically extracted from the Buckeye Corpus. I have 8978 tokens in total. A subset of that data, coded for Grammatical Class and log(Frequency) of each word in the corpus, is available here:

 $http://www.ling.upenn.edu/{\sim}joseff/rstudy/data/sbuck.csv$ 

## TD

## The Analysis

I'll step through a clearer case of an analysis that requires use and interpretation of interaction effects.

#### Order of Models

- 1. Slope
- 2. Category Effects
- 3. Interaction Effects
- 4. Refinement Across Subsets
- 5. Final Refined Model

## Model 1: Slope

First, we'll determine whether -td deletion is sensitive to frequency.

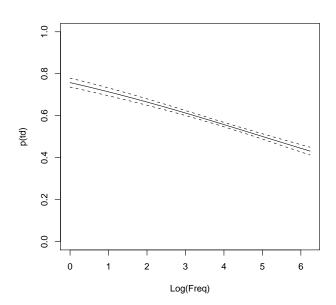
```
> model1 <- glm(DepVar ~ Log_Freq, data = td,
+ family = binomial)</pre>
```

	Df	Deviance	Resid. Df	Resid. Dev	P(> Chi )
NULL			8977	12303.55	
Log_Freq	1	276.86	8976	12026.68	3.622e-62

#### Result

There is a significant effect of Log(Frequency) on TD.

## Model 1: Slope



## Model 2: Category Effects

Next, we'll see if grammatical class has a uniform upward or downward effect on deletion.

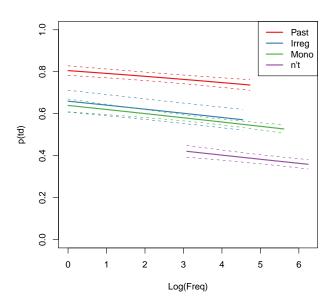
```
> model2 <- glm(DepVar ~ Gram + Log_Freq,
+ data = td, family = binomial)</pre>
```

	Df	Deviance	Resid. Df	Resid. Dev	P(> Chi )
NULL			8977	12303.55	
Gram	3	572.12	8974	11731.43	1.113e-123
Log_Freq	1	25.87	8973	11705.55	3.647e-07

#### Result

At least one grammatical class has a different intercept from the others

## Model 2: Category Effects



### Model 3: Interaction effects

Do all grammatical classes react to frequency the same way?

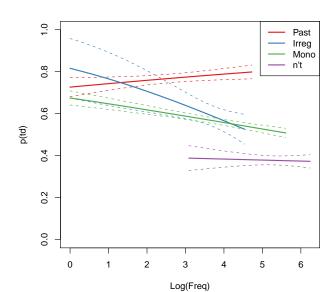
	Df	Deviance	Resid. Df	Resid. Dev	P(> Chi )
NULL			8977	12303.55	
Gram	3	572.12	8974	11731.43	1.113e-123
Log_Freq	1	25.87	8973	11705.55	3.647e-07
$Gram:Log_Freq$	3	27.37	8970	11678.18	4.927e-06

#### Result

There is a significant interaction effect between Grammatical Class and Frequency!

## Model 3: Interaction Effects

Wow!



### Model 3: Interaction Effects

Let's take a look at the interactions here

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	0.7005	0.1447	4.84	1.3e-06
GramPast	0.2756	0.1664	1.66	0.09765
Gramlrreg	0.7890	0.3703	2.13	0.03314
${\sf GramMono}$	0.0292	0.1549	0.19	0.8504
$Log_{F}Freq$	-0.0915	0.0360	-2.54	0.01101
GramPast:Log_Freq	0.1764	0.0455	3.88	0.0001041
GramIrreg:Log_Freq	-0.2143	0.0962	-2.23	0.02588
GramMono:Log_Freq	-0.0331	0.0385	-0.86	0.3897
GramMono:Log_Freq	-0.0331	0.0385	-0.86	0.3897

#### Result

It looks like the slope of the past tense over frequency is significantly positive!

### Refinement Across Subsets

The only way to see if the slopes of Monomorphemes, Irregular Past Tense and -n't clitics are parallel or not is to fit a model on a subset of the data, excluding the past tense.

```
> subTd <- subset(td, Gram != "Past")
> subTd <- drop.levels(subTd)
> subTd$Gram <- relevel(subTd$Gram, "Mono")
> subTd$Gram <- recontrast(subTd$Gram, "treatment")</pre>
```

#### Note

The function drop.levels() comes from a special package called gdata.

## Model 4: Refinement Across Subsets

```
> model4 <- glm(DepVar ~ Gram + Log_Freq +
+ Gram:Log_Freq, data = subTd, family = binomial)</pre>
```

	Df	Deviance	Resid. Df	Resid. Dev	P(> Chi )
NULL			7335	10161.33	
Gram	2	208.93	7333	9952.41	4.289e-46
Log_Freq	1	42.89	7332	9909.52	5.786e-11
Gram:Log_Freq	2	5.69	7330	9903.83	0.05811

#### Result

The interaction is non-significant.

## Model 5: Final Refined Model

It appears as if there is one important division between grammatical classes, Past Tense, and Other. Let's create a new factor group to reflect that:

```
> td$Gram2 <- "other"
> td[td$Gram == "Past", ]$Gram2 <- "past"
> td$Gram2 <- as.factor(td$Gram2)</pre>
```

#### **Subdivisions**

There may also be significant subdivisions within the "Other" class. I'm particularly interested to see if the irregular past or n't behave differently from monomorphemes, so I'll relevel, and recontast the Gram factor group to reflect this.

```
> td$Gram <- relevel(td$Gram, "Mono")
> td$Gram <- recontrast(td$Gram, "treatment")</pre>
```

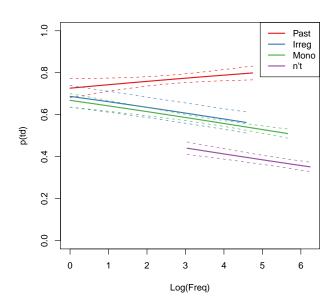
### Model 5: Final Refined Model

```
> model5 <- glm(DepVar ~ Gram2 + Gram2:Gram +
+ Gram2:Log_Freq - 1, data = td, family = binomial)</pre>
```

	Estimate	Std. Error	z value	Pr(> z )
Gram2other	0.7005	0.0732	9.57	1.107e-21
Gram2past	0.9761	0.1160	8.41	3.969e-17
Gram2other:GramIrreg	0.0871	0.1068	0.82	0.4148
Gram2other:Gramn't	-0.5853	0.0613	-9.54	1.365e-21
Gram2other:Log_Freq	-0.1168	0.0180	-6.51	7.688e-11
Gram2past:Log_Freq	0.0849	0.0393	2.16	0.03061

- ► Gram2 Intercepts for Past and Other
- Gram2:Gram Difference of intercepts within Other
- ► Gram2:Log(Freq) Slopes of Past and Other

## Model 5: Final Refined Model



### **TD Conclusions**

I really only have more questions after this analysis

- What does this tell us about the morphophonological complexity of the regular past tense?
- ▶ Why is the regular past's slope *positive*?
- Why is the irregular past tense's slope negative?
- What causes the stable differentiation of classes in the "other" category across all frequencies?

Was it worth it? Yes

Could I have done it without R? Maybe

Could I have done it with GoldVarb? No way.

## The End

Thanks!

### Selected References



Guy, Gregory R. 1980.

Variation in the group and the individual: the case of final stop deletion.

In Locating Language in Time and Space, ed. William Labov, 1-36. New York: Academic Press.



Labov. William, 1989.

The child as linguistic historian.

Language Variation and Change 1:85–97.



Pitt, M. A., L. Dilley, Keith Johnson, S. Kiesling,

W. Raymond, Elizabeth Hume, and E. Fosler-Lussier. 2007.

Buckeye corpus of conversational speech (2nd release).

www.buckeyecorpus.osu.edu.

### Selected References



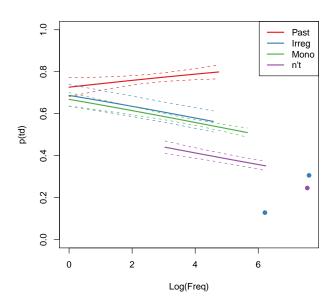
R Development Core Team. 2009.

R: A Language and Environment for Statistical Computing.

R Foundation for Statistical Computing, Vienna, Austria.

http://www.R-project.org, ISBN 3-900051-07-0.

## With Outliers



## 3 Slopes

## > xtable(model4)

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	0.7298	0.0781	9.35	0.0000
GramIrreg	0.7597	0.4884	1.56	0.1198
Gramn't	-1.1230	0.2989	-3.76	0.0002
$Log\_Freq$	-0.1246	0.0193	-6.45	0.0000
GramIrreg:Log_Freq	-0.1812	0.1276	-1.42	0.1557
Gramn't:Log_Freq	0.1041	0.0572	1.82	0.0688

## 3 Slopes

```
> td$Gram2 <- "im"
> td[td$Gram == "Past", ]$Gram2 <- "past"
> td[td$Gram == "n't", ]$Gram2 <- "n't"
> td$Gram2 <- as.factor(td$Gram2)</pre>
```

## 3 Slopes

```
> model6 <- glm(DepVar ~ Gram2 + Gram2:Gram +
+ Gram2:Log_Freq - 1, data = td, family = binomial)</pre>
```

	Estimate	Std. Error	z value	Pr(> z )
Gram2im	0.7461	0.0773	9.65	0.0000
Gram2n't	-0.3932	0.2885	-1.36	0.1730
Gram2past	0.9761	0.1160	8.41	0.0000
Gram2im:GramIrreg	0.0864	0.1068	0.81	0.4187
Gram2im:Log_Freq	-0.1290	0.0191	-6.75	0.0000
Gram2n't:Log_Freq	-0.0205	0.0538	-0.38	0.7031
Gram2past:Log_Freq	0.0849	0.0393	2.16	0.0306

## Model 6: 3 Slopes

