Supplementary semantic analyses for Vu et al.

Jeff Phillips (jeffrey.s.phillips@gmail.com)

9/26/2015

1 Overview

The Special Issue editors have asked us to provide evidence that we are classifying based on semantic features; they suggest analysis of the confusion matrix, to see whether the model's errors tend to be semantic in nature. Here I correlated the likelihood (across 10 iterations) of misclassification for each word pair. I then regressed this misclassification rate on the word pairs' semantic similarity, which was calculated as the correlation of feature vectors derived from our team's Mechanical Turk ratings experiment. A few details:

- Used 239 of 261 words: words like "The","an","through" dropped due to unavailability of Turk ratings.
- Turk data comprise continuous human ratings on a scale of 0–100 for 21 features:

Category	Features
Perceptual	body, building, color, face, motion, shape, smell, sound, touch, taste
Motor	action, manipulation
Abstract	abstract, composite, emotion, freedom, intent, natural, quantity, social, time

- Word pair similarity: Pearson's correlation for two 21-feature vectors, with Fisher's R-to-z transformation
- Linear regression to test for association between similarity (z-score) and cross-classification (misclassification) rate
- Included interaction of misclassification rate and model (static or dynamic)
- Misclassification rate expressed as a proportion of total misclassifications for each item. This measure is thus independent of mean classification accuracy for a given word, which may be confounded by factors such as frequency of occurrence in the Phase1a stimulus set.

2 Results

2.1 Semantic similarity and misclassification rate: 3T scanner

```
##
## Call:
## lm(formula = misclassnorm ~ fz * model, data = summ)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.637 -0.411 -0.071  0.250 14.288
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.09012    0.00378   -23.9   <2e-16 ***
## fz            0.41952    0.00862    48.7   <2e-16 ***
## modelstat     0.08982    0.00534    16.8   <2e-16 ***
## fz:modelstat -0.41810    0.01219   -34.3   <2e-16 ***
```

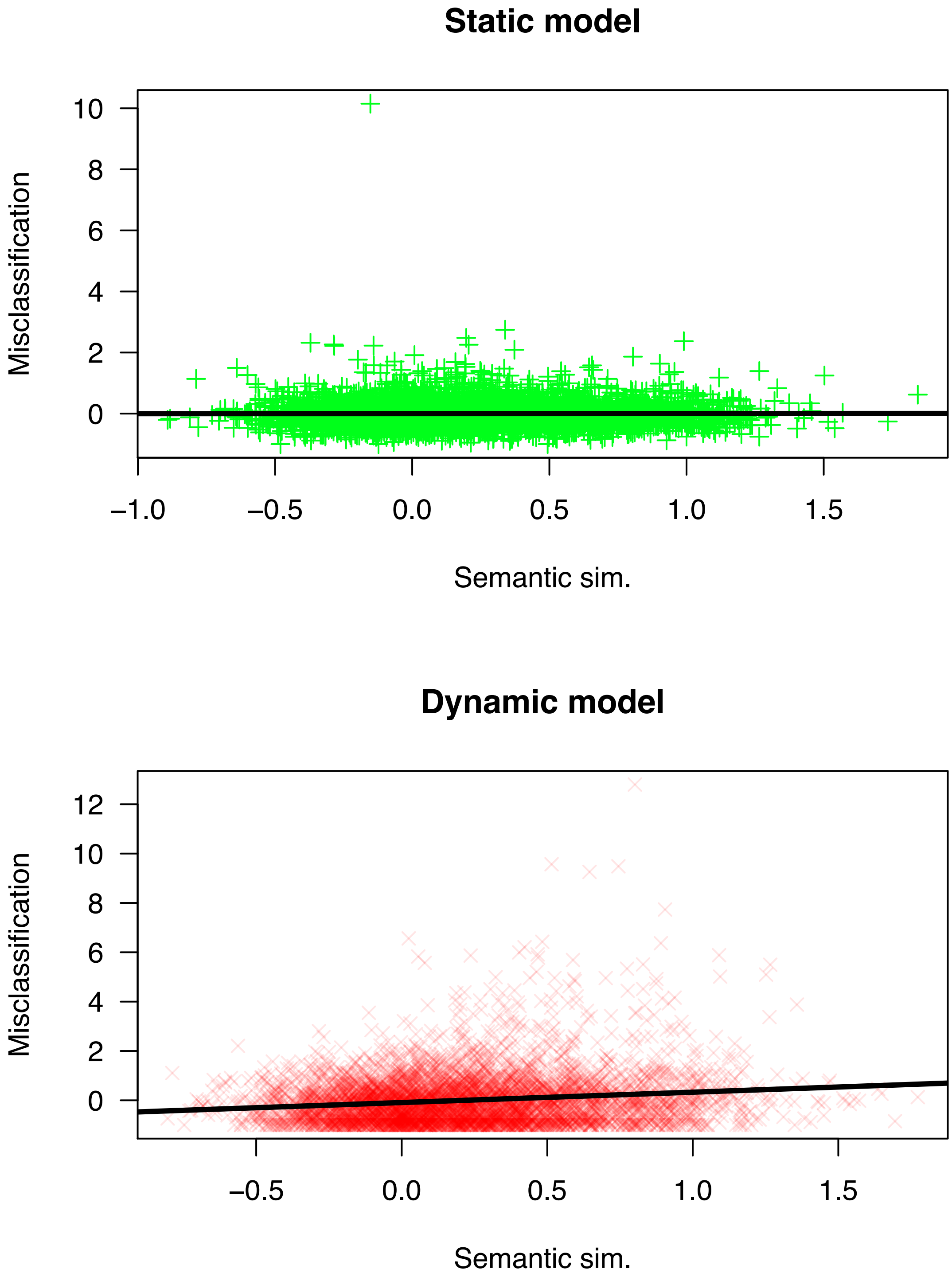


Figure 1: Cross-classification and NYU-AMT semantic features: 3T

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.785 on 113760 degrees of freedom
## Multiple R-squared:  0.0204, Adjusted R-squared:  0.0204
## F-statistic:  790 on 3 and 113760 DF,  p-value: <2e-16
```

```
## Static model: correlation of semantic similarity and misclassification rate
```

```
##
##  Pearson's product-moment correlation
##
## data:  summ$fz[summ$model == "stat"] and summ$misclassnorm[summ$model == "stat"]
## t = 0.32, df = 57000, p-value = 0.7
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.006856  0.009579
## sample estimates:
##      cor
## 0.001362
```

```
## Dynamic model: correlation of semantic similarity and misclassification rate
```

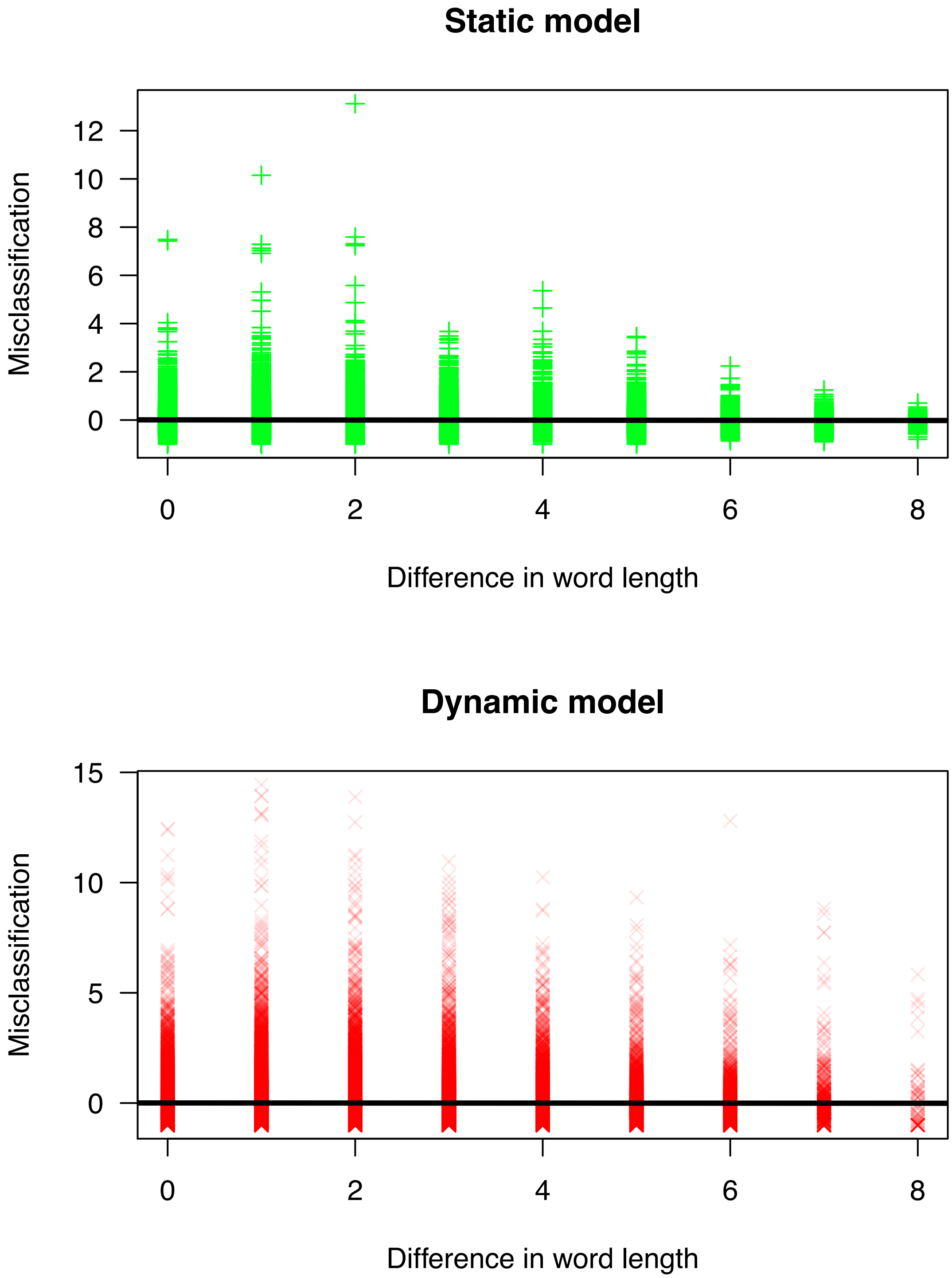
```
##
##  Pearson's product-moment correlation
##
## data:  summ$fz[summ$model == "dyn"] and summ$misclassnorm[summ$model == "dyn"]
## t = 37, df = 57000, p-value <2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1449 0.1610
## sample estimates:
##    cor
## 0.1529
```

- Static model has a significantly higher misclassification rate.
- No correlation of misclassification rate and semantic similarity in the static model.
- Misclassification in the dynamic model is significantly associated with semantic similarity (R=0.15).

2.2 Word length difference and misclassification rate: 3T scanner

For comparison, does difference in word length (a perceptual variable) predict cross-classification?

```
##
## Call:
## lm(formula = misclassnorm ~ diffchar * model, data = summ)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.006 -0.414 -0.072  0.245 14.443
##
```


Static model

Dynamic model

Figure 2: Cross-classification and word-length differences: 3T

```
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         0.00343    0.00540    0.63     0.53
## diffchar           -0.00168    0.00209   -0.81     0.42
## modelstat           0.00302    0.00764    0.40     0.69
## diffchar:modelstat -0.00148    0.00295   -0.50     0.62
##
## Residual standard error: 0.793 on 113760 degrees of freedom
## Multiple R-squared:  2.59e-05,   Adjusted R-squared:  -4.56e-07
## F-statistic: 0.983 on 3 and 113760 DF,  p-value: 0.4
```

```
##
##  Pearson's product-moment correlation
##
## data:  summ$diffchar[summ$model == "stat"] and summ$misclassnorm[summ$model == "stat"]
## t = -3, df = 57000, p-value = 0.003
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.020806 -0.004373
## sample estimates:
##      cor
## -0.01259
```

```
##
##  Pearson's product-moment correlation
##
## data:  summ$diffchar[summ$model == "dyn"] and summ$misclassnorm[summ$model == "dyn"]
## t = -0.61, df = 57000, p-value = 0.5
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.01078  0.00566
## sample estimates:
##       cor
## -0.002558
```

- No overall effect of word length, independent of mean item-wise classification accuracy.
- Post-hoc Pearson's correlation test suggests there could be a small effect of word length in the static model results (R=-0.01), but not in the dynamic model.

2.3 Semantic similarity and misclassification rate: 7T scanner

```
##
## Call:
## lm(formula = misclassnorm ~ fz * model, data = summ)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -1.65  -0.51  -0.09   0.27  38.96
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.90122    0.00446   202.2   <2e-16 ***
```

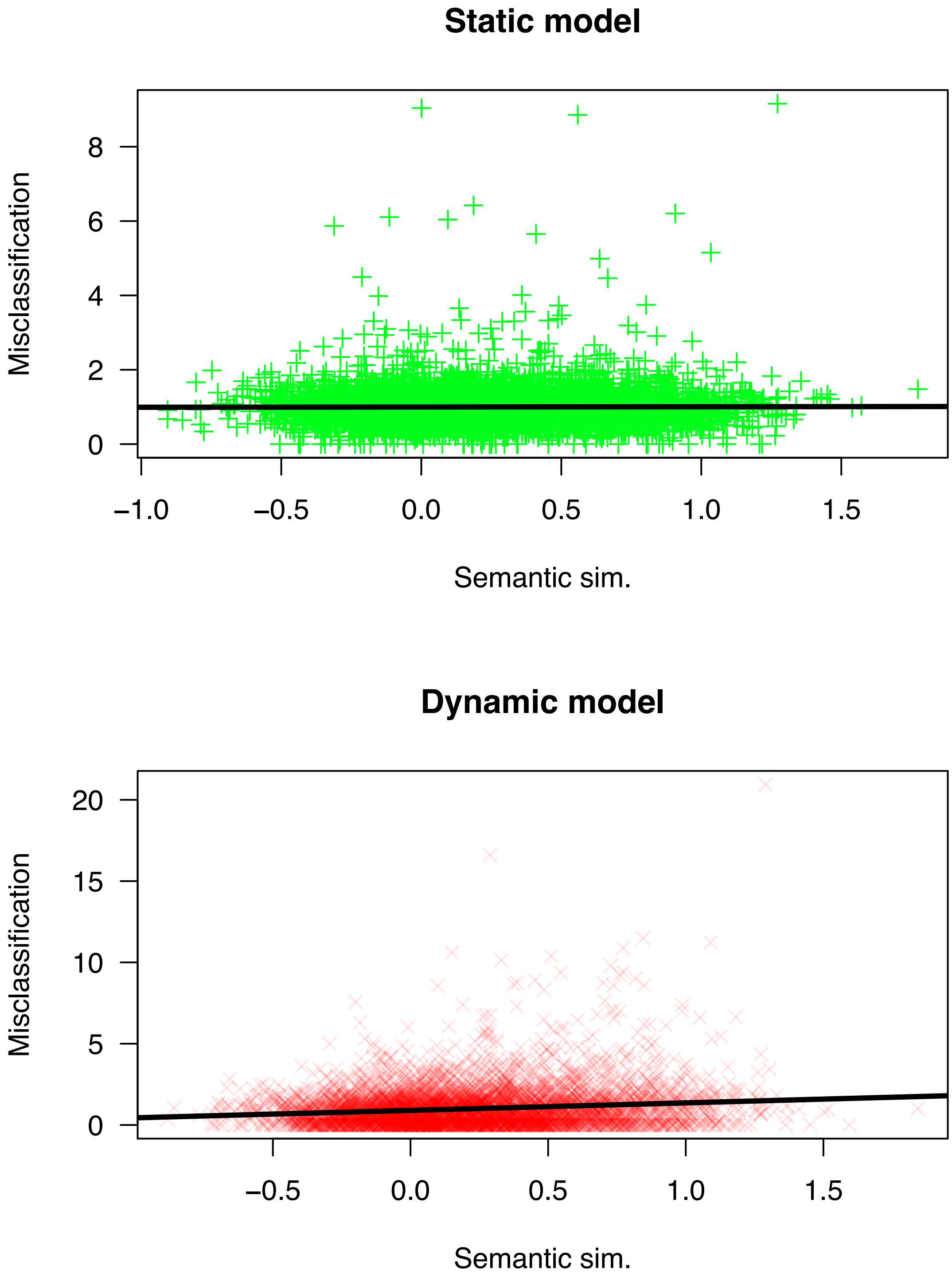


Figure 3: Cross-classification and NYU-AMT semantic features: 7T

```
## fz            0.45981    0.01017    45.2   <2e-16 ***
## modelstat     0.09731    0.00630    15.4   <2e-16 ***
## fz:modelstat -0.45296    0.01439   -31.5   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.927 on 113760 degrees of freedom
## Multiple R-squared:  0.0176, Adjusted R-squared:  0.0176
## F-statistic:  681 on 3 and 113760 DF,  p-value: <2e-16
```

```
## Static model: correlation of semantic similarity and misclassification rate
```

```
##
##  Pearson's product-moment correlation
##
## data:  summ$fz[summ$model == "stat"] and summ$misclassnorm[summ$model == "stat"]
## t = 1.3, df = 57000, p-value = 0.2
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.002947  0.013489
## sample estimates:
##      cor
## 0.005271
```

```
## Dynamic model: correlation of semantic similarity and misclassification rate
```

```
##
##  Pearson's product-moment correlation
##
## data:  summ$fz[summ$model == "dyn"] and summ$misclassnorm[summ$model == "dyn"]
## t = 35, df = 57000, p-value <2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1353 0.1514
## sample estimates:
##    cor
## 0.1433
```

- Static model has a significantly higher misclassification rate.
- No correlation of misclassification rate and semantic similarity in the static model.
- Misclassification in the dynamic model is significantly associated with semantic similarity (R=0.15).

2.4 Word length difference and misclassification rate: 7T scanner

For comparison, does difference in word length (a perceptual variable) predict cross-classification?

```
##
## Call:
## lm(formula = misclassnorm ~ diffchar * model, data = summ)
##
## Residuals:
```

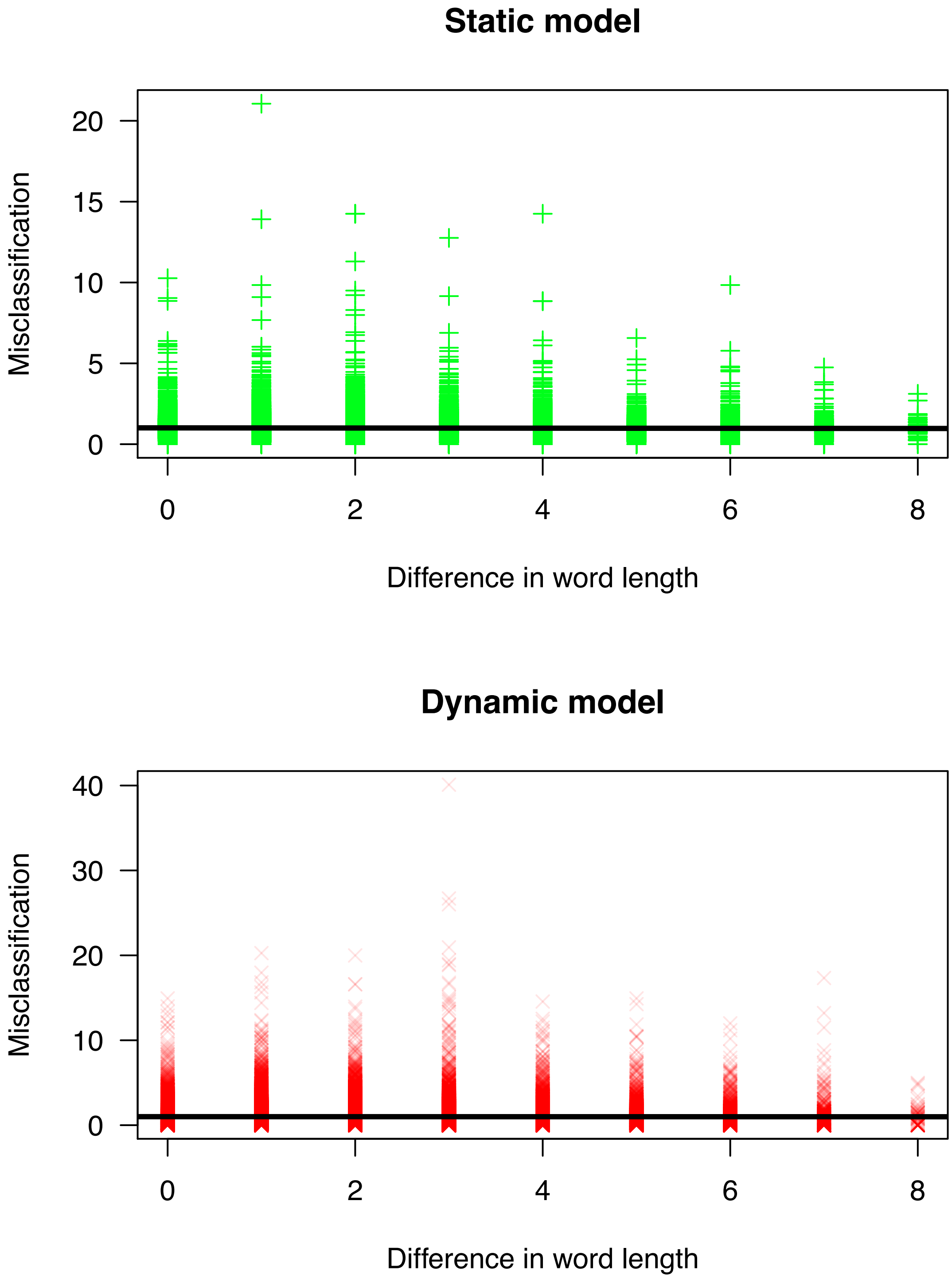


Figure 4: Cross-classification and word-length differences: 7T

```
##    Min     1Q Median     3Q    Max
##  -1.01  -0.51  -0.09   0.27  39.10
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.00300    0.00637  157.58   <2e-16 ***
## diffchar           -0.00147    0.00246   -0.60     0.55
## modelstat           0.00618    0.00900    0.69     0.49
## diffchar:modelstat -0.00303    0.00348   -0.87     0.38
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.935 on 113760 degrees of freedom
## Multiple R-squared:  3.26e-05,   Adjusted R-squared:  6.22e-06
## F-statistic: 1.24 on 3 and 113760 DF,  p-value: 0.295
```

```
##
##  Pearson's product-moment correlation
##
## data:  summ$diffchar[summ$model == "stat"] and summ$misclassnorm[summ$model == "stat"]
## t = -3.4, df = 57000, p-value = 6e-04
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.022675 -0.006242
## sample estimates:
##      cor
## -0.01446
```

```
##
##  Pearson's product-moment correlation
##
## data:  summ$diffchar[summ$model == "dyn"] and summ$misclassnorm[summ$model == "dyn"]
## t = -0.46, df = 57000, p-value = 0.6
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.010130  0.006306
## sample estimates:
##       cor
## -0.001912
```

3 Conclusions

1. Error analysis indicates that misclassifications are partially attributable to the semantic similarity of words, particularly in the results from the dynamic model analysis.
2. In the dynamic model analysis, errors are not attributable to similar word length between pairs of confused words, a perceptual variable raised as a possible confound by reviewers. There is a modest association with word length in the static model analysis.
3. Semantic similarity effects are comparable between the 3T (n=3) and 7T (n=2) datasets with a 500 ms TR.