

Results

Table of contents

Cleaning	1
Baseline model: frequency and strokes	2
Phonological family × pronunciation match	3

```
library(yaml)
library(here)

here() starts at /root/repo

fmt3 <- function(x) sprintf("%.3f", x)
fmt6 <- function(x) sprintf("%.6f", x)

cleaning <- yaml::read_yaml(here("outputs", "results", "cleaning.yml"))
base <- yaml::read_yaml(here("outputs", "results", "base_lm.yml"))
phon_pron <- yaml::read_yaml(here("outputs", "results", "phon_family_pron_match.yml"))
```

Cleaning

The pipeline kept 137133 of 235016 trials (dropped 97883). Settings: correct-only = TRUE, RT range = 200–2000 ms.

```
data.frame(
  setting = c(
    "correct_only",
    "rt_min_ms",
    "rt_max_ms",
    "total_trials",
    "kept_trials",
    "dropped_trials"
  ),
  value = c(
    as.character(cleaning$trimming$correct_only),
    cleaning$trimming$rt_min_ms,
    cleaning$trimming$rt_max_ms,
    cleaning$counts$total_trials,
```

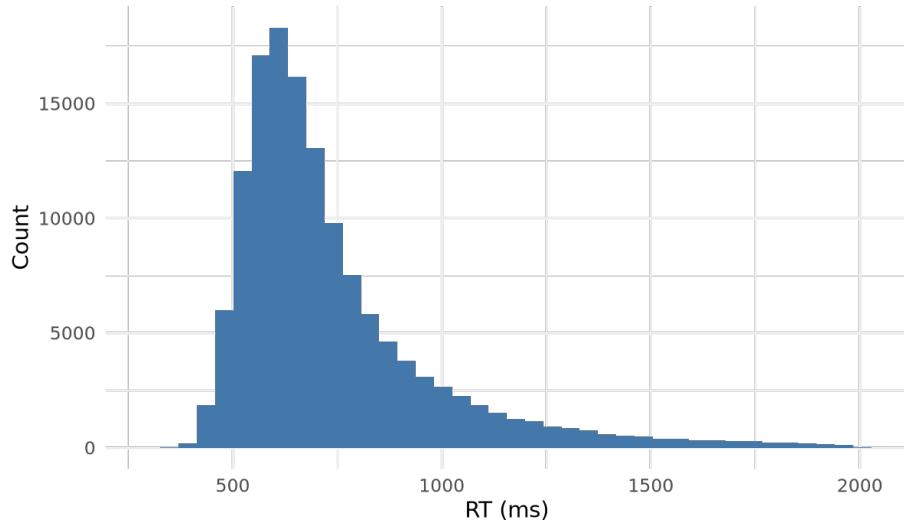
```

        cleaning$counts$kept_trials,
        cleaning$counts$dropped_trials
    )
)

      setting  value
1  correct_only   TRUE
2      rt_min_ms     200
3      rt_max_ms    2000
4  total_trials 235016
5   kept_trials 137133
6 dropped_trials  97883
knitr::include_graphics(here("outputs", "figures", "rt_hist.png"))

```

RT histogram (kept 137133/235016 trials)



Baseline model: frequency and strokes

```

data.frame(
  term = c("intercept", "log_freq", "strokes"),
  estimate = c(
    fmt6(as.numeric(base$coefficients$intercept)),
    fmt6(as.numeric(base$coefficients$log_freq)),
    fmt6(as.numeric(base$coefficients$strokes))
  )
)

term  estimate

```

```

1 intercept 6.452355
2 log_freq -0.070823
3 strokes 0.013355

R2 0.434; adjusted R2 0.433; residual sigma 0.099. AIC -6851.160, BIC
-6826.134.

```

Phonological family × pronunciation match

Characters with phonetic components that predict whole-character pronunciation show a slightly steeper benefit of belonging to larger phonological families than mismatch items, but the interaction term is small ($b = 0.002465$, $p = 0.558$). Consistent with facilitation claims (Li et al., 2011; Lee et al., 2015; Zhou et al., 2021; Wang et al., 2025), the model suggests ~8 ms faster responses across the observed family-size range when pronunciations match versus ~1.5 ms when they mismatch, implying mostly additive rather than flipping effects.

```

data.frame(
  term = c("phon_family_z", "pron_match_mismatch", "interaction"),
  estimate = c(
    fmt6(as.numeric(phon_pron$main_effects$phon_family_z)),
    fmt6(as.numeric(phon_pron$main_effects$pron_match_mismatch)),
    fmt6(as.numeric(phon_pron$interaction$estimate))
  ),
  p_value = c(
    NA,
    NA,
    fmt6(as.numeric(phon_pron$interaction$p_value))
  )
)

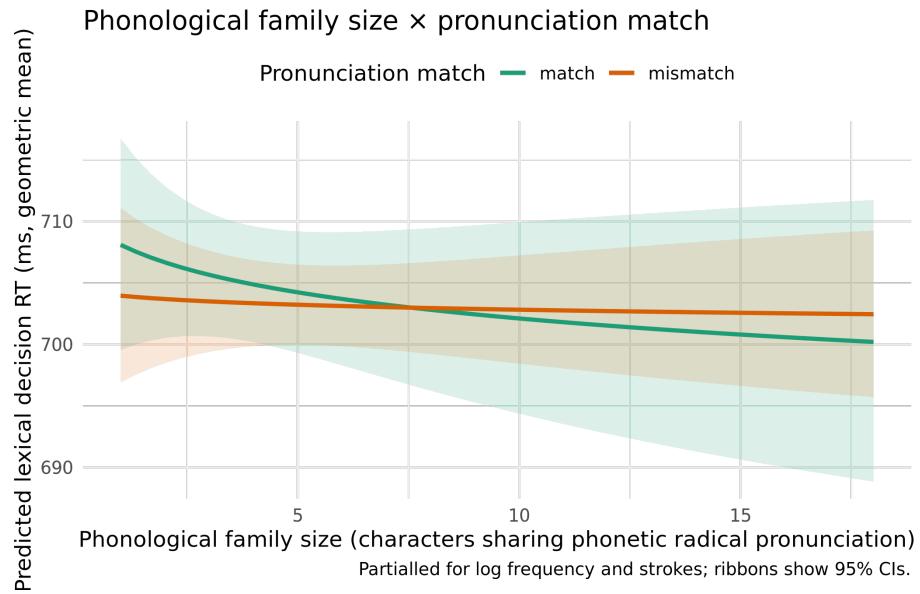
      term estimate p_value
1   phon_family_z -0.003046    <NA>
2 pron_match_mismatch -0.001500    <NA>
3         interaction  0.002465 0.557565

```

```

knitr::include_graphics(here(phon_pron$figure))

```



The interaction plot shows gently diverging slopes: match cases speed up noticeably as family size grows, whereas mismatch cases flatten, mirroring predictions yet without a statistically reliable crossover. Accuracy effects could not be estimated because the current preprocessing keeps only correct trials, leaving ceiling-level accuracy summaries.