# concepticon-analysis

```
library(knitr)
library(tidyboot)
library(tidyverse)
library(childesr)
library(here)
library(reticulate)
use_condaenv("r-reticulate")
theme_set(theme_classic(base_size = 16))
opts_chunk$set(echo = TRUE, message = FALSE, warning = FALSE,
               error = FALSE, cache = TRUE, tidy = FALSE)
childes_read_dir = here("model/trained_models/childes_adult_word2vec.model")
coca_read_dir = here("model/trained_models/coca_word2vec.model")
```

```
types <- get_types(collection = "Eng-NA")

parent_types <- types %>%
  mutate(gloss = str_to_lower(gloss)) %>%
  filter(speaker_role == "Mother" | speaker_role == "Father") %>%
  group_by(gloss) %>%
  summarise(sum = sum(count))
```

```
from gensim import utils
import gensim.models
import gensim.models.word2vec
from gensim.test.utils import datapath
childes_model = gensim.models.Word2Vec.load(r.childes_read_dir)
coca_model = gensim.models.Word2Vec.load(r.coca_read_dir)
coca_vocabulary = coca_model.wv.vocab
```

```
set.seed(111)

# NOTE: excluded "shortest" because it doesn't happen enough in childes to have a representation
childes_target_words = c("long","longer","longest","short","shorter")
coca_target_words = c("long","longer","longest","short","shorter","shortest")
```

```
coca_vocab <- unlist(py$coca_vocabulary) %>% as.list()
coca_vocab <- names(coca_vocab)
coca_vocab <- as_tibble(coca_vocab)

#coca_vocab <- py$coca_vocabulary %>% as.data.frame()

colnames(coca_vocab) = c("word")

childes_random_words <- read_csv(here("data/childes_random_words.csv")) %>%
  left_join(parent_types, by = c("word" = "gloss")) %>%
  filter(sum >= 50, !(word %in% coca_target_words), word %in% coca_vocab$word) %>%
  sample_n(500)
```

```r
concepticon_list <- read_csv(here("data/concepticon_words.csv")) %>%
  mutate(exclude = if_else(is.na(exclude), FALSE, TRUE),
         concept = str_to_lower(concept),
         concept = str_replace(concept, " \\(.*\\)", ""),
         domain = if_else(SemanticField == "Time", "time", "space"))

concepticon_list <- concepticon_list %>%
  left_join(parent_types, by = c("concept" = "gloss")) %>%
  rename(parent_tokens = sum) %>%
  filter(exclude == FALSE) %>%
  select(domain, concept, parent_tokens)

top_concepts <- concepticon_list %>%
  arrange(desc(parent_tokens)) %>%
  group_by(domain) %>% slice(1:20)

childes_pairs <- cross(list(top_concepts$concept, childes_target_words))
childes_random_pairs <- cross(list(top_concepts$concept, childes_random_words$word))
coca_pairs <- cross(list(top_concepts$concept, coca_target_words))
coca_random_pairs <- cross(list(top_concepts$concept, childes_random_words$word))
space_time_pairs <- combn(top_concepts$concept, 2, simplify = FALSE)
```

```python
from gensim import utils
import gensim.models
import gensim.models.word2vec
from gensim.test.utils import datapath
childes_model = gensim.models.Word2Vec.load(r.childes_read_dir)
coca_model = gensim.models.Word2Vec.load(r.coca_read_dir)
childes_dict = {}
childes_random_dict = {}
coca_dict = {}
coca_random_dict = {}
space_time_childes_dict = {}
space_time_coca_dict = {}

for word, target_word in r.childes_random_pairs:
   childes_random_dict[word + " " + target_word] = childes_model.wv.similarity(word, target_word)

for word, target_word in r.childes_pairs:
   childes_dict[word + " " + target_word] = childes_model.wv.similarity(word, target_word)

for word, target_word in r.coca_random_pairs:
   coca_random_dict[word + " " + target_word] = coca_model.wv.similarity(word, target_word)

for word, target_word in r.coca_pairs:
   coca_dict[word + " " + target_word] = coca_model.wv.similarity(word, target_word)

for word, target_word in r.space_time_pairs:
   space_time_childes_dict[word + " " + target_word] = childes_model.wv.similarity(word, target_word)
   space_time_coca_dict[word + " " + target_word] = coca_model.wv.similarity(word, target_word)
```

```r
childes_sims <- py$childes_dict %>% unlist() %>% as.list() %>% as_tibble() %>%
  t() %>% as_tibble(rownames = "name")



#
colnames(childes_sims) = c("name", "distance")
#
childes_sims <- childes_sims %>%
  mutate(word = gsub(" .*$", "", name), target_word = gsub(".* ", "", name)) %>%
  select(word, target_word, distance) %>%
  left_join(top_concepts %>% select(domain, concept), by = c("word" = "concept"))
```

```r
childes_random_sims <- py$childes_random_dict %>% unlist() %>% as.list() %>% as_tibble() %>%
  t() %>% as_tibble(rownames = "name") %>%
  dplyr::rename(distance = V1) %>%
  mutate(word = gsub(" .*$", "", name), target_word = gsub(".* ", "", name)) %>%
  select(word, target_word, distance) %>%
  left_join(top_concepts %>% select(domain, concept), by = c("word" = "concept"))

coca_sims <- py$coca_dict %>%unlist() %>% as.list() %>% as_tibble() %>%
  t() %>% as_tibble(rownames = "name") %>%
  dplyr::rename(distance = V1) %>%
  mutate(word = gsub(" .*$", "", name), target_word = gsub(".* ", "", name)) %>%
  select(word, target_word, distance) %>%
  left_join(top_concepts %>% select(domain, concept), by = c("word" = "concept"))

coca_random_sims <- py$coca_random_dict %>%unlist() %>% as.list() %>% as_tibble() %>%
  t() %>% as_tibble(rownames = "name") %>%
  dplyr::rename(distance = V1) %>%
  mutate(word = gsub(" .*$", "", name), target_word = gsub(".* ", "", name)) %>%
  select(word, target_word, distance) %>%
  left_join(top_concepts %>% select(domain, concept), by = c("word" = "concept"))

st_childes <- py$space_time_childes_dict %>% unlist() %>% as.list() %>% as_tibble() %>%
  t() %>% as_tibble(rownames = "name") %>%
  dplyr::rename(distance = V1) %>%
  mutate(word = gsub(" .*$", "", name), target_word = gsub(".* ", "", name)) %>%
  select(word, target_word, distance) %>%
  left_join(top_concepts %>% select(domain, concept), by = c("word" = "concept")) %>%
  rename("word_domain" = "domain") %>%
  left_join(top_concepts %>% select(domain, concept), by = c("target_word" = "concept")) %>%
  rename("target_word_domain" = "domain") %>%
  mutate(within_between = case_when(word_domain == "time" & target_word_domain == "time"
                                    ~ "within_time",
                                    word_domain == "space" & target_word_domain == "space"
                                    ~ "within_space",
                                    word_domain != target_word_domain ~ "between_space_time"
                                    ))

st_coca <- py$space_time_coca_dict %>% unlist() %>% as.list() %>% as_tibble() %>%
  t() %>% as_tibble(rownames = "name") %>%
  dplyr::rename(distance = V1) %>%
  mutate(word = gsub(" .*$", "", name), target_word = gsub(".* ", "", name)) %>%
```

```
  select(word, target_word, distance) %>%
  left_join(top_concepts %>% select(domain, concept), by = c("word" = "concept")) %>%
  rename("word_domain" = "domain") %>%
  left_join(top_concepts %>% select(domain, concept), by = c("target_word" = "concept")) %>%
  rename("target_word_domain" = "domain") %>%
  mutate(within_between = case_when(word_domain == "time" & target_word_domain == "time"
                                    ~ "within_time",
                                    word_domain == "space" & target_word_domain == "space"
                                    ~ "within_space",
                                    word_domain != target_word_domain ~ "between_space_time"
                                    ))

between_within_coca <- st_coca %>%
  group_by(within_between) %>%
  tidyboot_mean(distance) %>%
  mutate(domain = case_when(within_between == "within_time" ~ "time",
                            within_between == "within_space" ~ "space"))

between_within_childes <- st_childes %>%
  group_by(within_between) %>%
  tidyboot_mean(distance) %>%
  mutate(domain = case_when(within_between == "within_time" ~ "time",
                            within_between == "within_space" ~ "space"))
```

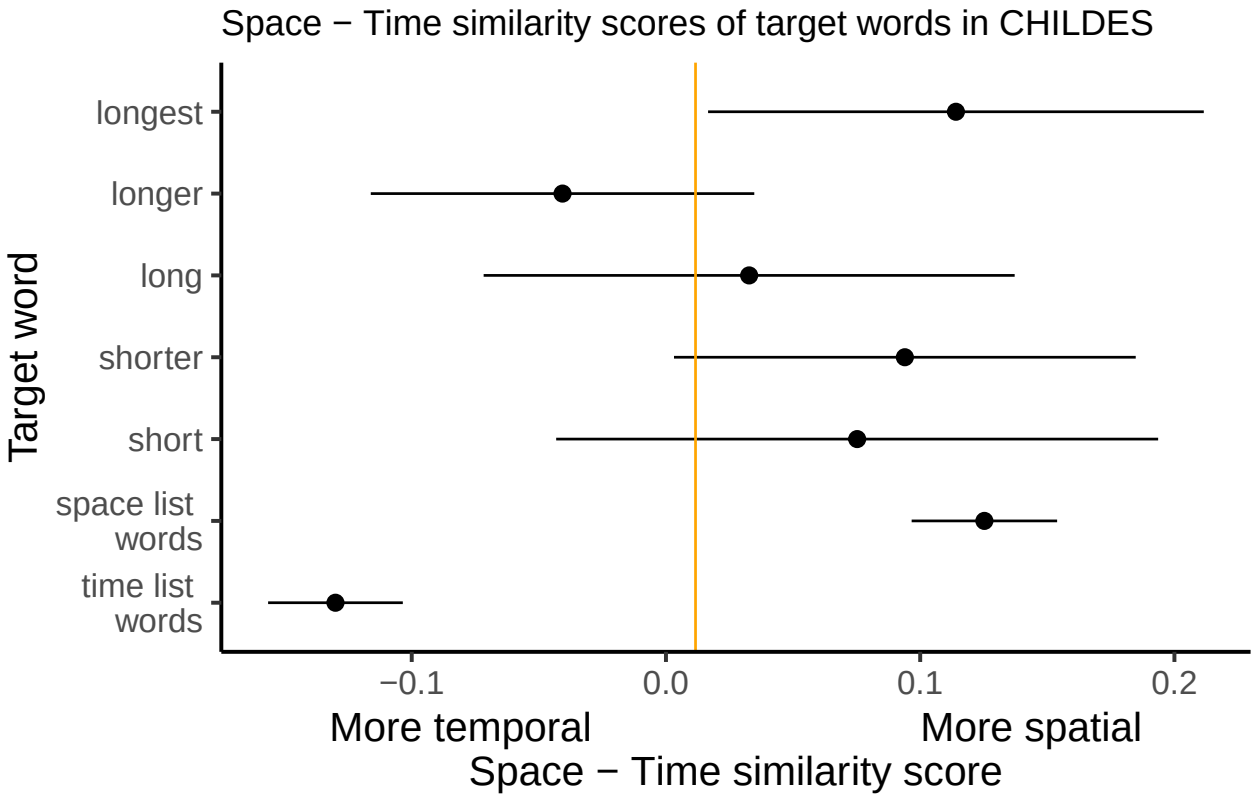
```
get_difference_CIs <- function(data) {
  data <- data %>%
    summarise(sd = sd(distance), mean = mean(distance), n = n()) %>%
    pivot_wider(names_from = domain,
                values_from = c(mean, sd, n)) %>%
    mutate(space_time_diff = mean_space - mean_time,
           se_diff = sqrt((sd_space^2)/n_space + (sd_time^2)/n_time),
           diff_ci_upper = space_time_diff + 1.96*se_diff,
           diff_ci_lower = space_time_diff - 1.96*se_diff) %>%
    ungroup()
  return(data)
}

get_difference_CIs_within_between <- function(data) {
  data <- data %>%
    summarise(sd = sd(distance), mean = mean(distance), n = n()) %>%
    pivot_wider(names_from = within_between,
                values_from = c(mean, sd, n)) %>%
    mutate(time_diff = mean_between_space_time - mean_within_time,
           space_diff = mean_within_space - mean_between_space_time,
           se_time_diff = sqrt((sd_between_space_time^2)/n_between_space_time +
                               (sd_within_time^2)/n_within_time),
           se_space_diff = sqrt((sd_within_space^2)/n_within_space +
                               (sd_between_space_time^2)/n_between_space_time),
           time_diff_ci_upper = time_diff + 1.96*se_time_diff,
           time_diff_ci_lower = time_diff - 1.96*se_time_diff,
           space_diff_ci_upper = space_diff + 1.96*se_space_diff,
           space_diff_ci_lower = space_diff - 1.96*se_space_diff) %>%
    ungroup()
  return(data)
```

```
}
```

```
childes_target_diffs <- childes_sims %>%
  group_by(target_word, domain) %>% get_difference_CIs() %>%
  select(target_word, space_time_diff, diff_ci_lower, diff_ci_upper)

childes_random_diffs <- childes_random_sims %>% group_by(domain) %>%
  get_difference_CIs() %>%
  select(space_time_diff, diff_ci_lower, diff_ci_upper) %>%
  mutate(target_word = "null")

childes_between_within_diffs <- st_childes %>% group_by(within_between) %>%
  get_difference_CIs_within_between() %>%
  select((contains("time_diff") | contains("space_diff")) &
         !contains("se_")) %>%
  pivot_longer(cols = starts_with("time") | starts_with("space"),
               names_to = c("word", "stat"),
               names_pattern = "([A-Za-z]+)_([A-Za-z_]+)",
               values_to = "value") %>%
  pivot_wider(names_from = stat, values_from = value) %>%
  rename("target_word" = "word", "space_time_diff" = "diff")

coca_target_diffs <- coca_sims %>%
  group_by(target_word, domain) %>% get_difference_CIs() %>%
  select(target_word, space_time_diff, diff_ci_lower, diff_ci_upper)

coca_random_diffs <- coca_random_sims %>% group_by(domain) %>%
  get_difference_CIs() %>%
  select(space_time_diff, diff_ci_lower, diff_ci_upper) %>%
  mutate(target_word = "null")

coca_between_within_diffs <- st_coca %>% group_by(within_between) %>%
  get_difference_CIs_within_between() %>%
  select((contains("time_diff") | contains("space_diff")) &
         !contains("se_")) %>%
  pivot_longer(cols = starts_with("time") | starts_with("space"),
               names_to = c("word", "stat"),
               names_pattern = "([A-Za-z]+)_([A-Za-z_]+)",
               values_to = "value") %>%
  pivot_wider(names_from = stat, values_from = value) %>%
  rename("target_word" = "word", "space_time_diff" = "diff")
```

```
childes_null <- rbind(childes_target_diffs,
              childes_random_diffs, childes_between_within_diffs) %>%
  filter(target_word == "null")

rbind(childes_target_diffs, childes_random_diffs, childes_between_within_diffs) %>%
  filter(target_word != "null") %>%
  mutate(target_word = factor(target_word,
                              levels = c("time", "space",
                                         "short", "shorter",
                                         "long", "longer", "longest"),
                              labels = c("time list \nwords", "space list \nwords",
                                         "short", "shorter",
```

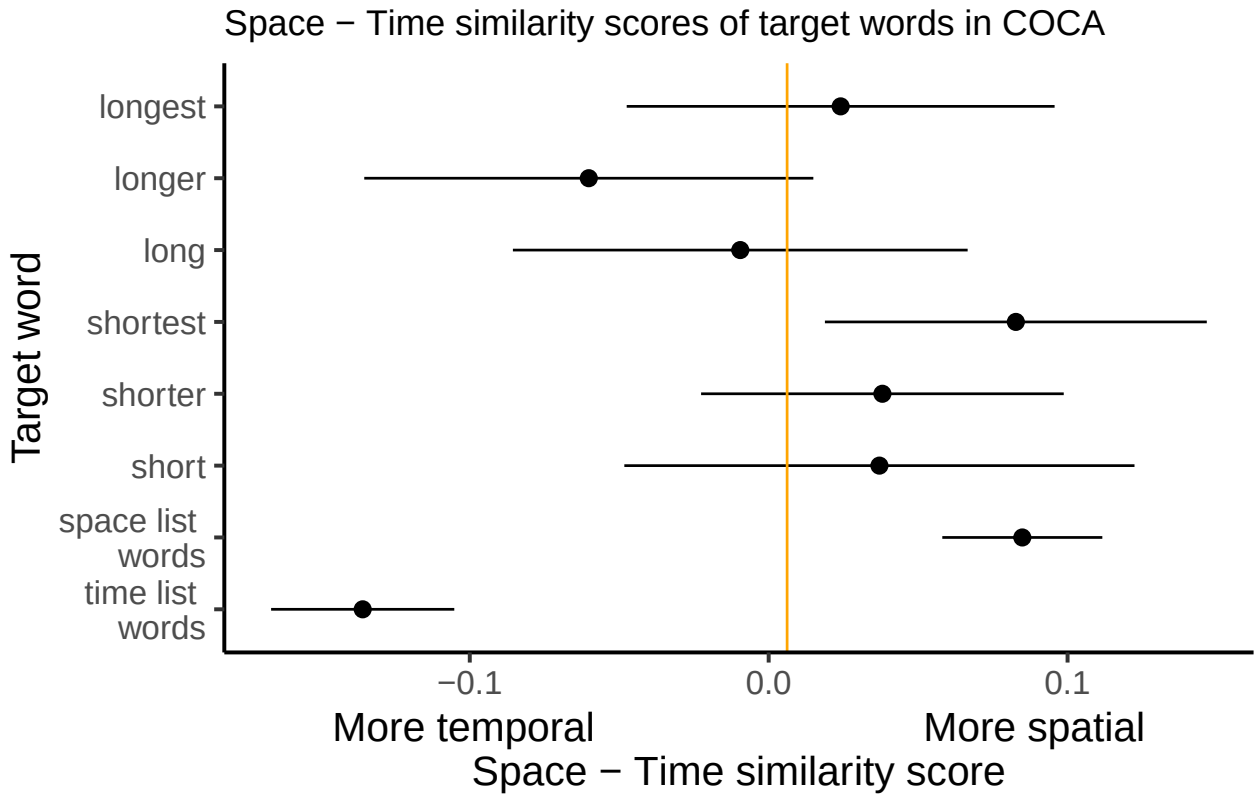
```
                                        "long", "longer", "longest"))) %>%
  ggplot() +
  geom_pointrange(aes(y = target_word, x = space_time_diff,
                      xmin = diff_ci_lower, xmax = diff_ci_upper),
                  position = position_dodge(width = 0.5)) +
  geom_vline(xintercept = childes_null$space_time_diff, color = "orange") +
  xlab("More temporal                                  More spatial\nSpace - Time similarity score") +
  ylab("Target word") +
  ggtitle("Space - Time similarity scores of target words in CHILDES")+
  theme(plot.title = element_text(size = 14))
```




CHILDES mean similarities between target words and space/time words.

```
coca_null <- rbind(coca_target_diffs, coca_random_diffs, coca_between_within_diffs) %>%
  filter(target_word == "null")

rbind(coca_target_diffs, coca_random_diffs, coca_between_within_diffs) %>%
  filter(target_word != "null") %>%
  mutate(target_word = factor(target_word,
                              levels = c("time", "space",
                                         "short", "shorter", "shortest",
                                         "long", "longer", "longest"),
                              labels = c("time list \nwords", "space list \nwords",
                                         "short", "shorter", "shortest",
                                         "long", "longer", "longest"))) %>%
  ggplot() +
  geom_pointrange(aes(y = target_word, x = space_time_diff,
                      xmin = diff_ci_lower, xmax = diff_ci_upper),
                  position = position_dodge(width = 0.5)) +
```

```
  geom_vline(xintercept = coca_null$space_time_diff, color = "orange") +
  xlab("More temporal                                More spatial\nSpace - Time similarity score") +
  ylab("Target word") +
  ggtitle("Space - Time similarity scores of target words in COCA") +
  theme(plot.title = element_text(size = 14))
```




COCA mean similarities between target words and space/time words.