# Expression of motion events in English: Manner and path verb usage in children's spontaneous narratives

Eric Xia, Revyn Kim, and Jared Giszack

CLPS 1360

Professor Uriel Cohen Priva

May 15, 2024

**Expression of motion events in English: Manner versus path verb usage in children's spontaneous narratives**

## 1. Introduction

Verbs are relatively difficult to acquire for children: while it is possible to simply point to a referent for many nouns, verbs are rather more multi-dimensional. For example, what comprises the action of eating? If you were to point at someone eating cereal, does the word *to eat* refer to the motion of spooning cereal into one's mouth, the action of chewing, or any of the other motions that the referent is doing? Verbs which encode movement are even more complex; the same motion event can be described as *running, leaping, entering,* or *going.* This project examines children's acquisition and usage of motion verbs through corpus linguistics methods.

## 2. Background

### 2.1 Manner and Path Verbs

An important category of verbs is that of motion verbs, which encode movement of an object. Such verbs of motion convey information about the displacement of a Figure relative to the Ground, where the Figure is some conceptually moveable point and the Ground is a stationary reference-point within the reference frame (Talmy 1975). Thus, in a sentence such as that in (1), the ball is the Figure, and the slope is the stationary Ground which provides the reference frame within which the ball moves.

(1) The ball rolled down the slope.

Languages are also able to encode features of movement such as path and manner. Path encodes where the Figure moves relative to the Ground – its trajectory – while manner encodes the way in which it moves. It is possible, then, to distinguish motion verbs into two groups: manner verbs (e.g. *run, dash, hop*) and path verbs (e.g. *go, enter, return*). In the sentence in (2a), the word *went,* a path verb, encodes that the man's movement happens in the direction of a place called *home* and that the end position of the man is seemingly at that home, but it tells nothing about how the man moved. Perhaps he hopped, skipped, or ran home. Compare this to the sentence in (2b), where the manner verb *walked* does encode the manner in which the man moved: he walked.

(2)

a. The man went home.
b. The man walked home.

### 2.2 Cross-Linguistic Differences in Encoding Spatial Information

These types of verbs encode spatial information, the mental representation of which is universally constrained by both human cognition and physical laws of the world. In addition, studies have shown that the ability to distinguish between manner and path as a spatial element appears fairly early in childhood development regardless of language (Smyder & Harrigan 2021). Thus, it would be reasonable to expect that languages encode such spatial information in similar ways; however, there is robust evidence for rather large differences between languages in how they encode movement in their verbs (Cappelle 2012; Slobin 1996). Languages can largely be split into two general categories: those which preferentially encode manner in their verbs, and those that preferentially encode path. The former of these two, called satellite-framed or S-framed languages, include languages such as English, Russian, and Chinese. These languages encode manner in their verbs, and path is encoded in a satellite element. In English, these satellite elements generally take the form of prepositional phrases: in (3), the manner by which the girl moves is conveyed through the verb *skipped*, while the path that the girl took is encoded within the satellite element *out of the bookstore*.

(3) The girl skipped out of the bookstore.

Languages that preferentially encode the path of a movement, called V-framed or verb-framed languages, feature opposite patterns to those of S-framed languages: the verb encodes path, while optional features such as manner are conveyed through satellite elements. Compare (3) to (4), where the path is encoded in *passa*, while the satellite element *à grande vitesse* shows the manner in which the UFO moved.

(4) French (Cappelle, 2012, p. 5)
un OVNI passa à grande vitesse.
a UFO pass-PAST at great speed
'a UFO passed at great speed.'

**2.3 Manner and Path Verb Usage in Children**

Research has shown that young children tend to favor expressing path rather than manner, regardless of the language that they and those around them speak (Maguire et al., 2010). This supports the idea present in typological literature that directional features such as path are a "core" lexical element, while features such as manner are optional (Talmy 1991). Nevertheless, as children develop linguistically, children begin to conform to the patterns of the language they speak; studies have shown that this bias begins to appear around the ages of three to seven (Maguire et al. 2010; Skordos & Papafragou 2014). English-speaking children, then, should show a bias for expressing manner over path, while children who speak a language such as Spanish should show a preference for using verbs to express path over manner, and this is indeed what studies have found (Smyder & Harrigan 2021; Maguire et al. 2010)

Smyder and Harrigan (2021), for example, presented children with a series of familiarization videos depicting an actor performing an action named by a novel verb, then asked

the children to match this novel verb with an action. In line with previous research, they found that English-learning children are more accurate at mapping the manner of an event to a novel verb than the path. Research in this vein has largely relied on such forced-choice tasks which involve mapping of actions onto novel verbs. This project extends these studies to see if this trend extends to natural speech: will English-speaking children show a developing preference for manner over path verbs? We hypothesize that the satellite-framed nature of English will indeed influence children's relative usage of manner versus path verbs, and older children will show a stronger preference for expressing manner of a motion in their narratives compared to younger children.

## 3. Data

### 3.1 ECSC Corpus

For our data, we used the FROGS subset of the Eugene Children's Story Corpus. This corpus is a part of the larger CHILDES corpora, and it consists of 180 structured spontaneous narratives. In the FROGS subset are 367 audio recordings and transcriptions based on the following picture-books: *A Boy, a Dog, and a Frog*; *Frog, Where are You?*; *Frog on His Own*; and *One Frog too Many*.

To collect this data, researchers adopted a longitudinal approach wherein they investigated broader acquisition of temporal patterns in language. Beginning in 2009, children aged 5 through 7 began the study, repeating it annually until a total of three measurements per child were reached. There were some children who did not participate in all three iterations of the study, but most did. Since our project is investigating an aspect of language acquisition, a longitudinal dataset such as this one is appropriate as it allows us to examine behaviors over time rather than just at a given instant.

To collect the data, children looked through one of the picture-books and told a narrative to their accompanying caregiver. After this, the caregiver looked through the same picture-book and told their own version of the story. Lastly, the child told a second story about the same picture-book, and this telling is what researchers ultimately used to generate the final transcriptions. By repeating the storytelling process, researchers hoped to minimize language planning and word-finding effects on the production of narrative prosody.

### 3.2 Data Annotation and Cleaning

In looking at patterns of children's uses of manner versus path verbs, we first needed to define what these classes mean. We coded manner verbs as words which express *how* the motion happened: examples of manner verbs from this corpus include *run*, *jump*, and *swim*. For our project, we defined path verbs as those which encode information about the path of a motion event without providing information about the manner. Examples from the corpus include verbs like *come* and *went*.

The actual method of determining the manner or path status of a verb is not straightforward, and requires context. There are many cases where verbs are nominalized within the corpus: for example, one narrative mentions “their morning walk” (107F_1001_YR2 #34), using “walk” as a noun and not a verb. Such cases make labeling words based on their lexical representation impractical. Thus, we opted to annotate the transcriptions by hand to generate our eventual data frame for our analyses. Initially, we each annotated 10 stories, marking all instances of manner and path verbs. From there, we used Cohen’s Kappa to assess our inter-annotator agreement. When using this statistic, a value of 0 suggests total disagreement, while a value of 1 suggests perfect agreement. After defining some assumptions about our annotations for consistency’s sake, we found that our Cohen’s Kappa value was about 0.925, suggesting high agreement between annotators.

After establishing that we had high agreement in our annotation judgments, we randomly selected transcripts to be representative of the ages and stories composing the total study population, balancing for gender. Because our initial agreement was high, we elected not to overlap our annotations, so we each annotated an equal portion of each age of our selected transcripts. From there, we had our working data to move on to our analyses.

## 4. Analysis

While we began with 172 selected transcripts, our final annotated dataset was 165 story transcripts: some from the 6 and 8 age groups were not annotated and so they were excluded from the dataset.

Some preprocessing was done in Python, primarily the aggregation of individual story narratives. This dataset, which represented each individual line and annotation, had six columns. The first was the original line of text of the narrative, followed by the annotation, which in the majority of cases was NA. Then the ID of the child, the study year, the age, and the gender followed. From this point onwards, the analysis was conducted in R.

The identification of manner and verb counts per story, which by our premises comprised the motion verb count per story, was a straightforward grepl() on the grouped dataframe: this counted individual instances of 0 or 1 (including multiple on the same line), which corresponded to path and manner respectively. This was converted to a manner frequency by dividing the manner count by the total number of manner and path verbs. With the manner frequency variable, we were able to test our original hypotheses, that the age of the speaker or the study year of the speaker correlated to the relative frequency of manner utterances in the narrative (which inherently controls for the speaker ID). This was done with a mixed effects model which accounted for the story as a random intercept, as well as an additional model which tested the hypothesis that study year correlated with manner frequency. With the hypothesis that age correlated to manner frequency up to the story choice, the model produced an extremely low fixed effect for the predictor (<0.001/month) with a low t-value (-0.575). With the hypothesis

that the study year correlated to manner frequency up to the story choice, the model also produced an extremely low fixed effect (<0.001/year) with a low t-value (0.001).

Subsequent work took on the task of finding predictors which could correlate to manner frequency. Metadata from the study was brought in and merged with the speaker IDs, to allow for comparison of Peabody Picture Vocabulary Test (PPVT) scores for each child. PPVT is a verbal aptitude test, consisting of stimulus words and image plates: participants listen to a word uttered by the interviewer and selects one of four pictures that best describes the word's meaning. The mean PPVT-R value for the selected children was 144.04, while the range was 77-199, and roughly follows a normal distribution. Mixed effect models with these predictors also failed to produce meaningful correlations for manner frequency.

In the end, no predictors were found to meaningfully correlate to the manner frequency, even with controls for story, story length, study year, and the speaker ID. When we moved to predicting other variables, strong correlations were revealed between the predicting variables and the manner, path, and verb count, and we focused on investigating these instead.

**5. Results**

Our results do not align with our original hypothesis that the proportion of manner verbs to path verbs used would increase with speakers' ages: the fixed effect for age as a predictor was extremely low, as was the fixed effect for study year. Indeed, we could not find any predictors which were significantly correlated with *relative* manner-path verb usage. What we were able to find is a statistically significant correlation between PPVT-R and manner verb count which is *not* present between PPVT-R and path verbs, which will be elaborated on below.

Analysis of overall verb counts yielded significant correlations, particularly between the age and the overall manner, path, and motion verb count. Our mixed effects model found on average an increase of 0.05 manner verbs per month of age, with story as random intercept. It also found on average an increase of 0.03 path verbs per month of age ($t = 2.4$) with story as random intercept. This corresponds with the mixed effects model predicting an increase of 0.08 motion verbs per month of age ($t = 4$).

We also found a strong correlation between the study year and the number of motion verbs. The study year is a variable associated with each speaker, reporting which longitudinal year they took the study (up to three). Out of our random sample, there were 93 participating in the study for the first time, 32 coming back for the second year, and 40 for the third. Our mixed effects model found on average an increase of 1.37 motion verbs per year of study, from year one to year three ($t = 3.1$), with the story effect modeled as random intercept.

This effect persists when the speaker ID and age are added as a random intercept, suggesting that the study year is an effective predictor independent from the identity of the speaker and the age of the speaker. The reason for this correlation is somewhat unclear, as it

cannot come from an increased familiarity with the story (each child was given a unique story each year they returned). It is most likely an effect specific to the testing methodology. One hypothesis is that such an effect comes from an increased familiarity with the format of the study, and a greater willingness to elaborate on the original narrative.

We were also able to find a correlation between the Peabody Picture Test Score - Revised (PPVT-R) for each speaker and the number of motion verbs. Our mixed effects model found on average an increase of 0.043 motion verbs per point PPVT-R when the story effect is modeled as random intercept (t= 2.8).

This effect persists when age was added as a random intercept, suggesting that PPVT-R scores are an independent predictor of the motion count in the story. When using age as the predictor variable and controlling for speaker ID, age, and the selected story, one month of age corresponds to 0.062 additional motion verbs in the story (t = 3.1), with 0.028 additional path verbs (t=2.3) and 0.047 additional manner verbs (t=2.9) respectively.

A striking difference between motion and path verb counts was found when using the PPVT-R as a predictor. W*hen using PPVT-R scores as the predictor variable and controlling for the speaker ID, the story length, and the selected story, one point increase in the PPVT-R corresponds to 0.032 additional manner verbs per story (t=2.6), but only 0.0076 additional path verbs (t=0.83).* This suggests that manner verbs represent a unique lexical category that is independently predicted by PPVT-R.

**6. Conclusion and Next Steps**

Our results did not confirm our initial hypothesis that the relative manner frequency corresponds to the age of the child up to the story chosen. There are several reasons for why we might have found these results. First, it is possible that inter-speaker variability was simply too high, meaning that there is no distinct trend that could be observed; in this case, it would be prudent to use increased sample size in a larger project. Further, since there is evidence that language-based biases begin to develop around the age of three, perhaps our youngest age group (age five) may have simply already reached adult-like levels of relative manner-path verb usage.

Extending our age range to include both older and younger speakers may reveal trends that align better with previous research: in the case that our youngest age group is already adult-like, then younger speakers might reveal a larger trend. Comparison to adult speakers would provide a means of confirmation; alternatively, if adult speakers demonstrate a different proportion of manner versus path verbs, then it would reveal that perhaps the change in manner-path verb usage occurs in older children, not younger. Our analysis also may have been erroneous. One oversight made during the process was neglecting to normalize the data: we did not transform variables to the same axis or perform any log transforms on underlying variables.

However, we were able to find a correlation between the age and the manner count, path count, and motion verb count, consistent across a range of controls. We were also able to find correlations for verb counts when using PPVT-R and study year as predictors, with controls. This suggests that our method of analysis is viable, but that the initial metric of manner frequency may not have been well-defined as a proxy for the quality we wanted to measure.

Additionally, we find that these correlations (manner, path, and motion count to age, study year) persist to some degree when the length of the story, and the speaker ID are added as random intercepts. These were done at the request of a commenter, and by Dr. Cohen Priva. This provides supportive evidence that motion verb usage is an independent category of lexical expression, and not correlated to either overall verb usage nor the overall amount of words used in a spontaneous narrative.

Finally, when revisiting the dataset, we discovered an extremely promising result. We find that manner verb usage is strongly correlated to PPVT-R scores, but not path verb usage. *This suggests that manner verbs represent a unique lexical category that is independently predicted by PPVT-R.* Further work with this dataset (starting with normalized variables) could confirm this result and lead to further insights.

**7. Responses to Questions**

Many of the questions we received involved possible directions that further research could take – both ones which we considered as well as some that we hadn't thought of. One of the most popular of these suggestions included extending the study to adults and to multilingual children (see Appendix for a full list of questions). As discussed in the previous section, extending the age range to include adults would enable us to confirm whether the children in our dataset were indeed adult-like or not and thus provide an explanation for our findings.

Expanding future research to examine multilingual speakers would enable cross-linguistic comparison between S-framed and V-framed languages. In addition, it would perhaps provide some insight for how multilingual speakers negotiate between the languages they speak: if children simultaneously learn an S-framed and a V-framed language, would the speaker show a bias towards encoding manner or towards encoding path in the verb? Further suggestions for expansion of this project include examining whether use of a pronoun versus a specific subject would result in significantly different patterns of manner-path verb usage, as well as more precise categorization of motion verbs into motion, path, and complex or ambiguous classes. This latter suggestion does indeed have precedents in the literature, although it was beyond the scope of this project.

The following clarifying questions are unrelated in theme, and they are listed in a simple question and answer format:

1. I'm wondering why the participants weren't used as a random effect?

- As discussed above, we did use participants as a random effect, by adding speaker ID as a random intercept in our models.

2. I was wondering what exactly is meant by "richer language distinctions" that could be made in future study and how exactly your current work can be extended to non-motion verbs.
    - The comment made about richer language distinctions specifically referred to further distinctions that could be made in encoding manner. In one descriptive passage, some participants describe frogs *riding* on a turtle, as opposed to *sitting*. Here, *riding* is a more articulate description and more “manner-like”. This distinction is actually somewhat captured by our current framework, as we agreed to classify sitting as a non-motion verb (so not annotated), while riding is a manner verb. Extending the current work would include encoding verbs like these or using a manner/path scale, or making other fine-grained distinctions.

3. I'm curious why you now think that manner frequency may not be the best metric. I'm also curious whether you could easily look into the length of the stories as a function of these variables.
    - When running models to predict manner frequency, none of the variables were successful no matter what control variables we assigned, except for the manner count for obvious reasons. Our initial conclusion is that this variable is not actually tied to any concept within linguistics, and that this particular line of prediction was not going in the right direction. We were able to find correlations to the manner and path counts, so we explored those further.
    - Regarding your second point, that is a great suggestion that we followed above.

4. It might be good to explain what the original longitudinal study did to prevent children from changing in their response patterns over time simply due to increased familiarity. Were the participants given different stories to describe each time?
    - Yes! As outlined in the methodology, ECSC participants were given different stories to describe each time.

5. In your analysis, do you need to account for the fact that there seem to be more manner verbs available in English than path verbs, it seems? Would this influence how speakers choose / are conditioned to choose verbs? Or does this not affect your analysis.
    - I think, more than count of verbs (as there are actually quite a lot of “basic” or elementary path verbs that kids would be familiar with), research has shown that just the language’s way of encoding motion in a sentence is more influential. Our

initial analysis accounted for this by using proportion of manner vs path, rather than count – we only looked at the actual word counts during further analysis, since our initial analysis gave us a null result.

6. I'm wondering if you had any way to take into account the frequency of path vs manner verbs usage in English in general? You mentioned in the beginning that English is a satellite-framed language, so it might not have been the best language to predict manner acquisition
    - We aren't predicting only manner acquisition, but instead the differences between manner and path verb acquisition/usage; for our initial analysis, we looked at the acquisition of both and how they change with age. Since we did end up looking at actual count as well, it would be interesting to see how the actual word counts are different in children that speak a different language.

# References

Cappelle, B. (2012). English is less rich in manner-of-motion verbs when translated from French. *Across Languages and Cultures*. *13*, 173-195. 10.1556/Acr.13.2012.2.3.

Dunn, Lloyd M. Peabody Picture Vocabulary Test-Revised (PPVT-R). Forms L and M. Circle Pines, Minn. :American Guidance Service, 1981.

Kallay, J. E., & Redford, M. A. (2021). Clause-initial AND usage in a cross-sectional and longitudinal corpus of school-age children's narratives. *Journal of Child Language*, *48*(1), 88–109. https://doi.org/10.1017/s0305000920000197

Maguire, M., Hirsh-Pasek, K., Golinkoff, R.M., Imai, M., Haryu, E., Vanegas, S., Okada, H., Pulverman, R., & Sanchez-Davis, B. (2010). A developmental shift from similar to language-specific strategies in verb acquisition: A comparison of English, Spanish, and Japanese. *Cognition,* 114, pp. 299-319. http://dx.doi.org/10.1016/j.cognition.2009.10.002

Skordos, D. & Papafragou, A. (2014). Lexical, Syntactic, and Semantic-Geometric Factors in the Acquisition of Motion Predicates. *Developmental Psychology. 50( 7)*, 1985–1998. http://dx.doi.org/10.1037/a0036970

Slobin, D. (1996). Two ways to travel: verbs of motion in English and Spanish. In M. Shibatani & S. A. Thompson (Eds.), *Grammatical Constructions: their form and meaning.* (pp. 195-219). Clarendon Press.

Smyder, R. & Harrigan, K. (2021). The influence of language-specific and universal factors on acquisition of notion verbs. *Proceedings of the Linguistic Society of America*. *6(1).* 927–937. https://doi.org/10.3765/plsa.v6i1.5036.

Speed, T. (2015). Manner/path typology of Bulgarian motion verbs. Journal of Slavic Linguistics, 23(1), 51–81. https://doi.org/10.1353/jsl.2015.0000

Talmy, L. (1975). Figure and Ground in Complex Sentences. *Proceedings of the First Annual Meeting of the Berkeley Linguistics Society.* 419-430. https://doi.org/10.3765/bls.v1i0.2322

Talmy, L. (1991). Path to Realization: A Typology of Event Conflation. *Proceedings of the Seventeenth Annual Meeting of the Berkeley Linguistics Society: General Session and Parasession on The Grammar of Event Structure*, pp. 480-519. https://doi.org/10.3765/bls.v17i0.1620

# FrogWords

## Frog Corpus

Here are scripts for working with the annotated ECSC Frog Corpus data.

### Old Read in Data (ignore)

```
# library(tidyverse)
# ec5 <- read_csv("Grouped/5.csv", TRUE)
# ec6 <- read_csv("Grouped/6.csv", TRUE)
# ec7 <- read_csv("Grouped/7.csv", TRUE)
# ec8 <- read_csv("Grouped/8.csv", TRUE)
# ec9 <- read_csv("Grouped/9.csv", TRUE)
# ec10 <- read_csv("Grouped/10.csv", TRUE)
#
# all_ec <- rbind(ec5, ec6, ec7, ec8, ec9, ec10)
#We want to determine if the study year has an effect on the respondent.
 ↪
###To do this:
# 1. we will create a manner_freq variable for each row, grouped by
 ↪  respondent and study_year

# all_ec <- all_ec %>% mutate(id = paste(id, yr, sep = "_")) %>%
 ↪  select(-yr)

#
# # 2. Then, we will group the data by study year and analyze with other
 ↪  variables like age accounted for.
# #grouped only by id.
```

```
# id_grouped_ec <- all_ec %>% group_by(id) %>%
 ↪  mutate(manner_ct=sum(grepl("1", annotation)), path_ct=sum(grepl("0",
 ↪  annotation)))
# id_grouped_ec <- id_grouped_ec %>% mutate(id.mn_freq=manner_ct /
 ↪  (manner_ct + path_ct))
# #
# all_ec <- all_ec %>% distinct(id, .keep_all=TRUE)
# all_ec
```

## Read in Data

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
newec <- read_csv("new_actual_merged_df.csv", TRUE)
```

```
Rows: 11676 Columns: 3
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (3): text, annotation, id

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
newec$yr <- as.numeric(substring(newec$id, nchar(newec$id)))
```

```
newec <- separate(newec, id, into = c("just_id", "year"), remove=FALSE,
 ↪  sep = "(?<=[0-9])_")
#newec <- newec %>% mutate(id = paste(id, yr, sep = "_")) %>%
 ↪  select(-yr)
# newec <- newec %>% distinct(id, .keep_all=TRUE)
distinct(newec, id)
```

```
# A tibble: 165 x 1
   id
   <chr>
 1 67M_4034_YR1
 2 64M_1055_YR1
 3 64M_1032_YR1
 4 61M_3007_YR1
 5 65M_4005_YR1
 6 71F_2030_YR1
 7 70F_2010_YR1
 8 69F_2015_YR1
 9 65F_4012_YR1
10 66F_2008_YR1
# i 155 more rows
```

### Add manner_freq and motion_ct variables

Here I also count adjacent MM/PM/MP/PP occurrrences. This part is mostly for fun and to validate the data.

```
newec <- newec %>% group_by(id) %>% mutate(
  manner_ct=sum(grepl("1", annotation)),
  path_ct=sum(grepl("0", annotation)),
  ct0_0 = sum(grepl("0.*0", annotation)),
  ct0_1 = sum(grepl("0.*1", annotation)),
  ct1_0 = sum(grepl("1.*0", annotation)),
  ct1_1 = sum(grepl("1.*1", annotation)),
  total_len=n())

newec <- newec %>% mutate(manner_freq = manner_ct / (manner_ct +
 ↪  path_ct), motion_ct = manner_ct + path_ct)

ct = grepl("1.*1", newec$annotation)
```

```
newec[ct, ]
```

```
# A tibble: 21 x 15
# Groups:   id [15]
   text       annotation id    just_id year     yr manner_ct path_ct ct0_0 ct0_1
   <chr>      <chr>      <chr> <chr>   <chr> <dbl>     <int>   <int> <int> <int>
 1 jumped ou~ 1jumped, ~ 68F_~ 68F_20~ YR1       1        10       0     0     0
 2 he walked~ 1WALKED, ~ 71M_~ 71M_10~ YR2       2         6       2     1     0
 3 and ran a~ 1RAN, 1RAN 68M_~ 68M_20~ YR2       2         7      11     0     1
 4 suddenly ~ 1JUMPED, ~ 70M_~ 70M_10~ YR2       2        14       0     0     0
 5 leap poun~ 1leap, 1p~ 76F_~ 76F_10~ YR2       2         4       7     1     0
 6 and when ~ 1FLEW, 1C~ 91M_~ 91M_10~ YR1       1         2      10     0     0
 7 and then ~ 0GO, 1SAI~ 94F_~ 94F_10~ YR1       1         8       3     1     1
 8 and hops ~ 1HOPS, 1H~ 94F_~ 94F_10~ YR1       1         8       3     1     1
 9 he hops a~ 1HOPS, 1H~ 87M_~ 87M_10~ YR2       2        15       3     0     0
10 when he s~ 1WALKING,~ 75F_~ 75F_10~ YR3       3        10       5     0     0
# i 11 more rows
# i 5 more variables: ct1_0 <int>, ct1_1 <int>, total_len <int>,
#   manner_freq <dbl>, motion_ct <int>
```

```
#30 * 5 + 22 = 172 speakers to 165
distinct_newec <- newec %>% group_by(id) %>% mutate(story_len = n())
distinct_newec <- distinct(distinct_newec, id, .keep_all=TRUE)
distinct_newec
```

```
# A tibble: 165 x 16
# Groups:   id [165]
   text       annotation id    just_id year     yr manner_ct path_ct ct0_0 ct0_1
   <chr>      <chr>      <chr> <chr>   <chr> <dbl>     <int>   <int> <int> <int>
 1 one time ~ <NA>       67M_~ 67M_40~ YR1       1         3       3     0     0
 2 one time ~ <NA>       64M_~ 64M_10~ YR1       1         7       2     0     0
 3 the frog ~ 0WENT      64M_~ 64M_10~ YR1       1         1       6     0     0
 4 he find a~ <NA>       61M_~ 61M_30~ YR1       1         4       2     0     0
 5 one .      <NA>       65M_~ 65M_40~ YR1       1         4       3     0     0
 6 there was~ <NA>       71F_~ 71F_20~ YR1       1         4       1     0     0
 7 there was~ <NA>       70F_~ 70F_20~ YR1       1         6       5     0     0
 8 the frog . <NA>       69F_~ 69F_20~ YR1       1         4       0     0     0
 9 one frog ~ <NA>       65F_~ 65F_40~ YR1       1         4       1     0     0
10 kay .      <NA>       66F_~ 66F_20~ YR1       1         1       5     0     0
# i 155 more rows
```

```
# i 6 more variables: ct1_0 <int>, ct1_1 <int>, total_len <int>,
#   manner_freq <dbl>, motion_ct <int>, story_len <int>
```

### Add per-speaker predictors: RS, PPVT R, PPVT SS

PPVT is a verbal aptitude test, consisting of stimulus words and image plates: participants listen to a word uttered by the interviewer and selects one of four pictures that best describes the word's meaning. The mean PPVT-R value for the selected children was 144.04, while the range was 77-199, and roughly follows a normal distribution. Mixed effect models with these predictors also failed to produce meaningful correlations for manner frequency.

```
# #ensure that id group is valid
id_stories <- read_csv('names_merged.csv')
```

```
Rows: 172 Columns: 2
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (2): names, story

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
distinct_newec_stories <- merge(distinct_newec, id_stories, by.x = 'id',
 ↪  by.y = 'names',  all.x = TRUE)
distinct_newec_stories$age <- as.numeric(sub("^([0-9]+).*", "\\1",
 ↪  distinct_newec_stories$id))
#distinct_newec_stories
speaker_data <- read_csv('speaker_data.csv')
```

```
Warning: One or more parsing issues, call `problems()` on your data frame for details,
e.g.:
  dat <- vroom(...)
  problems(dat)

Rows: 209 Columns: 20
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (20): Speaker ID, SUBJ, R/E C, R/E P, M/F C, M/F P, AGE/mo, Test Date, G...
```

```
i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
merged_ec <- merge(distinct_newec_stories, speaker_data, by.x="just_id",
 ↪  by.y="Speaker ID", all.x = TRUE)
#extract
# merged_all_ec <- merged_all_ec %>% mutate(study_yr =
 ↪  as.numeric(str_extract(study_yr, "\\d+$")))
merged_ec$RS = as.numeric(merged_ec$`RS`)
```

```
Warning: NAs introduced by coercion
```

```r
merged_ec$ppvt_r = as.numeric(merged_ec$`PPVT R`)
```

```
Warning: NAs introduced by coercion
```

```r
merged_ec$ppvt_ss = as.numeric(merged_ec$`PPVT SS`)
```

```
Warning: NAs introduced by coercion
```

```r
#merged_ec
```

```r
#this should be per story...
#merged_ec
```

```r
t <- table(merged_ec$just_id)
length(t)
```

```
[1] 125
```

**Attempting to restrict the dataset to repeat study years to visualize separately**

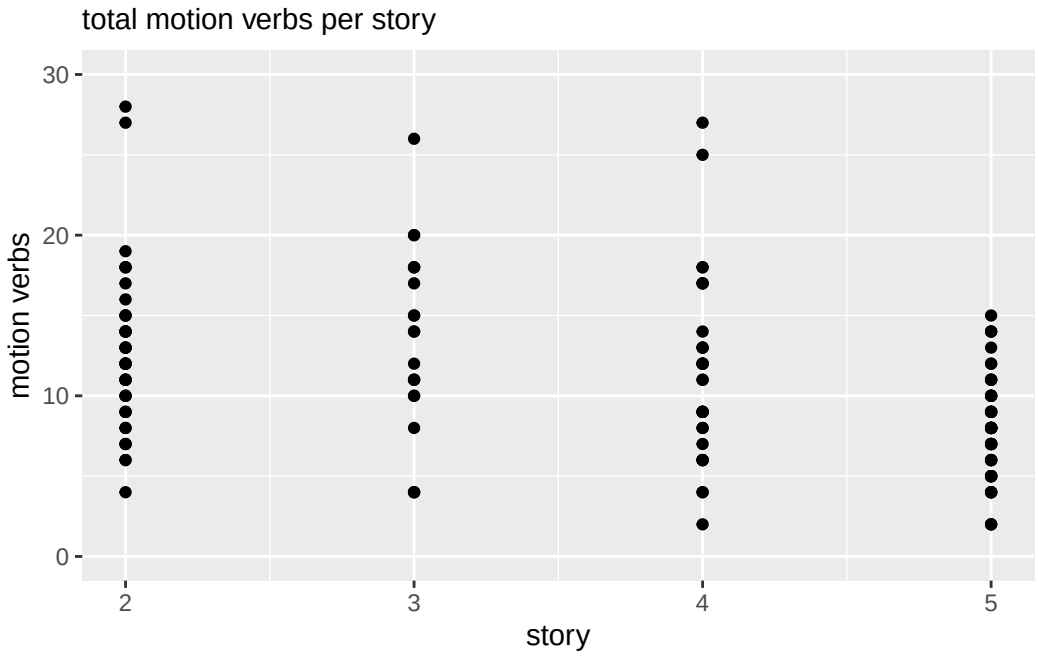
```
#Only a few different speakers were repeated in the data, so this is not
 ↪  a viable additional analysis.
# df <- all_ec
# # Group by 'id' and count the distinct 'study year' values for each
 ↪  'id'
# id_study_year_count <- df %>%
#   group_by(id) %>%
#   summarise(num_study_years = n_distinct(study_yr))
#
# # Filter 'id's where the number of distinct 'study year' values is
 ↪  greater than 1
# ids_with_multiple_study_years <- id_study_year_count %>%
#   filter(num_study_years > 1)

#ids_with_multiple_study_years$id
```

```
#Collapse groups into one
#colnames(all_ec)
#all_ec
```

**Plotting motion verbs per story (to demonstrate importance of controlling for story)**

```
merged_ec$story <- as.numeric(as.factor(merged_ec$story))
ggplot(data = merged_ec, aes(x = story, y = motion_ct)) +
  geom_point() +
  labs(title = "total motion verbs per story",
       x = "story",
       y = "motion verbs") + ylim(0, 30) + xlim(2, 5) + theme(plot.title
        ↪  = element_text(size = 11))
```

```
Warning: Removed 2 rows containing missing values or values outside the scale range
(`geom_point()`).
```




## Mixed-effects modeling

There are no positive effects associated with either age or study year when predicting manner frequency. However, the manner, path, and motion count are predictable from the data. Verb utterances did not get annotated, only the occurrence of motion verbs.

```
library(lme4)
```

```
Loading required package: Matrix
```

```
Attaching package: 'Matrix'
```

```
The following objects are masked from 'package:tidyr':

    expand, pack, unpack
```

```
summary(m.lmer <- lmer(formula = motion_ct ~ age + (1 | story),
 ↪  data=merged_ec))
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: motion_ct ~ age + (1 | story)
   Data: merged_ec

REML criterion at convergence: 986.4

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.9703 -0.6179 -0.0862  0.4320  4.3611

Random effects:
 Groups   Name        Variance Std.Dev.
 story    (Intercept)  8.263   2.875
 Residual             21.399   4.626
Number of obs: 165, groups:  story, 5

Fixed effects:
            Estimate Std. Error t value
(Intercept)  4.51249    2.28237   1.977
age          0.08216    0.02079   3.951

Correlation of Fixed Effects:
    (Intr)
age -0.777
```

**Plotting Manner Frequency against study year**

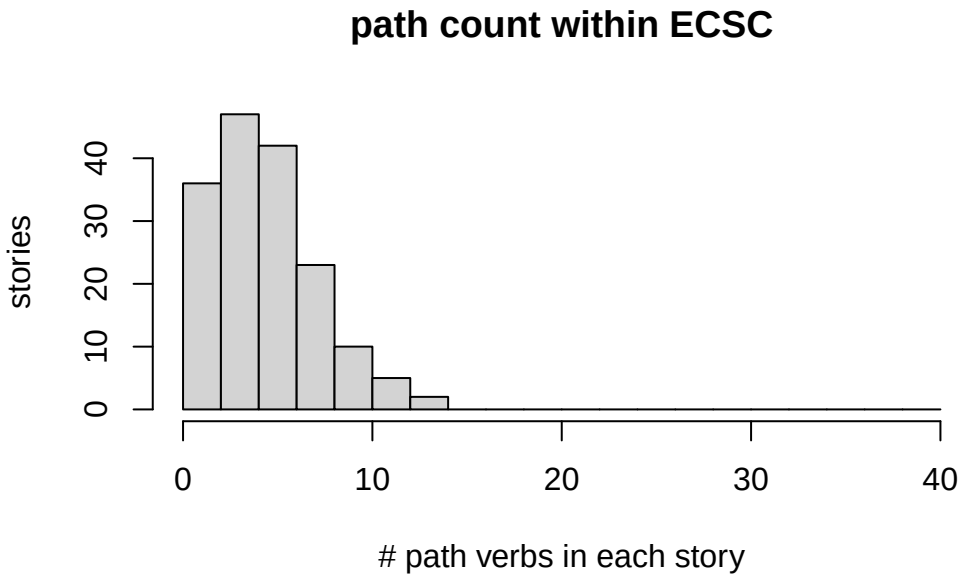
```
# ggplot(merged_ec, aes(x = yr, y = manner_freq)) +
#   stat_summary(fun.data = mean_cl_normal, geom = "bar") +
#   labs(x = "Study Year", y = "Manner Frequency")  + ylim(0.1, 3)
```

```
library(lme4)
#predicting overall manner_frequency with study_year is not a good
 ↪  idea...
#no correlation between total verb usage and manner_path freq
# m.lmer <- lmer(manner_freq ~ total_ct + (1 | age), data=all_ec)
# summary(m.lmer)

#all_ec$id <- as.factor(all_ec$id)
#all_ec$id
```

```
# m.lm <- lm(manner_freq ~ , data=all_ec)
# summary(m.lm)
```

**Plotting path count within the ECSC**

```
hist(merged_ec$path_ct, main = "path count within ECSC", xlab = "# path
 ↪  verbs in each story", ylab = "stories", breaks=seq(0, 40, by = 2))
```




```
# Linear mixed model fit by REML ['lmerMod']
#
# Formula: manner_freq ~ study_yr + (1 | all_ec$id)
#    Data: all_ec
#
# REML criterion at convergence: -15319.7
#
# Scaled residuals:
#     Min      1Q  Median      3Q     Max
# -3.0610 -0.0018 -0.0004  0.0017  9.4611
#
# Random effects:
#  Groups    Name        Variance Std.Dev.
```

```
#  all_ec$id (Intercept) 0.04860  0.22046
#  Residual              0.00125  0.03536
# Number of obs: 4095, groups:  all_ec$id, 52
#
# Fixed effects:
#              Estimate Std. Error t value
# (Intercept)  0.615821   0.030609   20.12
# study_yrYR2 -0.189788   0.004488  -42.29
# study_yrYR3 -0.086032   0.003502  -24.57
#
# Correlation of Fixed Effects:
#             (Intr) st_YR2
# study_yrYR2 -0.040
# study_yrYR3 -0.037  0.522

# Linear mixed model fit by REML ['lmerMod']
# Formula: manner_freq ~ study_yr + (1 | id) + (1 | age)
#    Data: all_ec
#
# REML criterion at convergence: -16187.7
#
# Scaled residuals:
#     Min      1Q  Median      3Q     Max
# -3.8926 -0.0018 -0.0001  0.0011  9.9006
#
# Random effects:
#  Groups   Name        Variance Std.Dev.
#  id       (Intercept) 0.052140 0.22834
#  age      (Intercept) 0.006789 0.08240
#  Residual             0.001004 0.03168
# Number of obs: 4095, groups:  id, 52; age, 6
#
# Fixed effects:
#             Estimate Std. Error t value
# (Intercept)  0.59449    0.04918  12.089
# study_yrYR2 -0.17724    0.04914  -3.607
# study_yrYR3 -0.01739    0.04365  -0.399
#
# Correlation of Fixed Effects:
#             (Intr) st_YR2
# study_yrYR2 -0.273
```

```
# study_yrYR3 -0.304  0.440
```

```
#plot of frequency against year of study
#reduce
# merged_all_ec <- merged_all_ec %>% group_by(id, study_yr) %>%
 ↪  summarize(id, study_yr, age, manner_freq)
# merged_all_ec <- merged_all_ec %>% distinct
# merged_all_ec
```

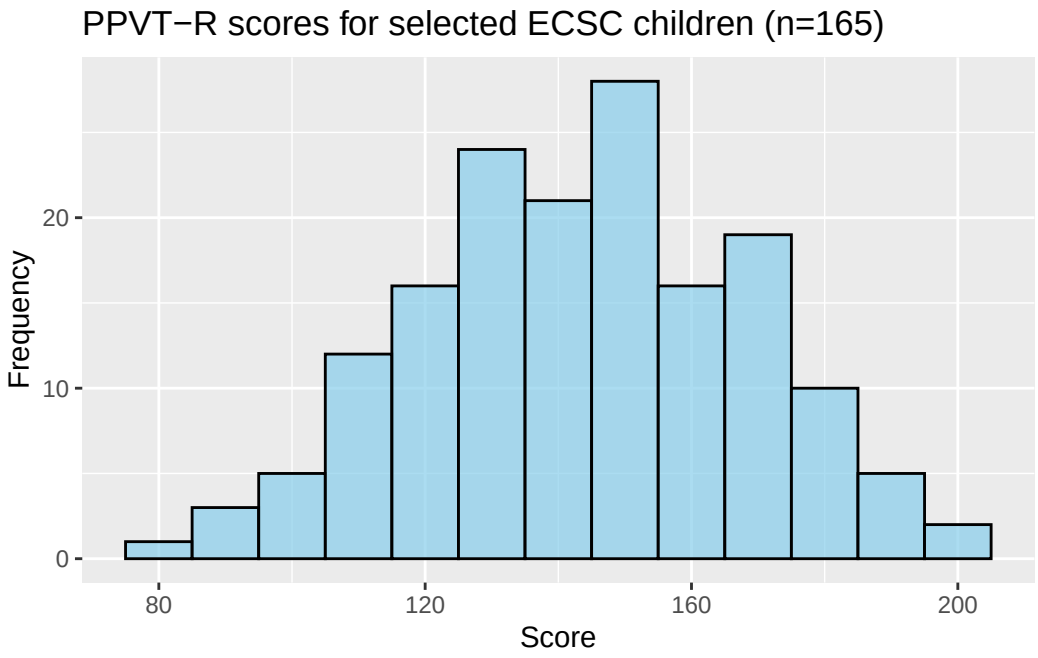
This is manner_freq per speaker..

```
#we should calculate an manner_freq for speakers, and a manner_freq for
 ↪  each story
#Then, lmer: predicting individual manner_freq by overall manner_freq
# m.lm <- lm(manner_freq ~ age, data=all_ec)
# summary(m.lm)
```

**Revision: Plotting PPVT-R per speaker**

```
data <- merged_ec$ppvt_r

# Create a ggplot object with just one variable
p <- ggplot() +
  geom_histogram(aes(x = data), data = data.frame(data), binwidth = 10,
   ↪  fill = "skyblue", color = "black", alpha = 0.7) +  # Customize
   ↪  histogram appearance
  labs(x = "Score", y = "Frequency", title = "PPVT-R scores for selected
   ↪  ECSC children (n=165)")  # Add labels and title
p
```

```
Warning: Removed 3 rows containing non-finite outside the scale range
(`stat_bin()`).
```




### Revision: Speaker as Random Intercept (with story-level data)

The just_id variable is the speaker id for each story, here we control for it and try to derive the predictions for manner_freq and motion count.

For predicting manner_freq, we continue to get a very small negative effect with a low absolute t-value. For predicting motion count overall, we get a strong positive effect of 0.047 per month. However, as several commentators have mentioned, this could be skewed by the total length of the story: we should see if this holds with length controlled for.

```
library(lme4)
m.lmer <- lmer(formula = motion_ct ~ yr + (1 | story) + (1 | just_id) +
 ↪  (1 | age) , data=merged_ec)
summary(m.lmer)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: motion_ct ~ yr + (1 | story) + (1 | just_id) + (1 | age)
   Data: merged_ec

REML criterion at convergence: 976

Scaled residuals:
```

```
    Min      1Q  Median      3Q     Max
-1.5709 -0.4858 -0.0698  0.3835  3.2177

Random effects:
 Groups   Name        Variance Std.Dev.
 just_id  (Intercept) 10.46944 3.2357
 age      (Intercept)  0.01223 0.1106
 story    (Intercept)  4.54709 2.1324
 Residual             11.95904 3.4582
Number of obs: 165, groups:  just_id, 125; age, 56; story, 5

Fixed effects:
            Estimate Std. Error t value
(Intercept)   9.3848     1.2920   7.264
yr            1.1686     0.3981   2.936

Correlation of Fixed Effects:
   (Intr)
yr -0.498
```

```
# Linear mixed model fit by REML ['lmerMod']
# Formula: manner_freq ~ age + (1 | just_id)
#    Data: merged_ec
#
# REML criterion at convergence: -22
#
# Scaled residuals:
#      Min       1Q   Median       3Q      Max
# -2.52713 -0.66234 -0.01424  0.72919  2.14621
#
# Random effects:
#  Groups   Name        Variance Std.Dev.
#  just_id  (Intercept) 0.002008 0.04481
#  Residual             0.044404 0.21072
# Number of obs: 165, groups:  just_id, 125
#
# Fixed effects:
#               Estimate Std. Error t value
# (Intercept)  0.6094984  0.0832619   7.320
# age         -0.0007148  0.0009497  -0.753
#
```

```
# Correlation of Fixed Effects:
#     (Intr)
# age -0.979
```

We see that the motion count is still well predicted by the age with the story_len as a random intercept.

```
library(lme4)
m.lmer <- lmer(formula = motion_ct ~ age + (1 | just_id) + (1 |
 ↪  story_len), data=merged_ec)
summary(m.lmer)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: motion_ct ~ age + (1 | just_id) + (1 | story_len)
   Data: merged_ec

REML criterion at convergence: 1007.8

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.3627 -0.4634 -0.1309  0.2869  2.7712

Random effects:
 Groups    Name        Variance Std.Dev.
 just_id   (Intercept) 15.922   3.990
 story_len (Intercept)  1.898   1.378
 Residual              10.773   3.282
Number of obs: 165, groups:  just_id, 125; story_len, 76

Fixed effects:
            Estimate Std. Error t value
(Intercept)  6.90544    2.17133   3.180
age          0.04240    0.02453   1.728

Correlation of Fixed Effects:
    (Intr)
age -0.975
```

```
#Linear mixed model fit by REML ['lmerMod']
# Formula: motion_ct ~ age + (1 | just_id) + (1 | story_len)
```

```
#    Data: merged_ec
#
# REML criterion at convergence: 1007.8
#
# Scaled residuals:
#     Min      1Q  Median      3Q     Max
# -1.3627 -0.4634 -0.1309  0.2869  2.7712
#
# Random effects:
#  Groups    Name        Variance Std.Dev.
#  just_id   (Intercept) 15.922   3.990
#  story_len (Intercept)  1.898   1.378
#  Residual              10.773   3.282
# Number of obs: 165, groups:  just_id, 125; story_len, 76
#
# Fixed effects:
#             Estimate Std.  Error t value
# (Intercept)  6.90544    2.17133   3.180
# age          0.04240    0.02453   1.728
#
# Correlation of Fixed Effects:
#     (Intr)
# age -0.975
```

Here we show that manner and path counts both increase with age. The rate at which manner verbs increase with age is higher than path verbs, but they are surprisingly balanced overall (0.32 manner verbs/yr vs 0.26 path verbs/yr).

```
library(lme4)
m.lmer <- lmer(formula = manner_ct ~ age + (1 | just_id) + (1 |
 ↪  story_len), data=merged_ec)
summary(m.lmer)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: manner_ct ~ age + (1 | just_id) + (1 | story_len)
   Data: merged_ec

REML criterion at convergence: 925

Scaled residuals:
    Min      1Q  Median      3Q     Max
```

```
-1.5318 -0.4732 -0.1408  0.3013  3.6202

Random effects:
 Groups    Name        Variance Std.Dev.
 just_id   (Intercept) 7.730    2.7803
 story_len (Intercept) 0.656    0.8099
 Residual              8.192    2.8621
Number of obs: 165, groups:  just_id, 125; story_len, 76

Fixed effects:
            Estimate Std. Error t value
(Intercept)  3.60922    1.64688   2.192
age          0.02657    0.01865   1.425

Correlation of Fixed Effects:
    (Intr)
age -0.976
```

```
#MANNER_CT
# Linear mixed model fit by REML ['lmerMod']
# Formula: manner_ct ~ age + (1 | just_id) + (1 | story_len)
#    Data: merged_ec
#
# REML criterion at convergence: 925
#
# Scaled residuals:
#     Min      1Q  Median      3Q     Max
# -1.5318 -0.4732 -0.1408  0.3013  3.6202
#
# Random effects:
#  Groups    Name        Variance Std.Dev.
#  just_id   (Intercept) 7.730    2.7803
#  story_len (Intercept) 0.656    0.8099
#  Residual              8.192    2.8621
# Number of obs: 165, groups:  just_id, 125; story_len, 76
#
# Fixed effects:
#             Estimate Std. Error t value
# (Intercept)  3.60922    1.64688   2.192
# age          0.02657    0.01865   1.425
#
```

```
# Correlation of Fixed Effects:
#     (Intr)
# age -0.976

#PATH_CT
# Linear mixed model fit by REML ['lmerMod']
# Formula: path_ct ~ age + (1 | just_id) + (1 | story_len)
#    Data: merged_ec
#
# REML criterion at convergence: 822
#
# Scaled residuals:
#     Min      1Q  Median      3Q     Max
# -1.8318 -0.6613 -0.1882  0.5120  2.9708
#
# Random effects:
#  Groups    Name        Variance Std.Dev.
#  just_id   (Intercept) 0.03564  0.1888
#  story_len (Intercept) 0.26070  0.5106
#  Residual              7.93866  2.8176
# Number of obs: 165, groups:  just_id, 125; story_len, 76
#
# Fixed effects:
#             Estimate Std. Error t value
# (Intercept)  2.87002    1.10157   2.605
# age          0.02149    0.01254   1.714
#
# Correlation of Fixed Effects:
#     (Intr)
# age -0.978
```

```
library(lme4)
m.lmer <- lmer(formula = motion_ct ~ ppvt_ss + (1 | story),
 ↪  data=merged_ec)
summary(m.lmer)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: motion_ct ~ ppvt_ss + (1 | story)
   Data: merged_ec

REML criterion at convergence: 602.8
```

```
Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.3573 -0.6818 -0.1416  0.4601  4.2290

Random effects:
 Groups   Name        Variance Std.Dev.
 story    (Intercept)  8.526   2.920
 Residual             23.606   4.859
Number of obs: 99, groups:  story, 4

Fixed effects:
            Estimate Std. Error t value
(Intercept) 18.61881    5.24027   3.553
ppvt_ss     -0.05156    0.04129  -1.249

Correlation of Fixed Effects:
        (Intr)
ppvt_ss -0.956
```

**Revision: PPVT-R correlates to manner but not path**

When using age as the predictor variable and controlling for speaker ID, age, and the selected story, one month of age corresponds to 0.062 additional motion verbs in the story (t = 3.1), with 0.028 additional path verbs (t=2.3) and 0.047 additional manner verbs (t=2.9) respectively.

A striking difference between motion and path verb counts was found when using the PPVT-R as a predictor. When using PPVT-R scores as the predictor variable and controlling for the speaker ID, the story length, and the selected story, one point increase in the PPVT-R corresponds to 0.032 additional manner verbs per story (t=2.6), but only 0.0076 additional path verbs (t=0.83). This suggests that manner verbs represent a unique lexical category that is independently predicted by PPVT-R. This is super interesting and should be explored further.

```
library(lme4)
m.lmer <- lmer(formula = manner_ct ~ ppvt_r + (1 | just_id) + (1 |
 ↪  story_len) + (1 | story), data=merged_ec)
summary(m.lmer)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: manner_ct ~ ppvt_r + (1 | just_id) + (1 | story_len) + (1 | story)
```

```
   Data: merged_ec

REML criterion at convergence: 877.9

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.6575 -0.5258 -0.0446  0.4077  3.7101

Random effects:
 Groups    Name        Variance Std.Dev.
 just_id   (Intercept) 4.5900   2.1424
 story_len (Intercept) 0.4867   0.6976
 story     (Intercept) 5.2109   2.2827
 Residual              7.5287   2.7438
Number of obs: 162, groups:  just_id, 122; story_len, 75; story, 4

Fixed effects:
            Estimate Std. Error t value
(Intercept)  1.92280    2.09256   0.919
ppvt_r       0.03222    0.01202   2.681

Correlation of Fixed Effects:
       (Intr)
ppvt_r -0.823
```

```
# Formula: manner_ct ~ ppvt_r + (1 | just_id) + (1 | story_len) + (1 |
 ↪  story)
#    Data: merged_ec
#
# REML criterion at convergence: 877.9
#
# Scaled residuals:
#     Min      1Q  Median      3Q     Max
# -1.6575 -0.5258 -0.0446  0.4077  3.7101
#
# Random effects:
#  Groups    Name        Variance Std.Dev.
#  just_id   (Intercept) 4.5900   2.1424
#  story_len (Intercept) 0.4867   0.6976
#  story     (Intercept) 5.2109   2.2827
#  Residual              7.5287   2.7438
```

```
# Number of obs: 162, groups:  just_id, 122; story_len, 75; story, 4
#
# Fixed effects:
#             Estimate Std. Error t value
# (Intercept)  1.92280    2.09256   0.919
# ppvt_r       0.03222    0.01202   2.681
#
# Correlation of Fixed Effects:
#        (Intr)
# ppvt_r -0.823

# Formula: path_ct ~ ppvt_r + (1 | just_id) + (1 | story_len) + (1 |
 ↪  story)
#    Data: merged_ec
#
# REML criterion at convergence: 805.7
#
# Scaled residuals:
#     Min      1Q  Median      3Q     Max
# -1.9593 -0.5979 -0.1356  0.4988  2.5447
#
# Random effects:
#  Groups    Name        Variance Std.Dev.
#  just_id   (Intercept) 0.4192   0.6475
#  story_len (Intercept) 0.8407   0.9169
#  story     (Intercept) 0.6883   0.8297
#  Residual              6.7419   2.5965
# Number of obs: 162, groups:  just_id, 122; story_len, 75; story, 4
#
# Fixed effects:
#             Estimate Std. Error t value
# (Intercept) 3.759087   1.398066   2.689
# ppvt_r      0.007612   0.009138   0.833
#
# Correlation of Fixed Effects:
#        (Intr)
# ppvt_r -0.937
```

### Generate word frequency graphs (not used)

```
# concat_ec5 <- apply(ec5, 2, function(x) paste(x, collapse = " "))
#
# #take all words, frequency from somewhere
# #plot word frequency against age and gender.
# #strsplit(ec5$71M_1083_YR1, split=" ")
# #concat_ec5 <- concat_ec5 %>% separate_rows(column1, sep = " ")
#
# #character 1:48 array
# #concat_ec5 <- separate_rows(concat_ec5, x, sep = " ")
# concat_ec5_v <- as.vector(concat_ec5)
# ec5_split_words <- unlist(str_split(concat_ec5_v, "\\s+"))
# word_counts <- table(ec5_split_words)
# word_counts <- word_counts[order(-word_counts)]
#
# #histogram of word_counts for ec5
# #1: get frequency from word_counts
# #2: get overall frequency
#
# # Vector of names to remove
# names_to_remove <- c("NA", ".")
# # Remove entries by name
# word_counts <- word_counts[!names(word_counts) %in% names_to_remove]
#
# total_sum <- sum(word_counts)
# word_percentages <- (word_counts / total_sum) * 100
# #barplot of word percentages
# #barplot(word_percentages, main = "Word Counts", xlab = "Words", ylab
 ↪  = "Frequency")
# word_percentages <- as.data.frame(word_percentages)
# word_percentages
```

```
library(tidyverse)

#table

# word_freq <- read_tsv("google-freq.tsv", FALSE)
#
# sum <- sum(word_freq$X3)
```

```
#
# # Add a new column with the percentage of each row's total
# gg_word_freq <- word_freq %>%
#   mutate(gg_freq = (X3 / sum) * 100)
#
# df <- merge(word_percentages, gg_word_freq, by.x="ec5_split_words",
 ↪  by.y="X1")
#
# df
```

```
# k1 <- as.numeric(df$Freq)
# k2 <- as.numeric(df$gg_freq)
# sp <- ggplot(df, aes(k1, k2, label=ec5_split_words))
# round(df$gg_freq, digits=5
```

```
#print(k1)
# sp + geom_point() + xlim(0,0.1) + ylim(0,0.2)
```

```
#
# FLAG = FALSE
# sp + geom_label(size=3, check_overlap=FLAG) + xlim(0,1000) +
 ↪  ylim(0,2000)
  # geom_text(size=3) +  # Adjust text position
  # labs(x = "freq", y = "gg_word_freq", title = "observed 5yo. ECSC
   ↪  frequencies against standard (Google) frequencies") +
  # scale_x_continuous(limits = c(0, 0.1)) +   # Set x-axis limits
  # scale_y_continuous(limits = c(0, 0.25))
```