Think to speak: a Bayesian statistical analysis

Luiz Max Carvalho
School of Applied Mathematics, Getulio Vargas Foundation
lmax.fgv@gmail.com

May 2025

1 Background

In this report we will employ Bayesian statistical methods to address a few questions about the so-called "manner verbs". We will analyse data from six languages: Spanish, Italian, French, German, Brazilian Portuguese (BP) and English. The goal is to understand differences between languages in the proportion and rate of occurrence of manner verbs, as well as understand within-language (i.e. between-speakers) variation when disaggregated data per speaker is available. Specifically, the experimental design consists of eliciting oral narratives (raw data) based on a wordless picture book - Frog, where are you? (Mayer 1969). This children's book presents a total of 24 scenes that indicate a sequence of events in which the main characters, a boy and his dog, engage in search of an escaped frog. We used the clause as our unit of analysis and identified the motion events' path and manner components, following Talmy's description (Talmy 1991).

First, we will show the data under analysis and then we will fit a few statistical models using Stan.

2 Part I: The data

Part of the data analysed here is from Hijazo-Gascon and Ibarretxe-Antunano (2013) ("Tabla I" therein). The disaggregated data on Brazilian Portuguese ($N_1 = 14$ speakers) and English ($N_2 = 12$ speakers) were collected in the present study.

Let us first look at the aggregated data, which contains: the total types of 'movement verbs', number of types of manner verbs, the number of speakers measured for each language and the number of occurrences of manner verbs in the texts analysed. Table 1 contains the aggregated data we will analyse.

Figure 1 helps to visualise the disaggregated (per speaker) data. Notice there is substantial variability between speakers in their use of movement verbs in general and manner verbs in particular. A proper statistical treatment will take that variability into account when assessing between-language differences.

Table 1: Aggregated on manner verb usage in six languages.

Language	Total types	Manner types	No. speakers	Occurrences (manner)
Spanish	39	7	12	11
Italian	60	10	12	16
French	53	3	12	14
German	67	23	12	62
BP	21	5	14	31
English	41	20	12	64

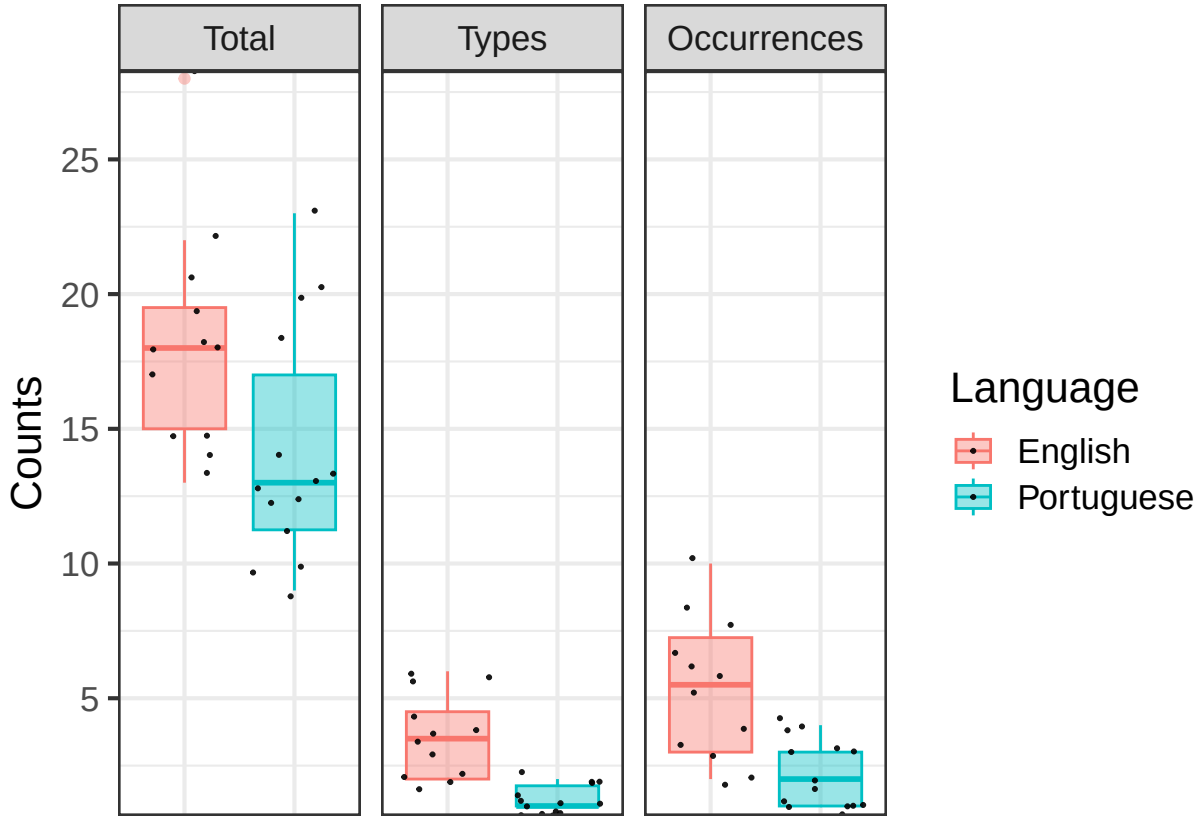

Figure 1: Boxplots of disaggregated counts for total types of movement verbs, types manner verbs and occurrences of manner verbs.

3 Part II: analysis of aggregated data

We will start by analysing the counts aggregated by language and thus ignore within-language variation for now.

3.1 Question 1: Do languages differ with respect to the proportion of manner verbs?

The first question we shall tackle is whether language differ in the proportion of motion (movement) verbs that are also manner verbs. To accomplish this, let us first fix some notation: let n_i denote the number of (total) movement verbs observed in language i and let x_i be the number of these verbs that are manner verbs. Statistically, we are thus interested in estimating the proportion, p_i, of movement verbs that are manner verbs for language i. A reasonable statistical model for this problem is the binomial distribution:

$$
\begin{aligned}
x_i &\sim \text{Binomial}(n_i, p_i); \\
p_i &= \frac{1}{1 + \exp(-[\theta + \delta_i])}; \\
\theta &\sim \text{Normal}(0, 5); \\
\delta_i &\sim \text{Normal}(0, 1)\,.
\end{aligned}
$$

where we parametrise the normal distribution in terms of mean and standard deviation. Let us break down the model: the proportion p_i is modelled as having two components: an overall effect, θ, which captures the overall proportion of manner verbs across all languages and a per-language effect, δ_i, which reflects the modification each language makes to the "grand mean" represented by θ. This is what is commonly called in

Table 2: Posterior mean and 95% BCIs for the proportion of manner verbs.

Language	Mean	Median	Lower	Upper
Spanish	0.18	0.18	0.09	0.31
Italian	0.17	0.17	0.09	0.27
French	0.08	0.08	0.03	0.16
German	0.34	0.33	0.23	0.45
BP	0.23	0.23	0.10	0.41
English	0.46	0.46	0.32	0.61

the statistical literature a "random effects" model. Other names, which mean the same thing, are "hierarchical model" and "multilevel model". The transformation $1/(1+\exp(-x))$ serves the purpose of mapping θ and δ_i to the space of probabilities, $(0, 1)$, and is called the *inverse logit* function.

Now that we have a model, let us fit it to the aggregated data and look at the results. We will use Stan, via cmdstanr to fit this model using four chains of 1000 iterations (500 warmup, 500 sampling) each. Details on the computing environment are given at the end of this document. We report the posterior mean and 95% Bayesian credible intervals (BCI) for the estimates of p_i for each language.

The results from fitting the Bayesian model discussed above are shown in Table 2 and clearly indicate that Spanish and Italian have very similar proportions of manner verbs, while French and Brazilian Portuguese (BP) show remarkably lower proportions. Moreover, as expected by theory, German and English show similar proportions; while around 1/3 of all movement verbs are manner verbs for German, in English this proportion is a bit higher. These results can also be visualised in Figure 2, which shows the breadth of the uncertainty around p for each language. The figure allows us to visualise patterns that might be hard to spot in the table. For instance, it becomes immediately apparent that while German and English have similarly higher proportions of manner verbs, the uncertainty for German is bigger.

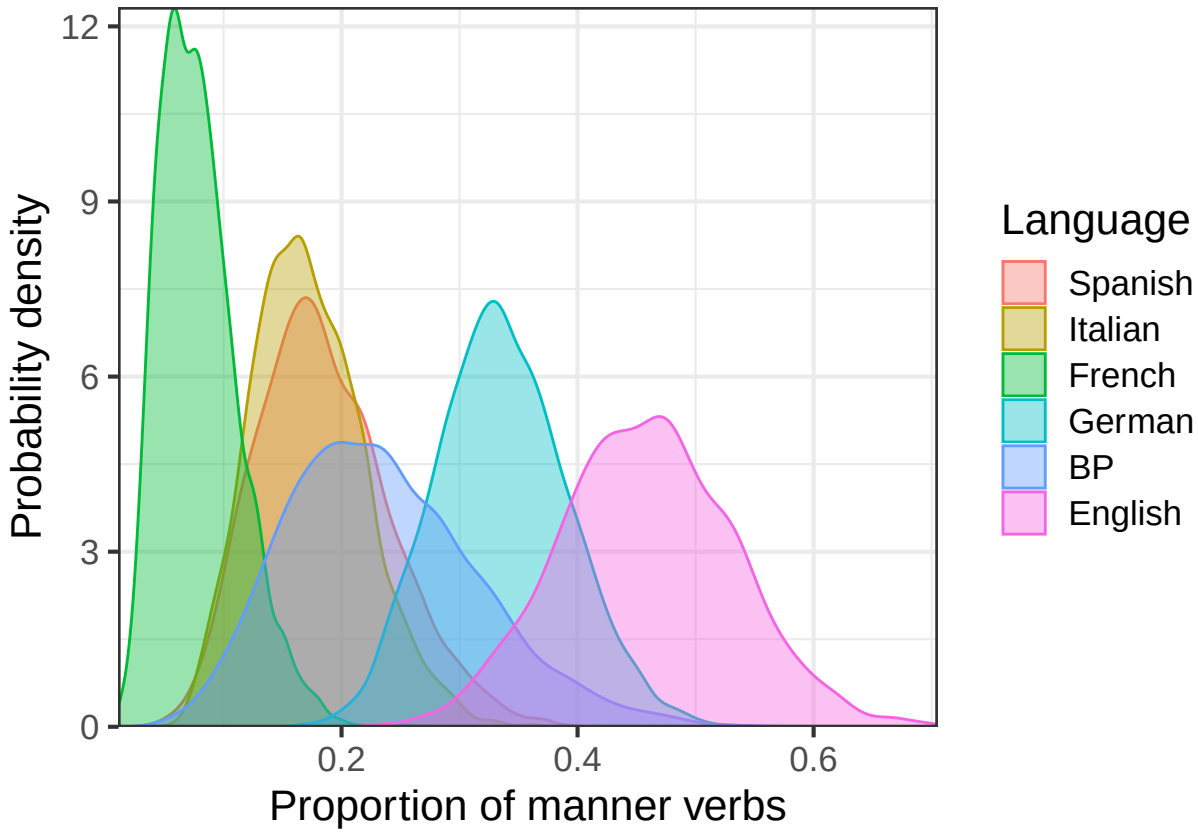

Figure 2: Posterior density of the proportion of manner verbs for each language.

3.2 Question 2: Do languages differ with respect to the average rate of manner verbs?

Having approached the question of the proportion of all movement verbs which are manner verbs, we now turn to the question of whether languages differ with respect to the *rate* of occurrence of manner verbs. To model this problem, we will use a Poisson distribution for the numbers of occurrences. We also account for the fact that different languages have different numbers of speakers by using the n_i as an *offset* and thus estimate a *per speaker* rate. Let y_i be the total number of times a manner verb was mentioned across all speakers. We are then interested in the rate λ_i of occurrence of manner verbs in language i. This setup leads to a model that reads:

$$
\begin{aligned}
y_i &\sim \text{Poisson}(n_i \lambda_i); \\
\lambda_i &= \exp(\alpha + \eta_i); \\
\alpha &\sim \text{Normal}(0, 5); \\
\eta_i &\sim \text{Normal}(0, 1)\,.
\end{aligned}
$$

This model is similar in spirit to the one in the previous section; we have an overall (log) rate, α, and language-specific effect η_i. Results of fitting this model are shown in Table 3 and Figure 3. As expected from theory, German and English show the highest rates, while the romance languages have lower rates, albeit with substantial variation between them. Brazilian Portuguese and Italian show the highest rates, Spanish shows a lower rate. French shows a rate compatible with Italian, with more uncertainty (wider BCIs).

Table 3: Posterior mean and 95% BCIs for the rate of occurrence of manner verbs estimated from aggregated data.

Language	Mean	Median	Lower	Upper
Spanish	0.98	0.96	0.52	1.61
Italian	1.36	1.33	0.80	2.09
French	1.21	1.18	0.69	1.88
German	5.09	5.06	3.86	6.39
BP	2.22	2.20	1.49	3.05
English	5.26	5.24	4.06	6.62

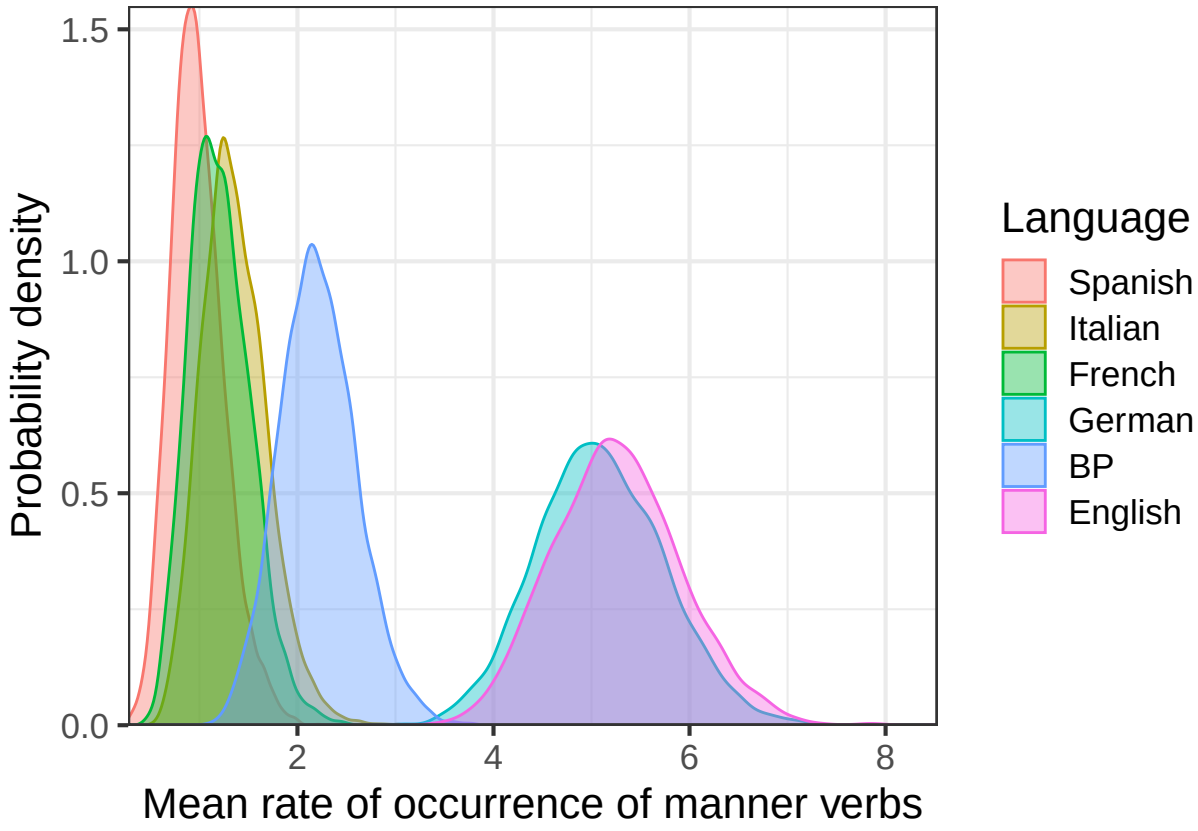

Figure 3: Posterior density for the rate of occurence of manner verbs estimated from aggregated data.

4 Part III: including disaggregated (by speaker) data

As show in Part I, we have disaggregated data per speaker for Brazilian Portuguese and English, and would like to take inter-speaker (i.e. within language) variation into account in our analysis. This is important because performing inference on aggregated counts underestimates the heterogeneity between speakers and thus exaggerates differences between languages. In this section we analyse all of the available data: language-level (aggregated counts) for Spanish, Italian, French and German and speaker-level counts for Brazilian Portuguese and English.

Table 4: Posterior mean and 95% BCIs for the rate of occurrence of manner verbs estimated from disaggregated data.

Language	Mean	Median	Lower	Upper
Spanish	0.97	0.94	0.51	1.54
Italian	1.37	1.35	0.82	2.09
French	1.21	1.17	0.67	1.94
German	5.05	5.01	3.84	6.40
BP	1.88	1.78	0.97	3.46
English	4.39	4.12	2.27	7.58

4.1 Question 3: Do languages differ with respect to the mean rate of occurrence of manner verbs taking inter-speaker variation into account?

We will modify our Poisson model to accommodate inter-speaker variation. Let y_{ij} be the number of manner verbs used by speaker j, $j = 1, 2, J_i$ in language i Then a model that includes speaker-level random effects is:

$$\begin{aligned} y_{ij} &\sim \text{Poisson}(\lambda_{ij}); \\ \lambda_{ij} &= \exp(\alpha + \eta_i + \epsilon_j); \\ \alpha &\sim \text{Normal}(0, 5); \\ \eta_i &\sim \text{Normal}(0, 1); \\ \epsilon_j &\sim \text{Normal}(0, 1)\,. \end{aligned}$$

Notice that the structure is very similar to what we already had, but now we include a speaker-level coefficient, ϵ_j, which accounts for variations between individuals. For compatibility with the previous analysis, we will first use $\lambda_i := \exp(\alpha + \eta_i)$ as our target quantity. Results are presented in Table 4 and Figure 4 and show the value of including speaker-level data. When within-language variation is taken into account, the posterior mean of rate of occurrence of manner verbs for English drops from around 5.2 to 4.3, lower than German. This indicates that one or more individuals might be pulling the mean upwards when the data are aggregated. Albeit less pronounced the same pattern of reduction (2.2 to 1.9) is present for Brazilian Portuguese. In addition, as expected, uncertainty in the estimates increased, because we are now propagating uncertainty about individual (speaker) level parameters to our language-level estimates. English seems to have much more uncertainty than BP, presumably due to more variability in the speaker-level data (Figure 1).

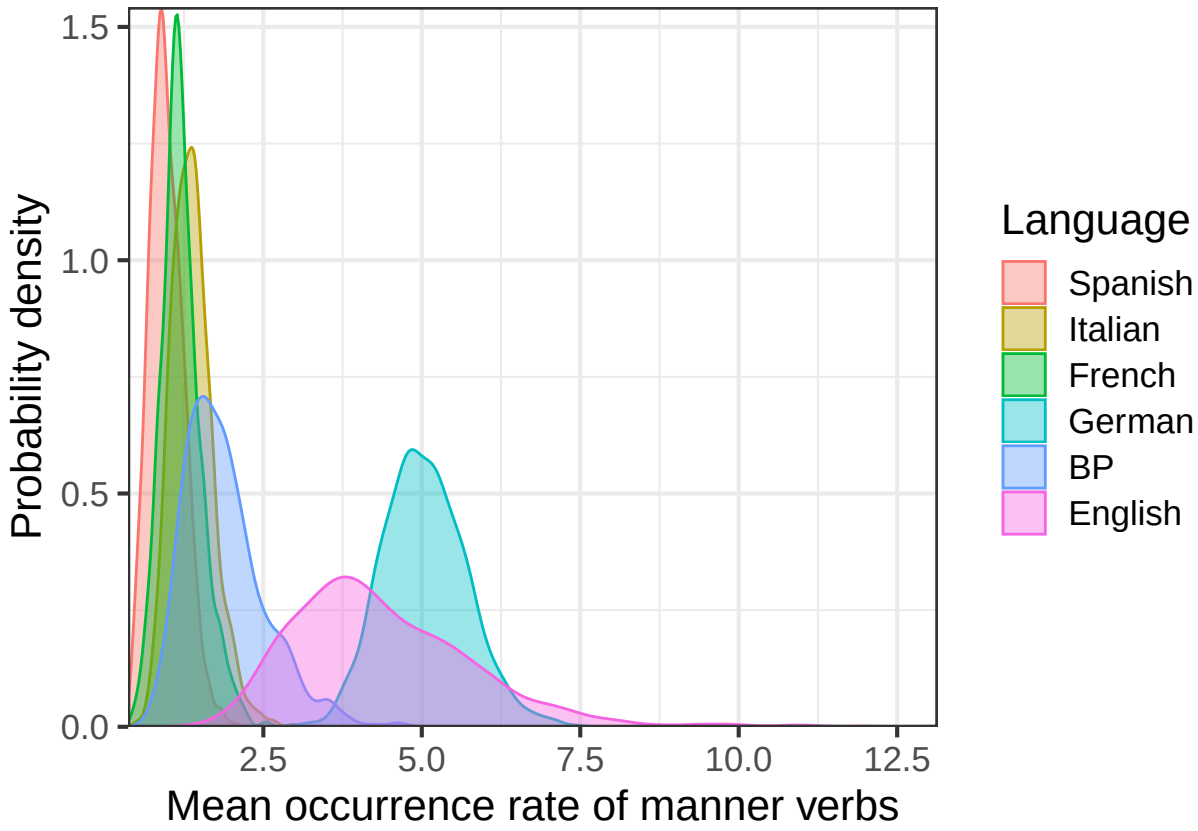

Figure 4: Posterior density for the rate of occurence of manner verbs estimated from disaggregated data.

To finish off our analysis, we will plot the individual-level rates, $\lambda_{ij} = \exp(\alpha + \eta_i + \epsilon_j)$, for both English and Brazilian Portuguese in Figure 5. These plots show that (i) there is indeed substantial within-language variation in the rate of manner verbs and; (ii) most individuals are close to the corresponding language-level mean, with only one speaker in English (speaker 3) substantially from the group median. This is evidenced by the BCI not do not cover the respective language-level median - shown with horizontal dashed lines. It is important to note that the speakers included in each language are **not** the same individuals.

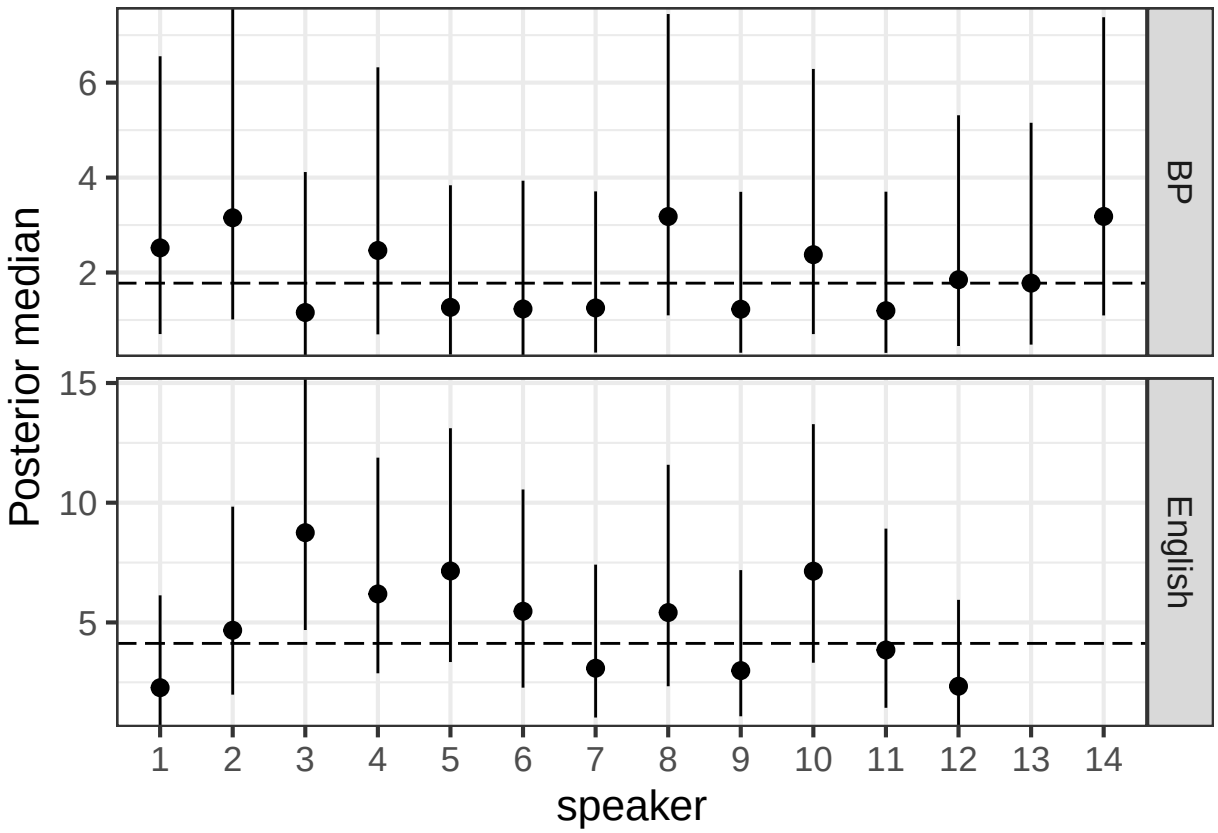

Figure 5: Posterior median and BCI for the rate of occurence of manner verbs by speaker. Horizontal lines mark the language-level median.

4.2 A note on uncertainty and the need for statistical modelling

Experimental data are rarely measured without error. Moreover, the very biological, socio-economic and behavioural processes giving rise to the data are subject to underlying variation. This induces uncertainty when, for instance, comparing population groups for differences in the rate of a given disease or literary genres for differences in the usage of a given class of words. Statistics is the language that allows us to describe and quantify this uncertainty.

An example will hopefully motivate the need for accounting for uncertainty: it is common knowledge that the average man is taller than the average woman. But when does this difference arise? Figure 6 shows data from the World Health Organisation on the growth of boys and girls.

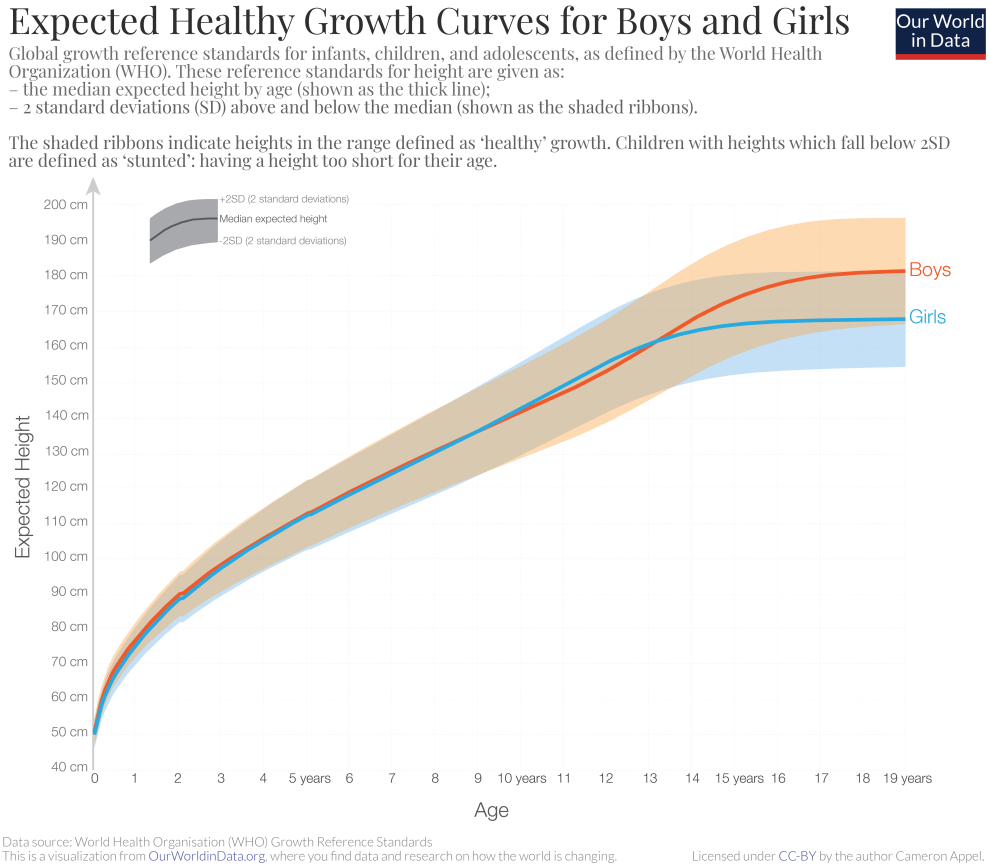

Figure 6: Expected height of boys/girls as a function of age. Taken from Our World in Data (https://ourworldindata.org/human-height). The solid lines show the median height and the ribbons show the median plus or minus two standard deviations.

At 11 years old, girls are typically more than two centimeters taller than boys, but they tend to stop growing a few years earlier. Boys achieve their adult height at around 18 years old, achieving, on average, 13 centimeters more than than girls.

Crucially, however, not all children grow at the same rate. The ribbons around the median growth lines in Figure 6 represent two standard deviations around the expected height. Heights which fall within two standard deviations of the median are considered to be 'healthy growth'. The ribbons show that boys really only "decouple" from girls after 16 years of age, as evidenced by the fact that the red curve (median boy height) only leaves the blue uncertainty band (girls) at the very end. This means that the range for 'healthy' growth in girls and boys, i.e. the uncertainty around the median, substantially overlaps. This highlights that when analysing estimates drawn from data one ought to also consider *interval* estimates (the ribbon), beyond *point* estimates (the lines).

4.3 A note on priors

In a Bayesian analysis, the choice of prior distribution for the unobserved quantities ("parameters") in model is crucial. Here we chose a Normal(0, 5) prior for the intercepts and Normal(0, 1) (standard normal) priors for the random effects. These are common prior choices in the Bayesian literature, called weakly-informative priors, and lead to well-behaved inferences without impacting the results in appreciable ways. See Gelman et al. (2013) for more details.

5 Conclusions

In this report we have fitted a few simple statistical models to answer questions about the pattern of occurence of manner verbs across languages. From a statistical perspective, we can draw the following conclusions:

- Spanish, Italian and BP are similar in their proportion of manner verbs;
- German and English are also similar. French has a lower proportion than all languages;
- Italian, French and BP have similar rates of occurrence of manner verbs, with Spanish having a slightly lower rate;
- English and German have the highest rates;
- Accounting for speaker-level variation reveals that a few individuals might pull the language-level rate upwards;
- Analysis of disaggregated data brings the rate of occurrence of manner verbs down for both English (closer to German than in the aggregated analysis) and Brazilian Portuguese.

6 References

Gelman, Andrew, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. 2013. *Bayesian Data Analysis*. CRC press.

Hijazo-Gascon, Alberto, and Iraide Ibarretxe-Antunano. 2013. "Las Lenguas Romanicas y La Tipologia de Los Eventos de Movimiento." *Romanische Forschungen* 125 (4): 467–94.

Mayer, Mercer. 1969. *Frog, Where Are You?* New York: Dial Books for Young Readers.

Talmy, Leonard. 1991. "Path to Realization: A Typology of Event Conflation." In *Annual Meeting of the Berkeley Linguistics Society*, 17:480–519. 1.

7 Computing environment

This document has been composed using the statistical computing language R in a GNU/Linux (Ubuntu) operating system. See below for a complete list of the packages used.

```
sessionInfo()
```

```
## R version 4.2.2 (2022-10-31)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Pop!_OS 22.04 LTS
##
## Matrix products: default
## BLAS:   /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.10.0
## LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.10.0
##
## locale:
##  [1] LC_CTYPE=en_GB.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_GB.UTF-8        LC_COLLATE=en_GB.UTF-8
##  [5] LC_MONETARY=en_GB.UTF-8    LC_MESSAGES=en_GB.UTF-8
##  [7] LC_PAPER=en_GB.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_GB.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
```

```
##  [1] scales_1.3.0        posterior_1.6.1     kableExtra_1.4.0
##  [4] lubridate_1.9.3     forcats_1.0.0       stringr_1.5.1
##  [7] dplyr_1.1.4         purrr_1.0.2         readr_2.1.5
## [10] tidyr_1.3.1         tibble_3.2.1        tidyverse_2.0.0
## [13] reshape2_1.4.4      ggplot2_3.5.1       rstan_2.32.6
## [16] StanHeaders_2.32.7  cmdstanr_0.8.1.9000
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.14          svglite_2.1.3        QuickJSR_1.1.3
##  [4] ps_1.9.0             digest_0.6.37        V8_4.4.2
##  [7] R6_2.6.1             plyr_1.8.9           backports_1.5.0
## [10] stats4_4.2.2         evaluate_0.23        pillar_1.10.1
## [13] rlang_1.1.6          curl_5.2.1           data.table_1.17.0
## [16] rstudioapi_0.16.0    checkmate_2.3.2      rmarkdown_2.26
## [19] labeling_0.4.3       loo_2.7.0            tinytex_0.50
## [22] munsell_0.5.1        compiler_4.2.2       xfun_0.43
## [25] systemfonts_1.0.6    pkgconfig_2.0.3      pkgbuild_1.4.4
## [28] htmltools_0.5.8.1    tidyselect_1.2.1     tensorA_0.36.2.1
## [31] gridExtra_2.3        bookdown_0.39        codetools_0.2-20
## [34] matrixStats_1.5.0    viridisLite_0.4.2    tzdb_0.4.0
## [37] withr_3.0.2          grid_4.2.2           distributional_0.5.0
## [40] jsonlite_1.9.1       gtable_0.3.6         lifecycle_1.0.4
## [43] magrittr_2.0.3       RcppParallel_5.1.10  cli_3.6.5
## [46] stringi_1.8.4        farver_2.1.2         xml2_1.3.6
## [49] generics_0.1.3       vctrs_0.6.5          tools_4.2.2
## [52] glue_1.8.0           hms_1.1.3            processx_3.8.6
## [55] abind_1.4-8          parallel_4.2.2       fastmap_1.2.0
## [58] yaml_2.3.8           timechange_0.3.0     inline_0.3.19
## [61] colorspace_2.1-1     knitr_1.46
```