Levshina, Natalia. In press. Comparing Bayesian and frequentist models of language variation: The case of help + (to) Infinitive. In Ole Schützler & Julia Schlüter (eds.), *Data and methods in corpus linguistics – Comparative Approaches*. Cambridge: Cambridge University Press.

Abstract

This chapter compares frequentist (maximum likelihood) and Bayesian approaches to mixed-effect logistic regression, which is *de facto* the standard method for modelling linguistic variation. The main conceptual differences between frequentist and Bayesian inference are discussed, and the key components of Bayesian inference – the priors, data (evidence) and posteriors – are explained using a simple example. We also discuss the epistemological and practical advantages of Bayesian inference, such as the ability of testing the research hypothesis directly, instead of trying to reject the null hypotheses, and the use of informative priors, which helps to fit adequate models in difficult cases. The Bayesian approach to generalized mixed-effect models is illustrated by a multifactorial case study of *help* + (*to*-)infinitive in U.S. magazines. Statistical analyses are performed with the help of *brms*, a popular R package for Bayesian regression The paper shows also how to recycle the results of previous analyses as informative priors in a new Bayesian model. This enables the linguist to use small samples, building on previous research.

Keywords: Bayesian regression, frequentist statistics, generalized mixed-effect models, multifactorial grammar, sample size, American English, *help*

1. Aims of this paper

The present paper compares the frequentist and Bayesian approaches to modelling language variation. The frequentist approach is often referred to as classical statistics, sampling theory

statistics, maximum likelihood estimation (especially in the context of regression modelling), or null hypothesis significance testing. Although frequentist methods are still predominant in most fields, some statisticians are very optimistic about the future of Bayesian methods. For example, John Kruschke (2011b: 272) writes, "whereas the 20th century was dominated by NHST [null hypothesis significance testing - Author], the 21st century is becoming Bayesian".

But what are the benefits of Bayesian statistics and why are some people so enthusiastic about it? This paper aims to demonstrate the main advantages of Bayesian inference. It is targeted mostly at those linguists who are familiar already with the traditional regression analysis (see Chapters … and … in this volume), but are curious about the Bayesian alternative.

I will focus on multiple logistic regression with mixed effects. This is probably the most popular method of investigating language variation in contemporary corpus linguistics. To illustrate the main differences between Bayesian and frequentist logistic regression, I will discuss the use of the bare and *to*-infinitive with the verb *help*. This alternation can be exemplified by the following two book titles:

(1) a. Beginning with Birding: 25 tips to **help** you **become** a better birdwatcher[1]

b. How to Become an Amazing Couple: Daily Simple Steps that Will **Help** you **to Become** a Better Lover[2]

This alternation represents one of the few cases where the choice between the bare and *to*-infinitive is still possible in Present Day English, although it is constrained by numerous contextual factors. Also, the *to*-infinitive after *help* is becoming increasingly rare, especially in American English (see Section 3).

Using examples of this alternation from the Magazines subsection of the Corpus of Contemporary American English (Davies 2008–), I discuss here the main distinctive characteristics of Bayesian regression. First, it provides the researcher with an opportunity to

[1] by Bob Hole (2018)
[2] by Oana Nicolau (2018)

test the research hypothesis directly, instead of trying to reject the null hypothesis. The typical research hypothesis in regression models of language variation is that certain contextual factors increase or decrease the chances of one variant in comparison with the other(s). Most importantly, the Bayesian approach does not rely on *p*-values and does not encourage binary decisions, enriching our knowledge about the impact of the contextual factors.

Second, one can use information from previous research as priors for subsequent models. This allows the researcher to use smaller samples, which may be attractive in some areas in which data are costly – for example, in studies based on spoken and multimodal corpora, investigations of rare typological phenomena, and experimental linguistics. Recycling one's knowledge in this way can also help us to overcome the recent crisis of reproducibility (Goodman et al. 2016) because the plausibility of the old findings can be estimated in the light of new data when these findings are incorporated into a new model (van de Schoot et al. 2014). The priors also help to avoid problems, such as loss of statistical power, overfitting and convergence issues, which often arise when one fits generalized mixed-effect models with complex structure.

This chapter is structured as follows. First, I will present the main conceptual differences between frequentist and Bayesian statistics (Section 2). In Section 3, I will discuss the previous studies of *help* + (*to*-)infinitive. After that, I will describe the data (Section 4) and present several models. Section 5 discusses a frequentist model and a Bayesian model with weak informative priors – both based on a large dataset. Section 6 describes a frequentist model and a Bayesian model with strong informative priors based on a small dataset. Section 7 summarizes the main ideas and suggests some literature for further reading.

The analyses can be reproduced with the help of the data and R code provided in the supplementary materials.[3] Note that the results will be slightly different every time you use the Bayesian approach because of the Monte Carlo sampling from the posterior (see Section 2).

[3] Available from https://github.com/levshina/FreqBayes

2. Frequentist and Bayesian inference in contrast

Although the overwhelming majority of statistical models in linguistics have been created with the help of frequentist methods, there are a few recent studies that employ Bayesian regression, e.g. Vasishth et al. (2013), who analyse how Chinese speakers process relative clauses, Scrivner's (2015) investigation of VO and OV word order patterns in Latin and Old French infinitival complements, and Levshina's (2016) multifactorial analysis of English permissive constructions. It is argued that there are some epistemological and practical advantages in using Bayesian models. This section discusses the differences between the approaches.

When fitting a model of language variation, we want to know which of available factors help us to predict the linguistic choices of speakers, and which are irrelevant. In order to answer these questions, we need inference. In frequentist statistics, inference involves null hypothesis significance testing. The null hypothesis (e.g. no correlation, no difference, no impact on the choice between the variants) is rejected when the probability of observing the test statistic and more extreme values under the null hypothesis is smaller than some pre-determined level (usually 0.05). In other words, we are interested in the likelihood of data given the null hypothesis, or P (Data|H_0).

In contrast, Bayesian inference allows us to estimate the probability of the research hypothesis given the data, or P (H_1|Data). But how can we estimate this probability? The trick is to apply the famous Bayes rule P (A|B) = P (B|A) P (A) / P (B), which allows one to compute the conditional probability of event A given event B if one knows the probability of B given A, the probability of A and the probability of B. By itself, Bayes rule has nothing particularly Bayesian about it. For example, one can use it to estimate the probability of a patient having COVID-19 if the test is negative, or P (COVID-19|negative). One can easily do it if one using the following probabilities:

- probability of a negative test if the patient has COVID-19, or P (negative|COVID-19);
- probability of COVID-19 in the population, or P (COVID-19);
- probability of a negative test, or P (negative).

Using Bayes rule in this way does not make you Bayesian. It is only when we substitute the concepts of data and hypotheses for A and B, we switch our thinking to the Bayesian mode.

According to Bayes rule, if we have some research hypothesis *H* and some Data, the probability of the hypothesis given the data, or *P* (H|Data), is proportional to the prior probability of the hypothesis *P* (H) and the conditional probability of the data given the hypothesis *P* (Data|H).

(2) $P(H|Data) \propto P(H)\, P(Data|H)$

where ∝ means 'proportional to'. The probability of the hypothesis before the data are taken into account is called the prior probability, or simply prior. The probability of the data given the hypothesis is called likelihood. Finally, the probability of the hypothesis given the data, or, in other words, the modified expectations after the data are taken into account, are called the posterior probability, or just posterior. Therefore, the formula in (2) is equivalent to the following expression:

(3) Posterior ∝ Prior x Likelihood

All this implies important differences between the frequentist and Bayesian approaches in our thinking about the world and our knowledge. In frequentist statistics, probable events are the ones that happen (or are expected to happen) frequently. Let us take a coin. If the coin is fair, this means that the probability of both heads and tails is 50%. If we toss the coin many, many times, we expect to get the heads 50% of the times and the tails 50% of the times. The more times we repeat the experiment, the closer the observed proportions of heads and tails will be to these numbers. Frequentist statistics is based on the idea that the population parameter is an unknown but fixed constant. If an experiment is repeated an infinite number of times, one obtains the true value.

Bayesian statisticians, in contrast, think about all unknown parameters as random variables, which can be represented as probability distributions, in which some of the values are believed to be more plausible than others. Consider Figure 1, which displays several possible priors, likelihoods and posteriors. Let us start with the priors, which are shown in the

top row. They represent our beliefs about the value of *x* (the horizontal axis) before we look at the data. The graphs are probability density plots, which show which values are more likely and which are less. Imagine that the priors on the left represent the beliefs of Ann that the verb *help* in some variety of English is equally frequently used with the bare and the *to*-infinitive. This means that the most likely value is *P* (bare) = 0.5. She is not very committed to this belief, however. She does not exclude a range of other values, including those close to 0 or 1. Such priors are called weak.

Similar to Ann, Bill believes that both variants are equally likely, i.e. the most plausible value is *P* (bare) = 0.5. Yet, he is more sceptical of scenarios where one or the other variant is predominant. His priors, which are shown in the centre, are therefore stronger than Ann's.

Finally, Clara, whose priors are on the right, has no idea at all. All values look equally possible to her. Such 'agnostic' priors are called 'flat', 'non-informative', as opposed to informative priors of Ann and Bill. Strictly speaking, the word 'non-informative' is a misnomer. Non-informative priors do carry information - that is, information about ignorance as the current state of our knowledge.

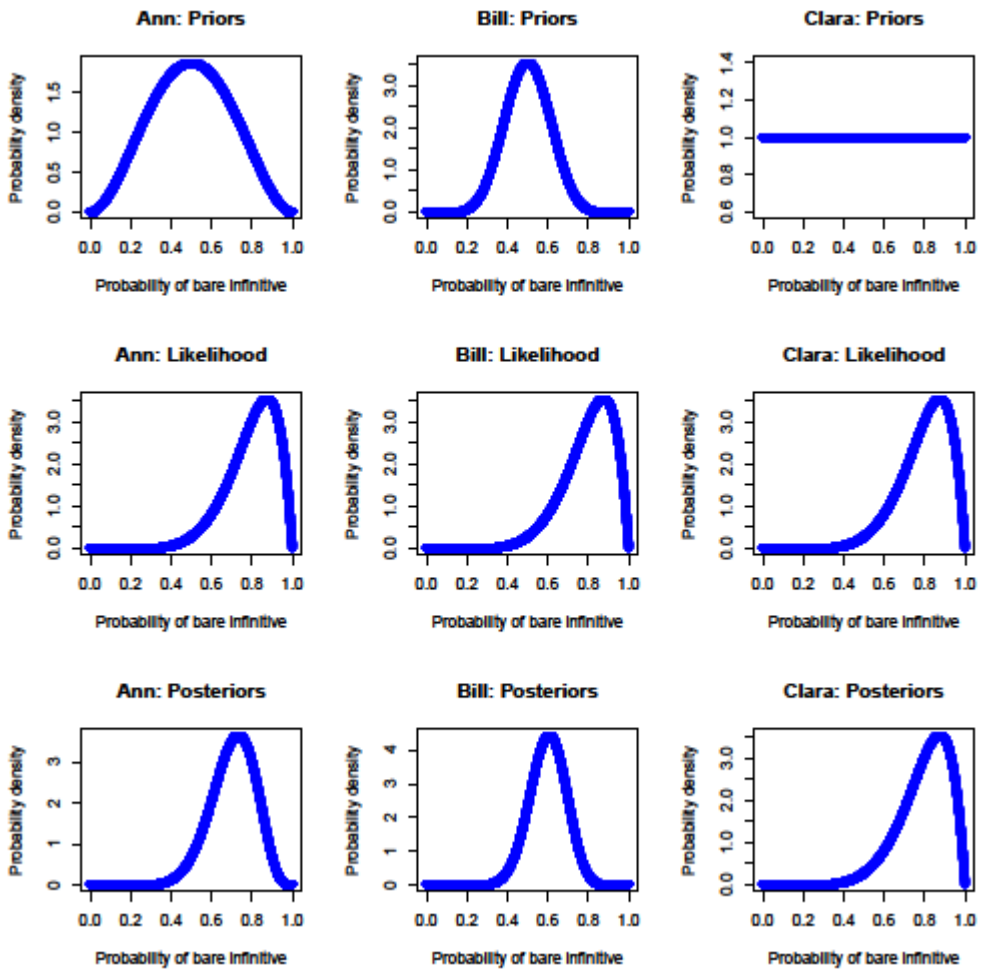

Figure 1. Different types of priors and their effects on posteriors, given the same data.

Now imagine that Ann, Bill and Clara have observed the same data with *help* in their corpus. Let us say, they have seen 8 examples with the bare infinitive and 2 examples with the *to*-infinitive. The likelihood (i.e., the probability distribution based on the data), which is displayed in the second row, is the same for all three. The posteriors, which are the beliefs after the data have been taken into account, are shown in the bottom row.

Two important facts should be mentioned. First, the stronger the priors, the greater their effect on the posteriors, given the same data. This is why Bill's posteriors are more shifted to the middle than Ann's posteriors. Flat priors have no impact on the posteriors. Therefore, Clara's posteriors are identical with her likelihood. This is why the results of Bayesian models with uninformative priors converge with those obtained with the help of frequentist methods (see Section 5).

Second, the more data we have, the less noticeable the effect of priors. If Bill observes 800 examples of bare infinitive against 200 examples of the *to*-infinitive, his posteriors would shift to the right and become very specific, overriding his prior expectations, as shown in Figure 2 (left). If Clara, who had no specific expectations whatsoever, observes the same data, her posteriors would look very similar (Figure 2, right). Importantly, Bayesian models based on very large datasets have low sensitivity to priors. As a result, when your dataset is large, the frequentist and Bayesian methods will converge, especially if your priors are weakly or moderately informative. It makes sense to perform prior sensitivity analysis in order to see whether different types of priors have an effect (see Appendix 6).

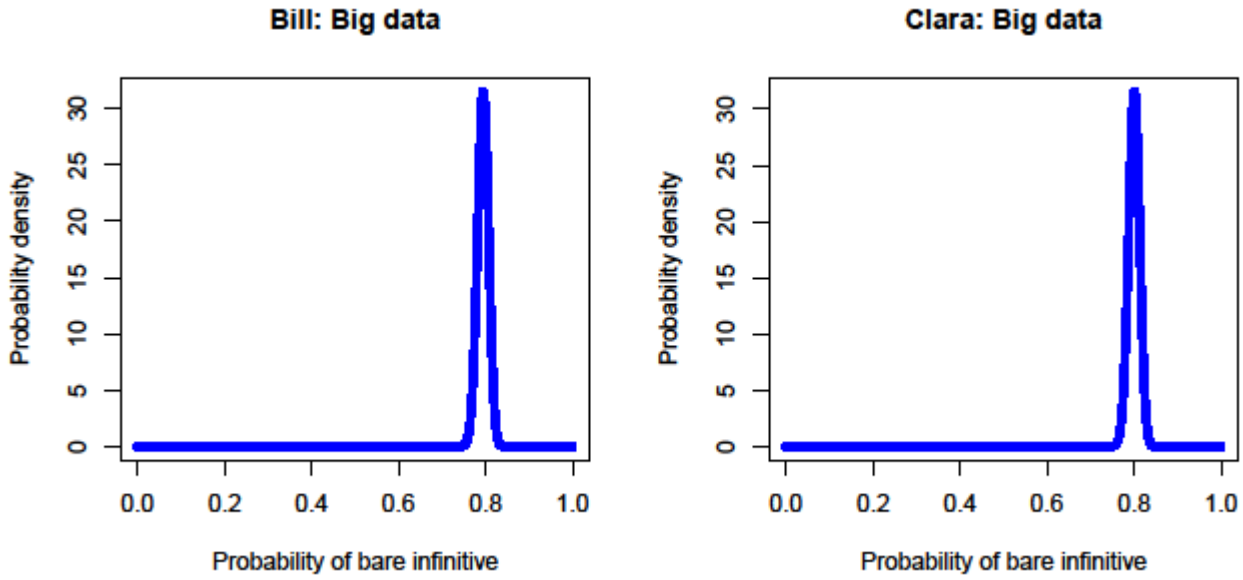

Figure 2. Effects of big data on posteriors for different types of priors.

The posteriors of one analysis can be later recycled as priors in a new study. If we use informative priors from previous analyses, we need less data. This makes our work as scientists more efficient.

There are practical benefits of using weakly informative priors, similar to Ann's. They help to exclude unrealistic values of coefficients in logistic regression, which can be useful in the situations of data sparseness. As pointed out by Gelman et al. (2008), it is reasonable to

specify that the values of predictor coefficients in a logistic regression model are very likely to be in the interval between -5 and 5, and unlikely to be outside that range.

How to combine the priors and the data and obtain the posteriors? Usually, we approximate them by sampling a large number of representative points from the posterior distribution with the help of a Markov Chain Monte Carlo (MCMC) algorithm. A Monte Carlo simulation is any simulation that draws random values from a distribution. A Markov Chain process is a random walk when the next step does not depend on the steps before the current position. In Bayesian MCMC algorithms, such as the Metropolis-Hastings algorithm, the initial value is selected randomly, then a next one is proposed randomly. Whether or not this new value will be accepted, depends probabilistically on whether or not it has a higher posterior probability than the current value, computed from its likelihood and priors (Kruschke 2011; McElreath 2016). Amazingly, an MCMC is supposed to converge to the target distribution regardless of where the initial position was. In Bayesian regression modelling, we usually create several chains, which consist of thousands of iterations. Often, we need to discard the initial steps (the so-called 'warm-up' or 'burn-in' period). This is why in some cases Bayesian regression is more costly from the computational point of view than frequentist models. However, with the development of powerful computers and smart algorithms, these issues are increasingly becoming negligible.

The posterior samples produced by an MCMC algorithm represent a distribution, from which we can compute diverse statistics: the mean, the median, the mode, the interquartile range, etc., which can tell us which values (or regions) of the regression coefficients are likely and unlikely, and how much variation exists. Consider an example, which is based on a simple binary logistic regression model of *help* + *to*- vs. bare infinitive, in Figure 3, which illustrates the effect of time on the odds of the *to*-infinitive vs. bare infinitive. Flat uninformative priors were used. The horizontal axis represents the log-odds. The further to the right, the more popular the *to*-infinitive becomes with time (per year). The more to the left, the more popular the bare infinitive becomes. There were 4,000 posterior samples based on the MCMC algorithm. The mean value of the posteriors, represented by the dot, is around -0.056. It overlaps with the median. The maximum is approximately -0.02, and the minimum is about -0.09. This means that all posterior samples are negative. Therefore, the chance that the *to*-infinitive becomes less and less probable with time is 100%. This is the probability of the research hypothesis after the data have been taken into account.

Figure 3. Log-odds of the effect of time (per year) on the chances of the *to*-infinitive (vs. the bare infinitive).

Figure 3 also displays the 95% credible interval (also known as the posterior probability interval), which should not be confused with confidence intervals used in frequentist statistics. A 95% **confidence** interval means the following. If we calculate a confidence interval in this way many times on repeated samples, the true mean (a fixed but unknown value) will be between the boundaries 95% of the time. In other words, a confidence interval reflects the reliability of the procedure, rather than the probability of finding the population parameter in a specific range. Obviously, such a complex concept is doomed to be misinterpreted. A Bayesian **credible** interval is much easier to understand. It simply means that the parameter of interest (e.g. a regression coefficient) falls in this interval with the probability of 95%. In fact, this is how many people interpret confidence intervals when they use frequentist statistics.[4]

[4] There are different kinds of credible intervals: quantile-based, or symmetric interval, and Highest Density Interval (HDI). A symmetric interval lies between the values that correspond to 2.5% and 97.5% of the distribution. A 95% HDI is the narrowest interval which contains 95% of all data points. If the distribution is symmetric, the intervals will coincide.

Importantly, Bayesian statistics does not employ *p*-values. We are not tempted to make binary decisions of the type "Significant" – "Not significant" or "Accept" – "Reject". In the age of *p*-value hacking, which is one of the causes of the reproducibility crisis, this can be seen as an advantage. Moreover, *p*-values are a notoriously difficult concept, which is easy to misinterpret (Goodman 2008).

For a long time, Bayesian models required special skills, such as programming in special languages that are tailored for specific software, such as Stan or BUGS. This was a hindrance for many linguists, who seldom have a background in programming. Nowadays, we are lucky to have the R package *brms* (Bürkner 2017), which uses a syntax very similar to the expressions used in the package *lme4*, which has become the default tool for modelling linguistic variation. This package is based on Stan, a statistical platform in C++ and programming language for Bayesian and other types of inference.

These ideas will be illustrated by a case study of *help* + bare or *to*-infinitive in American magazines. The next section summarizes previous work on this alternation.

3. Previous research on *help*

3.1. General remarks

The construction with *help* has received substantial attention in the literature. Many studies are descriptive (Rohdenburg 1996; Biber et al. 1999; Mair 2002; McEnery & Xiao 2005; Rohdenburg 2009). There are also multivariate models: Lohmann's (2011) logistic regression model based on the British National Corpus, and Levshina's (2018) models representing seven geographic varieties of web-based English. Despite the differences between the methods and the data, the results of these studies converge to a great extent. They demonstrate that there are many different factors that co-determine the use of the bare or *to*-infinitive, mostly in a subtle, probabilistic way. These factors are discussed below.

3.2. Formal factors

An important role is played by the form of the verb *help*. It has been observed that the base form *help* is the one that occurs the most frequently with the bare infinitive, while the form *helping* shows the highest proportion of *to*-infinitives (Rohdenburg 2009; Lohmann 2011; Levshina 2018). Compare the following examples from COCA (here and below the examples are taken from the Magazines section, unless specified otherwise):

(4) a. *It is his job to see through the contracts that will **help rebuild** Iraq.*

b. *Many are partners with South African companies on projects that are **helping to rebuild** the country's infrastructure.*

Another important variable is the linguistic distance between *help* and the infinitive. The more words between them, the higher the chances of the *to*-infinitive (Rohdenburg 1996; Lohmann 2011; Levshina 2018). Consider an example from COCA with a very long chunk of text between *help* and the infinitive:

(5) *HUD will provide $70 million for 1,300 rental vouchers to **help** people in public housing projects in Los Angeles, Chicago, Boston, Baltimore, and New York **to move** into surrounding middle-class and affluent suburbs.*

This tendency can be explained by the principle of minimization of cognitive complexity, which says, "[i]n the case of more or less explicit grammatical options the more explicit one(s) will tend to be favoured in cognitively more complex environments" (Rohdenburg 1996: 151).

Next, we should mention a universal tendency to avoid repetition of identical elements, or *horror aequi* (Rohdenburg 2003). If the verb *help* is preceded by *to*, the second infinitive is usually without *to* (see also Biber et al. 1999: 737; Lohmann 2011; Levshina 2018):

(6) *With yoga, find a teacher who will make adjustments to **help prevent** injuries.*

There is an interaction between this factor and linguistic distance. The more words there are between *help* and the infinitive, the weaker the influence of *horror aequi*. The sentence in (5), where *help* is preceded by *to*, but the second verb is still marked, serves as an example.

The presence or absence of the explicit Helpee plays a role, as well (Biber et al. 1999: 735; Lohmann 2011). The chances of the marked infinitive are higher if the Helpee is implicit. Compare (7a) with the explicit Helpee and (7b) without an implicit Helpee:

(7) a. […] *physical therapy has* ***helped*** *me* ***improve*** *my posture…*

b. *His encouragement and guidance* ***helped*** *[Ø]* ***to improve*** *my health and self-confidence.*

At the same time, as shown by Levshina (2018), there is an interaction between this variable and the form of *help*. In particular, the form *helping,* which has been reported to be followed predominantly by the *to*-infinitive (see above) does not differ substantially from the other forms if the Helpee is explicit.

According to McEnery & Xiao (2005), the passive infinitive should always be marked with *to*. However, I could find examples of the bare passive infinitive in COCA, as in the following sentence:

(8) *They want to* ***help*** *you* ***be saved****.*

At the same time, the passive infinitive of *help* is always accompanied by the *to*-infinitive, according to my observations:

(9) *They need to* ***be helped to reflect*** *on their shadow side -- for their own sake and the sake of others.*

Recently, it has been demonstrated that the *to*-infinitive is more frequently used (Levshina 2018; Forthcoming) if the verb in the infinitival slot is less predictable given the construction, and/or the construction is less predictable given the slot. For instance, *help* + *understand* is more likely to have the bare form than *help* + *undress, be* or *do*. The effect is not very strong, however.

3.3. Semantic factors

Dixon (1991: 199) argues that the variant with the bare infinitive is used when the Helper is more actively involved in carrying out the event. Thus, the sentence in (10a), which contains a bare infinitive after *help*, describes a cooperative effort of John and Mary, who ate the pudding together, whereas the sentence in (10b), where the *to*-infinitive is used, represents John as a facilitator for Mary's action (Dixon 1991: 199; 230).

(10) a. John helped Mary eat the pudding (he ate half).

b. John helped Mary to eat the pudding (by guiding the spoon to her mouth, since she was still an invalid). (Dixon 1991: 199)

If this is true, then the formal difference between help with the bare and *to*-infinitive is iconically motivated: the semantic integration of events is matched by the formal integration of *help* and the infinitive. At the same time, the importance of this semantic distinction has been questioned by some researchers (e.g. Huddleston & Pullum 2002: 1244; McEnery & Xiao 2005).

It has also been argued that animate Helpers have a potentially greater involvement in the event (Lind 1983). In fact, Lohmann (2011) finds that animate Helpers increase the chances of the bare infinitive in comparison with inanimate Helpers, although the effect is rather weak. This criterion, however, would not help to explain the difference between (10a) and (10b) because the Helper (John) is animate in both sentences.

3.4. Historical, geographic and stylistic variation

Finally, one should mention the factors related to geographic and stylistic variation. Although the bare variant is the more common one in Present-Day English across the world, there are some geographical differences. For example, U.S. corpus data yielded the lowest proportion of the *to*-infinitive, and the Jamaican variety the highest proportion of the seven varieties

studied by Levshina (2018). The constructions from Australia, Ghana, Great Britain, Hong Kong and India were in-between. This supports the previous observations about the differences between American and British English (e.g. Biber et al. 1999: 735).

From a historical perspective, the bare infinitive is on its way to replace the *to*-infinitive in American and British English (Mair 2002). Rohdenburg (2009: 318–319) shows that the bare infinitive was used only very rarely by authors born at the end of the 18th century. By the end of the 19th century, however, there was a significant increase. This tendency continued in American English in the 20th century, and, with some delay, also in British English.

Finally, we should mention a stylistic difference. The variant with the bare infinitive is used in less formal discourse than the variant with the *to*-infinitive (e.g. Rohdenburg 1996: 159; see also Biber et al. 1999: 736–737).

4. Data and variables

For the case studies that are described in this paper, I used a local (offline) copy of the Magazines section from the Corpus of Contemporary American English (COCA). The written data were chosen because the distribution of the variants is less skewed in favour of the bare infinitive there (i.e., there is more variability left). Below I describe the procedure of data extraction.

First, I extracted all sentences with the forms *help*, *helps*, *helped* and *helping* in upper or lower case from the text-only version of the corpus. Next, these instances were parsed syntactically with the Universal Dependencies parser using the R package *udpipe* (Strakov & Straková 2017). After that, I took all sentences in which there was a lemma *help* with the part of speech VERB, and this lemma had a dependency *xcomp*. This dependency represents infinitival complements in English. Next, I wrote a Python script to annotate the data using the parsing information, extracting automatically all subjects of *help* (the Helper), the objects of *help* (the Helpee), the distance between *help* and the infinitival complement in words, the presence or absence of *to* immediately before *help*, and the presence or absence of *to* as a

dependent element of the infinitive complement. I also extracted the information about the corpus file from which the sentence originates (using file IDs) and the year (from 1990 to 2012). There were approximately 17,700 observations with automatic annotation.

Since the full sample was too large for the purposes of this study, random sampling was performed, such that every observation came from a unique corpus file. I drew two samples, which will be used in the analyses below:

a) a large sample with approximately 2300 occurrences for the first two models. After cleaning the data manually and excluding some spurious hits, there were 2050 examples left.

b) a small sample, which resulted in 400 occurrences after the manual cleaning.

During the manual cleaning, I excluded instances of erroneous parsing, constructions with passive *help*, and the instances of a formally identical construction with the pronoun *it* as a dummy subject (cf. McEnery & Xiao 2005):

(11) *In some cases, it may* ***help to see*** *a therapist.*

In such cases, the particle *to* cannot be omitted.

Next, the datasets were checked in order to ensure that the automatic annotation was correct, and additional information was added manually. The variables are listed below.

1) Response variable (labelled as *Response* in the R code and output): the bare or *to*-infinitive.

2) Year (*Year_new*): the year when the text was published, from 1990 to 2012. In order to have an interpretable intercept value corresponding to 0, I subtracted 1990 from every number. As a result, I had numbers from 0 (1990) to 22 (2012). This was a discrete quantitative variable. The research hypothesis related to this variable was that the proportion of the *to*-infinitive slightly decreases with time.

3) The variable representing the *horror aequi* principle (*Horror*): whether there is *to* immediately before *help* or not. The values are 'Yes' and 'No'. The presence of *to* in front of *help* is expected to decrease the chances of *to* before the next infinitive, especially if the distance between *help* and the infinitive (see below) is small.

4) Log-transformed distance (*Distance_log*): the number of words between *help* and the infinitive, disregarding *to*. If the infinitive followed *help* immediately (e.g. *help [to] understand*), the distance was 1; if there was one word (e.g. *help him [to] understand*), the distance was 2, and so on. The original values were from 1 to 11 words, although most distances were small. According to the principle of cognitive complexity, I expected the chances of the *to*-infinitive increase with distance. The numbers were log-transformed after model diagnostics, which showed that a couple of outliers with the maximum values happened to have a particular configuration of features by chance, and had too much leverage on the estimates. The log-transformation makes the differences between very large values (e.g. 10 and 11 words) smaller than the differences between small values (e.g. 1 and 2), and in this way reduces the leverage. The natural logarithm was used.

5) Morphological form of *help* (*MorphForm*): *help*, *helps*, *helping* and *helped*. According to the research hypothesis, I expected the highest chances of *to* after the form *helping*, but only if there is no Helpee and the infinitive follows immediately. The base form *help* is expected to have the highest chances of being used with the bare infinitive.

6) The syntactic expression of the Helpee (*Helpee*): whether the Helpee is expressed explicitly in the sentence as a pronoun or nominal phrase or not. The values are 'Yes' and 'No'. One could expect the *to*-infinitive to be more likely when the Helpee is absent and less likely when the Helpee is present.

7) The Helper's semantic class (*Helper*): animate (humans, organizations, animals), inanimate or missing. This variable was manually coded. The category 'missing' was used when the Helper was left unexpressed, which often happens when *help* is non-finite, as in (12), or when it was impossible to identify the semantics from the context.

(12) *She even went through a rehabilitation program in an effort* ***to help*** *her* ***walk*** *again.*

I expected higher likelihood of the bare infinitive when the Helper was animate in comparison with inanimate Helpers.

8) The individual verbs (*Verb*), which fill in the infinitival slot of the construction.

In addition, I coded the smaller sample manually for the stress patterns in order to check if they influence the use of the bare or *to*-infinitive. I followed the approach proposed by Schlüter (2003), who describes the principle of rhythmic alternation, which is a prosodic tendency to avoid sequences of stressed syllables (so-called stress clash), as well as sequences of unstressed syllables (stress lapse). There are some reasons to expect the use of *to* in different constructions to depend on prosody. In particular, Wasow et al. (2015) found that the use of the bare or *to*-infinitive in the DO-BE construction (e.g. *All we want to do is (to) celebrate*) is influenced by stress. They show that *to* is preferred when both the copula and the first syllable of the infinitive after *be* were stressed. This helps to avoid stress clash.

Therefore, we can hypothesize that the use or omission of *to* after *help* may depend on the prosodic factors. Consider some constructed sentences below. The stressed syllables are in bold:

(13) a. *I **helped** (to) **fix** that issue* [Stress clash]

b. *I helped the **go**vernment (to) **fix** that issue* [Stress lapse]

c. *I **helped** him (to) **fix** that issue* [Neither clash, nor lapse]

Following the principle of rhythmic alternation, the addition of *to* would be the most likely in the contexts like (13a), where it helps to avoid the stress clash, and the least likely in the contexts like (13b), where adding it would increase the number of unstressed syllables.[5]

In order to test this hypothesis, it was necessary to code each sentence with *help* for the presence of the stress clash or stress lapse in the absence of *to*. This was very time-consuming; therefore, it has been done only for the smaller dataset. Function words, including personal pronouns, were considered unstressed. Secondary stress was ignored. The variable is called *Stress* and has the values 'Clash', 'Lapse' and 'Good' (neither stress, nor clash).

[5] Lohmann (2011) tested two phonetic variables, namely, if the infinitive begins with a vowel, and whether the first syllable of the infinitive is stressed. Neither of the variables had a significant effect on the choice between the forms of the infinitive. At the same time, his study does not test the principle of rhythmic alternation directly.

5. Models based on the large dataset

5.1. A maximum likelihood model based on the large dataset

The purpose of this section is to demonstrate the differences and similarities between the frequentist and Bayesian regression models fitted on a large dataset. We will begin with the frequentist, or a maximum likelihood (ML) model. More exactly, it is a mixed-effects Generalized Linear Model fitted with the help of the package *lme4* (Bates et al. 2015). The individual verbs are included as random intercepts. Due to the low frequencies of most verbs, the random slopes were not modelled. The response variable was whether the bare or *to*-infinitive was used. The predictors were introduced in the previous section. Two interactions were included, based on model comparisons using the likelihood ratio test, which represents a widely accepted method of model selection, as well as on previous research: the interaction between *horror aequi* and linguistic distance, and the interaction between the presence or absence of the Helpee and the form (see the reasons explained in Section 4). Other two-way interactions were tested, as well, but were not found to be useful, or produced deficient models due to insufficient data (e.g. it is difficult to model the interaction between the presence of a Helpee and linguistic distance because an explicit Helpee implies non-zero distance). The *C*-index (also known as the concordance index), a popular metric of goodness of fit for logistic models, is 0.854. This suggests that the model has explanatory power. The other goodness-of-fit statistics and the table of coefficients are provided in Appendix 1.

According to the likelihood ratio test, the Helper's semantic class did not make a significant contribution to the model. Still, it was included, in order to facilitate the comparison with the other models discussed in the present study. Appendix 2 contains the table of coefficients of the parsimonious model without the Helper. As one can see, the estimates are very similar to the larger model discussed here. Therefore, the presence of this variable does not distort the picture.

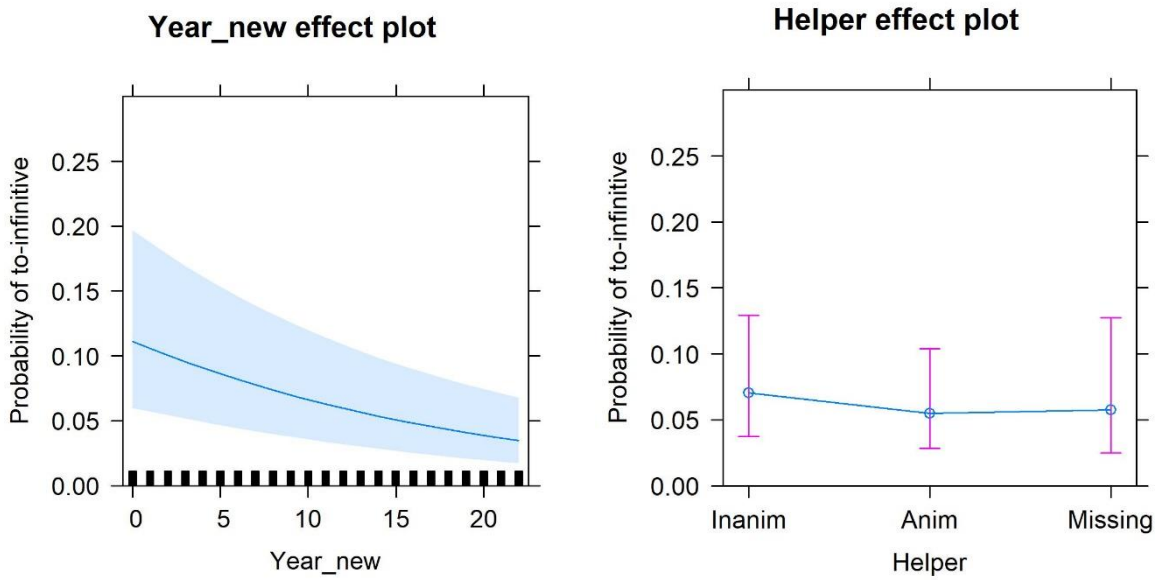

Figure 4. Partial effects of *Year_new* and *Helper*.

Let us begin with the variables that do not participate in an interaction. These are *Year_new* and *Helper*. Their effects are shown in Figure 4, which was produced with the help of the package *effects* (Fox 2003). The plot shows that the chances of the *to*-infinitive decrease with time. They are also lower for animate than for inanimate Helpers, but the difference is very small.

The other terms take part in the interactions. Figure 5 displays the interaction between the log-transformed distance *Distance_log* and the presence or absence of *to* before *help*, represented as *Horror*. The left-hand panel shows the effect of *Distance_log* in the absence of *to*. The effect is positive, but very weak. Note that the corresponding coefficient of *Distance_log* in the table is only marginally significant. The right-hand panel shows the effect of Distance in the presence of *to*. When the distance is small, the probability of the second *to* in front of the infinitive is extremely low. When the distance is large, the chances of the *to*-infinitive are at least just as great as in the absence of *to help* (note that the confidence bands are very broad due to the low number of observations with very large distances).

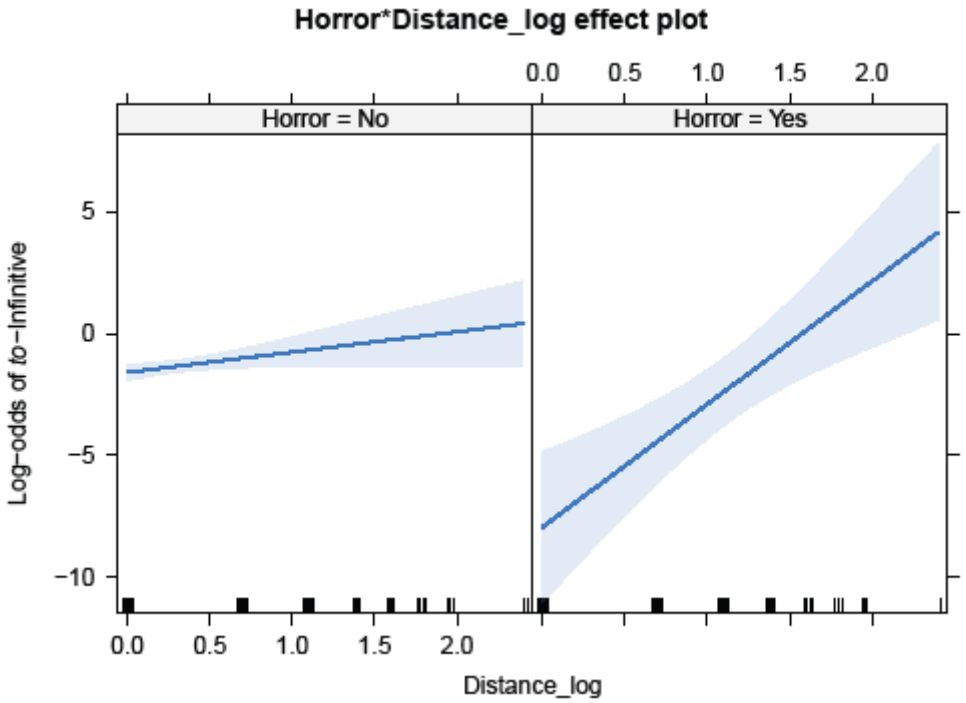

Figure 5. Interaction between log-transformed Distance and the presence or absence of *to* before *help* (*horror aequi*) in the maximum likelihood model based on the larger sample.

The second interaction in the model is between the morphological form of *help* and the presence and the absence of Helpee. This interaction is displayed in Figure 6. Overall, the Helpee indeed somewhat increases the chances of the bare infinitive, but differently for different forms. The effect is particularly striking for *helping*. It is the only form that still somewhat favours the *to*-infinitive, but only when there is no explicit Helpee. In the presence of the Helpee, the behaviour of different forms is similar.

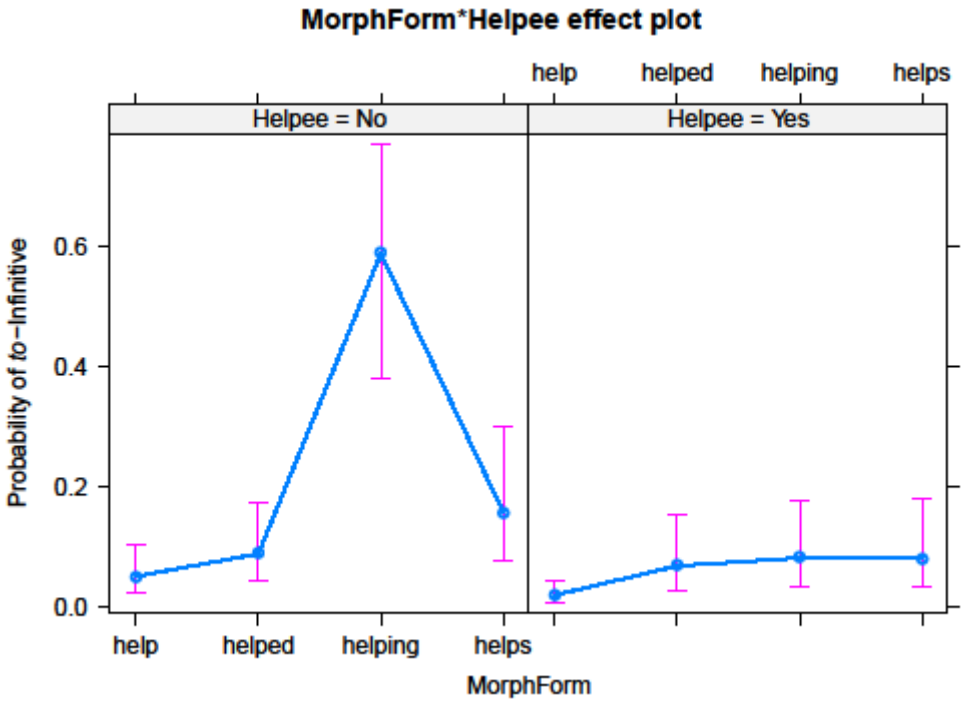

Figure 6. Interaction between the morphological form of *help* and the presence or absence of the Helpee in the maximum likelihood model based on the larger sample.

5.2. A Bayesian model with weakly informative priors based on the large dataset

When fitting a Bayesian model, the fundamental question is which priors to use. The decision depends on the type of model parameters. For residual variances and covariances, the default non-informative priors should be used. This is done because the residuals pick up omitted variables, which almost by definition are unknown (van de Schoot et al. 2014). For the standard deviation of the random intercepts, weakly informative priors are used,[6] and the number is fixed to be non-negative because variance cannot be negative by definition. As for the fixed effects, different options are available. By default, *brms* uses non-informative flat priors over the real numbers (i.e. any negative or positive numbers, including zero). A sensitivity analysis (see Appendix 6) reveals that the effects of different types of priors are

[6] More exactly, Student's *t* prior with 3 degrees of freedom and a scale parameter that depends on the standard deviation of the response after applying the link function.

very similar. When having a large sample with thousands of observations, this is mostly likely to be the case We will use Cauchy priors in the model presented below. These priors are very weak. They only 'tell' the algorithm that large negative and positive values of the coefficients are less likely than moderate ones. See more details in Appendix 6 In the next section, we will see that priors can change the picture radically if the sample is small.

There were four Markov chains with 2000 steps in each. However, 1,000 steps in each chain were used for adjusting the algorithm for efficient sampling (the so-called 'warm-up'), and were discarded. Therefore, in the end we are left with (2,000 – 1,000) x 4 = 4,000 posterior samples. When fitting a Bayesian regression model, one needs to check if Markov chains converge (see Appendix 7). The R-hat values (Appendix 5) and diagnostic plots suggest that everything is fine (see Appendix 7).

The main results of the modelling can be found in Appendix 3. Figure 7 displays the mean posteriors and 95% credible intervals. It also shows the estimates and confidence intervals of the ML model, for the sake of comparison. The estimates are very similar. One subtle difference is that the most extreme values are more moderate in the Bayesian model. They are constrained by the priors.

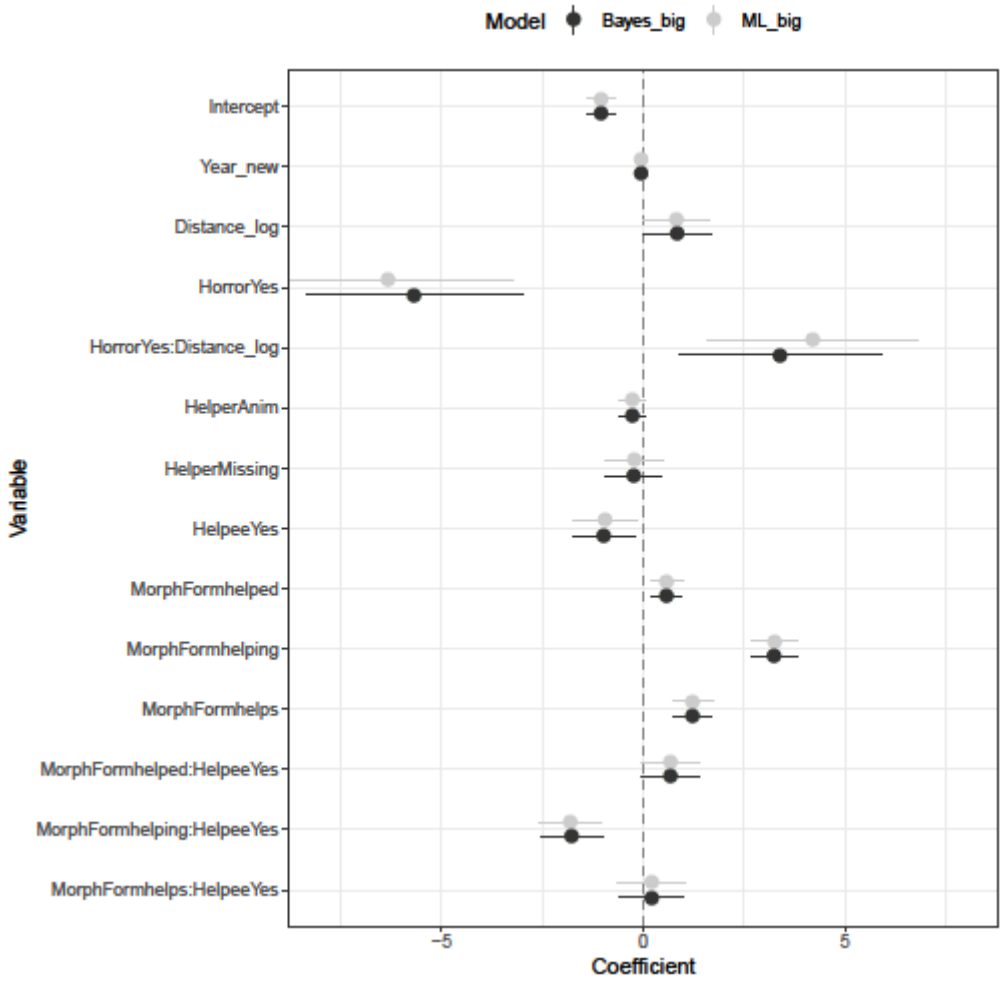

Figure 7. The coefficients of the maximum likelihood and Bayesian models.

The reader may start feeling uneasy now. If the results are similar to the ones obtained with the help of the ML approach, why bother? The reason is the Bayesian approach has some important epistemological advantages. The hallmark of the Bayesian approach is the possibility to test the research hypothesis directly. We can obtain the probabilities of the positive or negative effects of the contextual factors on the outcome. The probabilities of the regression terms having a positive effect on the odds of the *to*-infinitive are shown in Table 1.

Table 1. Posterior probabilities of positive effects in the model with the Cauchy priors

Regression parameter	**Probability of $b > 0$**
Intercept	0%
Year_new	0%
Horror = Yes (when Distance_log = 0)	0%
Distance_log (when Horror = No)	98.3%

MorphForm = helped (when Helpee = No)	99.9%
MorphForm = helping (when Helpee = No)	100%
MorphForm = helps (when Helpee = No)	100%
Helpee = Yes (when MorphForm = help)	0.7%
Helper = Animate (vs. Inanimate)	4.8%
Helper = Missing (vs. Inanimate)	26.7%
Horror= Yes : Distance_log	99.8%
MorphForm = helped:Helpee = Yes	96.8%
MorphForm = helping:Helpee = Yes	0%
MorphForm = helps:Helpee = Yes	69.7%

We see that most estimates are close to 0% or 100%, with the exception of *Helper = Missing* and the last interaction term between *MorphForm = helps*:*Helpee = Yes*. The closer the probabilities to 0 or 1, the more confident we can be about the direction of the corresponding effects. It is noteworthy that the chances of animate Helpers having a positive effect are only 4.8%. This means that their chances of having a negative effect are 95.2%, which is quite substantial. Recall that this term was not significant at the conventional level of significance ($\alpha = 0.05$) in the ML model. Traditional frequentist analysts might simply say that it was not significant and throw it away. In the Bayesian approach, we can still see that the probability of a negative effect is quite high. Probably, this effect, which was significant in Lohmann's (2011) model based on the British English data, is weaker in the American sample. This reflects the gradual nature of language change. Apparently, as the language change goes on, and the bare infinitive becomes increasingly predominant, the effects of semantic variables become weaker and weaker.

In order to evaluate how well the data are fitted by the model, we need to obtain the goodness-of-fit statistics. We can compute the predicted values based on the posterior mean estimates of the regression parameters for each of the 2,050 observations. The C-index based on such predicted values is nearly identical to the ML model: $C = 0.857$. This is not surprising. We have already seen that the posterior means and ML estimates are very similar. Moreover, the correlation between the values predicted by the ML and Bayesian models is nearly perfect ($r = 0.999$).

However, this is not a true Bayesian approach because it does not take into account the variability of the posteriors. In order to do so, we need to compute the *C*-index for each of the 4,000 random samples. For this purpose, I wrote an R script (see the *Rcode.R* file in the supplementary materials). Finally, I computed the mean and the 95% credible interval of the posterior samples. The result is very close to the one reported above: the mean *C*-index is

0.837, with the lower boundary at 0.82 and the upper boundary at 0.855. The average is slightly lower than the previous statistic.

6. Models based on the small sample

The aim of this section is to compare the ML and Bayesian approaches using the smaller sample with 400 observations. The reasons for having less data is that we are going to add one more variable, which is very time-consuming to code manually. This is a rather frequent problem in corpus-based research. We will test if stress patterns (stress clash or stress lapse) influence the use of the bare or *to*-infinitive after *help*, when controlling for the other contextual variables. For this purpose, I fitted an ML model with the same structure as the previous one, as well as a Bayesian model with informative priors based on the posteriors from the Bayesian model reported above. Normal priors were used (see Appendix 6), with the means and the standard deviations of the posterior distributions from the previous model. In this way, we recycle the results from the previous model for the new one. For *Stress*, weakly informative Cauchy priors were used, as in the previous section.

How well do the models fare? Figure 8 displays the estimates of the ML model and the posterior means of the Bayesian model with informative priors. For convenience, it also displays the estimates for the models based on the large sample (with the exception of the new variable *Stress*). 95% confidence intervals are added around the estimates of the ML models, and 95% credible intervals around the posterior means of the Bayesian models.

Note that R returns a convergence warning for the ML model. The convergence tests in *lme4* often generate false positives.[7] In other words, they find more problematic issues than there actually are. Finding out whether the problems are serious or not is usually beyond the competence of an average linguist. One of the advantages of the Bayesian approach is that the model diagnostics, which involves inspecting Markov chains, is quite straightforward (see Appendix 7).

The plot demonstrates that the small-sample ML model is very bad. Its coefficients are all over the place, and the confidence intervals are so broad that they even do not fit in the

[7] https://bbolker.github.io/mixedmodels-misc/glmmFAQ.html

boundaries of the plot between -8 and 8 log odds. Logistic regression estimates that do not fit in this interval are highly suspicious, for a dataset of this size. At the same time, the *C*-index of the small-sample ML model is 0.99. All this indicates that the model strongly overfits the data. In other words, it fits the noise, and will be useless if we take another sample. The small-sample Bayesian model, in contrast, behaves similar to the large-sample models. The coefficients of the ML and Bayesian models are provided in Appendices 4 and 5, respectively.

Now let us turn to the question whether the stress patterns play a role. The estimates in both models are very close to zero. No significant effect is found in the ML model. If there is a stress clash, the Bayesian model predicts the probability of 40.8% that the chances of the *to*-infinitive will increase. There is no evidence, therefore, that the particle *to* is inserted in order to avoid stress clash, other factors controlled for. A possible reason might be that *help* has become a true auxiliary and has lost stress itself. If there is a stress lapse, the model says that we have a probability of 63.3% that this has a positive effect on the chances of the bare form, as one could expect. This is very weak support for the importance of rhythm, however.

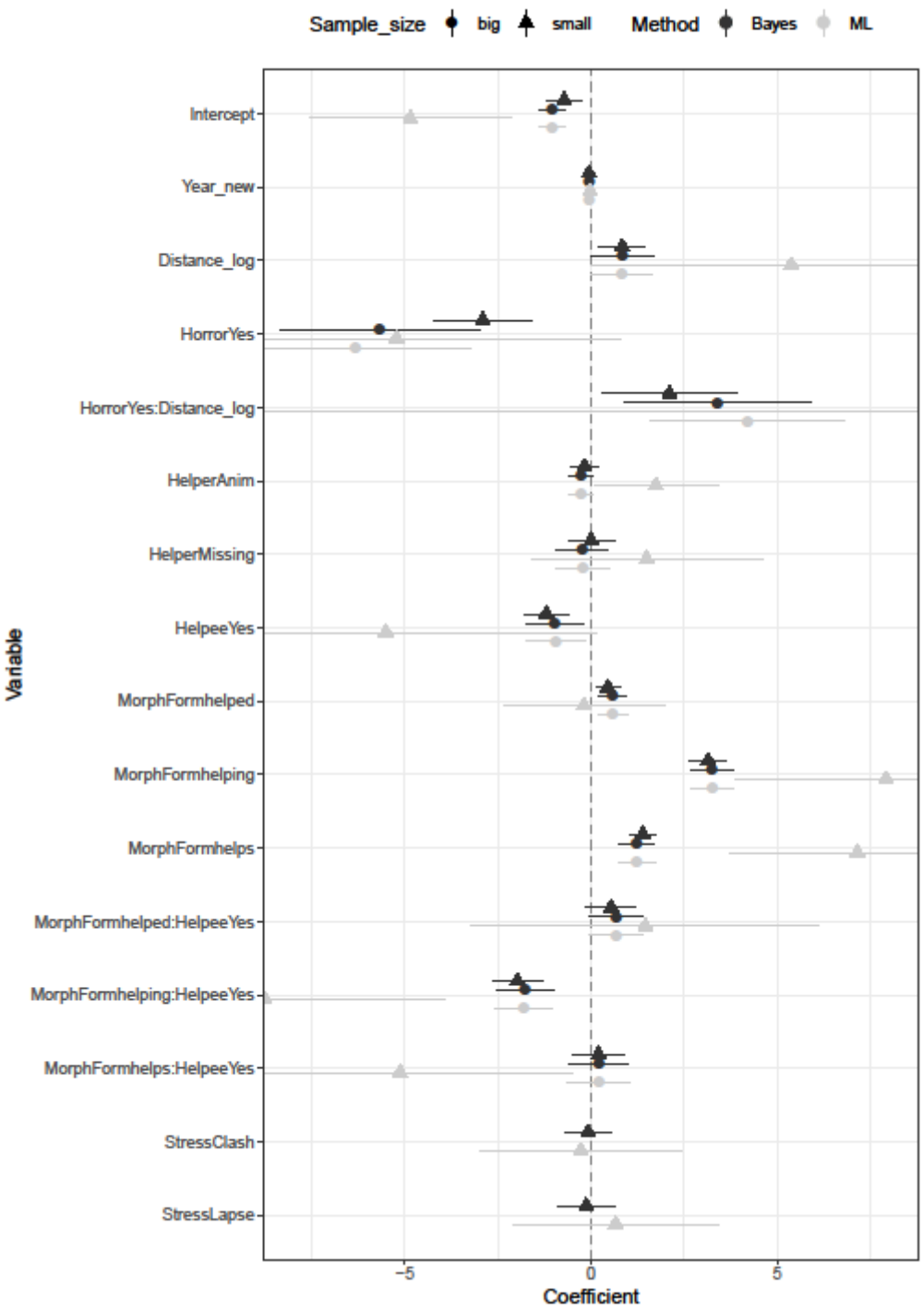

Figure 8. Coefficients of fixed effects of the large-sample maximum likelihood model, large-sample Bayesian model with weakly informative Cauchy priors, small-sample maximum likelihood model and small-sample Bayesian model with informative priors.

This example illustrates how we can build on our knowledge from previous studies and recycle the results. The informative priors also help to keep the model reasonable in case

of data sparseness. Given the hard, tedious work of manual annotation, this can help us test new hypotheses using smaller samples. This will speed up the process of accumulating knowledge and lead to more efficient use of resources.

7. Summary and reading suggestions

This chapter has compared the maximum likelihood and Bayesian approaches to regression, as well as the general principles of frequentist and Bayesian statistics. The following main ideas were discussed:

1) The Bayesian approach allows one to test the research hypothesis directly, instead of testing the null hypothesis, as in the frequentist approach. One can compute the posterior probabilities of a contextual variable having an effect on the choice of one or the other variant. This is why, potentially, Bayesian regression can be particularly good for capturing subtle differences and gradual changes in different time periods and language varieties.

2) When the dataset is large, the Bayesian model is not sensitive to the influence of priors. In this case, the statistics based on the posterior samples in Bayesian regression are very similar to the estimates in the traditional maximum likelihood approach.

3) The information from previous studies can be recycled as priors in a new Bayesian model. This can provide an advantage in cases when the data are costly. It also allows one to use the scientific resources more efficiently and increase the reproducibility of one's results.

4) In case of numerical problems (e.g. due to data sparseness), informative priors can provide a solution, ensuring that the model makes sense.

If these advantages sound attractive, the reader can explore the Bayesian methods further. More details about Bayesian logistic regression can be found in Gelman & Hill (2007) and Gelman et al. (2008). There are numerous publications on the main principles of Bayesian inference. I would recommend the following reader-friendly introductions: Kruschke (2011a), van de Schoot et al. (2014), van de Schoot & Depaoli (2014) and Nicenboim & Vasishth (2016) and McElreath (2016).

Table 2 contains a summary of the differences between the frequentist and Bayesian approaches to regression.

Table 2. Direct comparison of frequentist and Bayesian modelling

	Frequentist	Bayesian
Hypothesis testing	null hypothesis significance testing	testing the research hypothesis directly
The basis for inference	'let the data speak for themselves'	combine the data with prior expectations
Types of models	Structure is restricted by available packages	Virtually any structure (type of variables, link functions, parameters)
R script writing	Somewhat less extensive	Somewhat more extensive (specifying priors and sampling parameters)
Computational expense	usually lower	higher, especially for large datasets
Convergence issues	can be problematic (e.g. 'false alarm' messages)	usually not an issue (although diagnostics is still necessary)
Model comparison	log-likelihood test, AIC, BIC	WAIC, leave-one-out cross-validation
visualization of output	Additional packages (e.g. *effects*)	built-in visualization tools in *brms*
Verbalization	currently still accepted as default	currently still seems to require legitimization

There are many aspects that were not discussed in this introductory article due to the space limitations. For example, what to do when a Markov chain does not converge, how to use DIC (Deviance Information Criterion), LOOIC (Leave-one-out Information Criterion), WAIC (Widely Applicable Information Criterion) for variable selection, how to use Bayes factor for hypothesis testing, and how to model language variation with more than two outcomes. To find relevant information, the reader would be advised to check numerous

online publications and Internet fora dedicated to Stan-related packages, such as *brms*, *rstan* and *rstanarm*.

References

Bartoń, Kamil. 2018. MuMIn: Multi-Model Inference. R package version 1.42.1. URL: https://CRAN.R-project.org/package=MuMIn

Bates, Douglas, Martin Maechler, Ben Bolker & Steve Walker. 2015. Fitting Linear Mixed-Effects Models using lme4. *Journal of Statistical Software* 67(1): 1-48. DOI: https://doi.org/10.18637/jss.v067.i01

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman grammar of spoken and written English*. Harlow: Longman.

Bürkner, Paul-Christian. 2017. brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software* 80(1): 1-28. DOI: https://doi.org/10.18637/jss.v080.i01

Davies, Mark. (2008-) The Corpus of Contemporary American English (COCA): 560 million words, 1990-present. URL: https://corpus.byu.edu/coca/.

Dixon, R.M.W. 1991. *A new approach to English grammar, on semantic principles.* Oxford: Clarendon Press.

Fox, John. 2003. Effect Displays in R for Generalised Linear Models. *Journal of Statistical Software* 8(15): 1-27. URL: http://www.jstatsoft.org/v08/i15/

Gelman, Andrew & Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

Gelman, Andrew, Aleks Jakulin, Maria Grazia Pittau & Yu-Sung Su. 2008. A weakly informative prior distribution for logistic and other regression models. *The Annals of Applied Statistics* 2(4): 1360-1383. DOI: https://doi.org/10.1214/08-AOAS191

Gelman, Andrew, Ben Goodrich, Jonah Gabry & Aki Vehtari. 2018. R-squared for Bayesian regression models. *The American Statistician*. URL: http://www.stat.columbia.edu/~gelman/research/published/bayes_R2_v3.pdf (last access April 6 2019).

Goodman, Steven. 2008. A dirty dozen: Twelve p-value misconceptions. *Seminars in Hematology* 45(3): 135-140. DOI: https://doi.org/10.1053/j.seminhematol.2008.04.003

Goodman, Steven N., Daniele Fanelli and John P. A. Ioannidis. 2016. What does research reproducibility mean? *Science Translational Medicine* 8(341): 12. DOI: https://doi.org/10.1126/scitranslmed.aaf5027

Huddleston, Rodney & Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/9781316423530

Kruschke, John K. 2011a. *Doing Bayesian data analysis: A tutorial with R and BUGS*. Oxford: Elsevier.

Kruschke, John K. 2011b. Introduction to special section on Bayesian data analysis. *Perspectives on Psychological Science* 6(3): 272-273. DOI: https://doi.org/10.1177/1745691611406926 .

Levshina, Natalia. 2016. When variables align: A Bayesian multinomial mixed-effects model of English permissive constructions. *Cognitive Linguistics* 27(2): 235-268.

Levshina, Natalia. 2018. Probabilistic grammar and constructional predictability: Bayesian generalized additive models of *help* + (*to*) Infinitive in varieties of web-based English. *Glossa* 3(1). 55. 1-22. DOI: https://doi.org/10.5334/gjgl.294/

Levshina, Natalia. Forthcoming. Slot-filler predictability and communicative efficiency: A new approach to grammatical alternations.

Lind, Age. 1983. The variant forms of help to/help Ø. *English Studies* 64. 263–275. DOI: https://doi.org/10.1080/00138388308598255

Lohmann, Arne 2011. Help vs. help to – a multifactorial, mixed-effects account of infinitive marker omission. *English Language and Linguistics* 15(3). 499–521. DOI: https://doi.org/10.1017/S1360674311000141

Lunn, David, Christopher Jackson, Nicky Best, Andrew Thomas & David Spiegelhalter. 2013. *The BUGS book: A practical introduction to Bayesian analysis*. Boca Raton, FL: CRC Press.

Mair, Christian. 2002. Three changing patterns of verb complementation in Late Modern English: a real-time study based on matching text corpora. *English Language and Linguistics* 6(1). 105–131. DOI: https://doi.org/10.1017/S1360674302001065

McElreath, Richard. 2016. *Statistical Rethinking: A Bayesian course with examples in R and Stan*. Boca Raton, FL: CRC Press.

McEnery, Anthony & Zhonghua Xiao. 2005. HELP or HELP to: What do corpora have to say? *English Studies* 86(2). 161–187. DOI: https://doi.org/10.1080/0013838042000339880

Nakagawa Shinichi, Paul C. D. Johnson, & Holger Schielzeth. 2017. The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of The Royal Society Interface* 14(134). DOI: http://doi.org/10.1098/rsif.2017.0213

Nicenboim, Bruno & Shravan Vasishth. 2016. Statistical methods for linguistic research: Foundational Ideas - Part II. *Language and Linguistics Compass* 10: 591–613. DOI: https://doi.org/10.1111/lnc3.12207

Rohdenburg, Günther. 1996. Cognitive complexity and increased grammatical explicitness in English. *Cognitive Linguistics* 7(2). 149–182. DOI: https://doi.org/10.1515/cogl.1996.7.2.149

Rohdenburg, Günther. 2003. Horror aequi and cognitive complexity as factors determining the use of interrogative clause linkers. In Günther Rohdenburg & Britta Mondorf (eds.), *Determinants of Grammatical Variation in English*, 205–250. Berlin: Mouton de Gruyter. DOI: https://doi.org/10.1515/9783110900019.205

Rohdenburg, Günther. 2009. Grammatical divergence between British and American English in the nineteenth and early twentieth centuries. In Ingrid Tieken-Boon van Ostade & Wim van der Wurff (eds.), *Current issues in Late Modern English* (Linguistic Insights 77), 301–330. Bern: Peter Lang.

Schlüter, Julia. 2003. Phonological determinants of grammatical variation in English: Chomsky's worst possible case. In: Günter Rohdenburg & Britta Mondorf (eds.), *Determinants of Grammatical Variation in English*, 69-118. Berlin/New York: Mouton de Gruyter.

Scrivner, Olga B. 2015. *A Probabilistic Approach in Historical Linguistics. Word Order Change in Infinitival Clauses: From Latin to Old French*. PhD diss., Indiana University.

Straka, Milan, & Jana Straková. 2017. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In: *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Vancouver, Canada, August 2017.

van de Schoot, Rens & Sarah Depaoli. 2014. Bayesian analyses: where to start and what to report. *The European Health Psychologist* 16(2): 75–84.

van de Schoot, Rens, David Kaplan, Jaap J. Denissen, Jens B. Asendorpf, Franz J. Neyer & Marcel A. G. van Aken. 2014. A gentle introduction to Bayesian analysis: Applications to developmental research. *Child Development* 85: 842- 860. DOI: https://doi.org/10.1111/cdev.12169

Vasishth, Shravan, Zhong Chen, Qiang Li & Guelian Guo. 2013. Processing Chinese relative clauses: Evidence for the subject-relative advantage. *PLoS ONE* 8(10). 1–14. URL: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0077006

Wasow, Thomas, Roger Levy, Robin Melnick, Hanzhi Zhu & Tom Juzek. 2015. Processing, prosody, and optional to. In Lyn Frazier & Edward Gibson (eds.), *Explicit and Implicit Prosody in Sentence Processing*, 133–158. New York: Springer. DOI: https://doi.org/10.1007/978–3-319–12961–7_8

Appendix 1. Table of coefficients for the ML model based on the large sample, and some other details.

```
Random effects:
 Groups Name        Variance Std.Dev.
 Verb   (Intercept) 0.2147   0.4634
Number of obs: 2050, groups:  Verb, 594

Fixed effects:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               -1.03992    0.17451  -5.959 2.54e-09 ***
Year_new                  -0.05659    0.01048  -5.401 6.63e-08 ***
HorrorYes                 -6.33335    1.58414  -3.998 6.39e-05 ***
Distance_log               0.83140    0.41989   1.980  0.04770 *
```

```
MorphFormhelped             0.58174    0.19855   2.930  0.00339 **
MorphFormhelping            3.26033    0.29139  11.189  < 2e-16 ***
MorphFormhelps              1.23052    0.24966   4.929 8.27e-07 ***
HelpeeYes                  -0.94924    0.39858  -2.382  0.01724 *
HelperAnim                 -0.26949    0.16123  -1.672  0.09462 .
HelperMissing              -0.21942    0.36459  -0.602  0.54729
HorrorYes:Distance_log      4.19892    1.33709   3.140  0.00169 **
MorphFormhelped:HelpeeYes   0.67234    0.37968   1.771  0.07660 .
MorphFormhelping:HelpeeYes -1.81554    0.40431  -4.491 7.11e-06 ***
MorphFormhelps:HelpeeYes    0.20592    0.42235   0.488  0.62586
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

General properties:

AIC: 1554.6
BIC: 1639.0

The goodness-of-fit statistics are as follows. The conditional pseudo-R^2 (for the entire model) is 0.7, while the marginal equivalent, which takes into account only the fixed effects is 0.68. The statistics are computed based on the approach described in Nakagawa et al. (2017) with the help of MuMIn package (Bartoń 2018).[8] The *C*-index of concordance is 0.854. This suggests that the model has explanatory power.

Appendix 2. Table of coefficients for the ML model based on the large sample, without the Helper (parsimonious model), and some other details.

```
Random effects:
 Groups Name        Variance Std.Dev.
 Verb   (Intercept) 0.2132   0.4618
Number of obs: 2050, groups:  Verb, 594

Fixed effects:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               -1.08956    0.17217  -6.328 2.48e-10 ***
Year_new                  -0.05661    0.01046  -5.415 6.14e-08 ***
HorrorYes                 -6.46193    1.56893  -4.119 3.81e-05 ***
```

[8] More exactly, the ‘theoretical’ statistic for generalized linear mixed models is reported. There was a warning message about the convergence of the null model with the random intercepts only. A check on slightly modified samples (with several observations excluded) demonstrated that the results remain stable regardless of the presence or absence of the warning message.

```
Distance_log                 0.85528    0.41934   2.040  0.04139 *
MorphFormhelped              0.50307    0.19266   2.611  0.00902 **
MorphFormhelping             3.14587    0.28112  11.190  < 2e-16 ***
MorphFormhelps               1.25889    0.24905   5.055 4.31e-07 ***
HelpeeYes                   -0.99159    0.39752  -2.494  0.01261 *
HorrorYes:Distance_log       4.16951    1.33237   3.129  0.00175 **
MorphFormhelped:HelpeeYes    0.72105    0.37751   1.910  0.05613 .
MorphFormhelping:HelpeeYes -1.80414    0.40270  -4.480 7.46e-06 ***
MorphFormhelps:HelpeeYes     0.23383    0.42164   0.555  0.57918
```

General properties:

AIC: 1553.5
BIC: 1626.7
R^2 conditional: 0.7
R^2 marginal: 0.68
C-index: 0.853

Appendix 3. Bayesian model based on the large sample with weakly informative Cauchy priors

```
Group-Level Effects:
~Verb (Number of levels: 594)
              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk ESS Tail ESS
sd(Intercept)     0.48      0.19     0.07     0.83 1.00      508      470

Population-Level Effects:
                         Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept                   -1.05      0.17    -1.39    -0.70 1.00     4230     3397
Year_new                    -0.06      0.01    -0.08    -0.04 1.00     6178     2568
HorrorYes                   -5.68      1.36    -8.83    -3.44 1.00     2650     2159
Distance_log                 0.84      0.42     0.04     1.70 1.00     3907     2952
MorphFormhelped              0.58      0.20     0.18     0.96 1.00     4124     3018
MorphFormhelping             3.25      0.29     2.70     3.84 1.00     3232     2892
MorphFormhelps               1.22      0.24     0.74     1.69 1.00     3759     3257
HelpeeYes                   -0.98      0.39    -1.78    -0.23 1.00     3365     3143
HelperAnim                  -0.27      0.16    -0.59     0.04 1.00     6019     3503
HelperMissing               -0.24      0.36    -0.95     0.45 1.00     6764     2854
HorrorYes:Distance_log       3.40      1.29     0.99     6.02 1.00     2561     2113
MorphFormhelped:HelpeeYes    0.68      0.37    -0.04     1.43 1.00     4313     3422
MorphFormhelping:HelpeeYes  -1.77      0.39    -2.56    -1.02 1.00     4242     3101
MorphFormhelps:HelpeeYes     0.22      0.41    -0.58     1.03 1.00     4575     3169
```

Appendix 4. An ML model based on the small sample

```
Random effects:
 Groups Name        Variance Std.Dev.
 Verb   (Intercept) 31.21    5.587
Number of obs: 400, groups:  Verb, 224

Fixed effects:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                -4.85378    1.38783  -3.497 0.000470 ***
Year_new                   -0.03084    0.05155  -0.598 0.549697
HorrorYes                  -5.22132    3.08028  -1.695 0.090061 .
```

```
Distance_log                  5.37834    2.74684   1.958 0.050229 .
MorphFormhelped              -0.18552    1.10893  -0.167 0.867137
MorphFormhelping              7.92242    2.05237   3.860 0.000113 ***
MorphFormhelps                7.15671    1.74473   4.102  4.1e-05 ***
HelpeeYes                    -5.51405    2.89168  -1.907 0.056538 .
HelperAnim                    1.74596    0.84881   2.057 0.039691 *
HelperMissing                 1.49432    1.58555   0.942 0.345958
StressClash                  -0.27537    1.38856  -0.198 0.842797
StressLapse                   0.65872    1.40248   0.470 0.638583
HorrorYes:Distance_log      -21.57260 7020.50041  -0.003 0.997548
MorphFormhelped:HelpeeYes     1.46234    2.38238   0.614 0.539338
MorphFormhelping:HelpeeYes   -8.77337    2.47410  -3.546 0.000391 ***
MorphFormhelps:HelpeeYes     -5.12706    2.36549  -2.167 0.030201 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation matrix not shown by default, as p = 16 > 12.
Use print(x, correlation=TRUE)  or
    vcov(x)        if you need it

convergence code: 0
unable to evaluate scaled gradient
Model failed to converge: degenerate  Hessian with 1 negative eigenvalues

Warning messages:
1: In vcov.merMod(object, use.hessian = use.hessian) :
  variance-covariance matrix computed from finite-difference Hessian is
not positive definite or contains NA values: falling back to var-cov estimated from
 RX
2: In vcov.merMod(object, correlation = correlation, sigm = sig) :
  variance-covariance matrix computed from finite-difference Hessian is
not positive definite or contains NA values: falling back to var-cov estimated from
 RX
```

General properties:

AIC: 308.9
BIC: 376.7
R^2 conditional: 0.64
R^2 marginal: 0.97
C-index: 0.99

Appendix 5. A Bayesian model with informative priors based on the small sample

```
Group-Level Effects:
~Verb (Number of levels: 224)
             Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)     0.35      0.24     0.02     0.91 1.01      917     1454

Population-Level Effects:
                         Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept                   -0.72      0.24    -1.20    -0.25 1.00     5055     2990
Year_new                    -0.05      0.01    -0.07    -0.03 1.00     7103     2825
HorrorYes                   -2.91      0.67    -4.28    -1.64 1.00     3118     2785
Distance_log                 0.84      0.32     0.21     1.45 1.00     5630     3453
MorphFormhelped              0.45      0.17     0.12     0.78 1.00     6488     2939
MorphFormhelping             3.15      0.25     2.64     3.64 1.00     5150     2943
MorphFormhelps               1.40      0.18     1.05     1.76 1.00     5801     2298
HelpeeYes                   -1.19      0.30    -1.78    -0.61 1.00     4381     2988
HelperAnim                  -0.17      0.18    -0.52     0.17 1.00     6914     2730
HelperMissing                0.00      0.32    -0.62     0.63 1.00     5082     3015
StressClash                 -0.08      0.32    -0.69     0.56 1.00     5161     3156
StressLapse                 -0.13      0.40    -0.93     0.66 1.00     4970     2996
```

```
HorrorYes:Distance_log          2.11      0.92      0.30      3.91 1.00     4146     2926
MorphFormhelped:HelpeeYes       0.54      0.34     -0.13      1.22 1.00     5019     2618
MorphFormhelping:HelpeeYes     -1.98      0.34     -2.63     -1.32 1.00     5298     2829
MorphFormhelps:HelpeeYes        0.20      0.35     -0.50      0.88 1.00     6547     2609
```

Appendix 6. Different priors and sensitivity analysis

We can choose weakly informative priors based on well-known distributions, which are shown in Figure A1:

- Student's *t* priors (e.g. with 3 degrees of freedom, centred around 0, as shown in Figure 6);
- Cauchy priors (e.g. centred around zero and with the scaling parameter of 2.5);
- Normal priors (e.g. with the mean of zero and the standard deviation of 2).

These priors are weakly informative because they constrain the possible values only somewhat. Moreover, they do not specify the direction of an effect because they are centred around zero. One can see that the Cauchy priors have particularly fat tails, which means that they allow for more extreme values than the other priors.

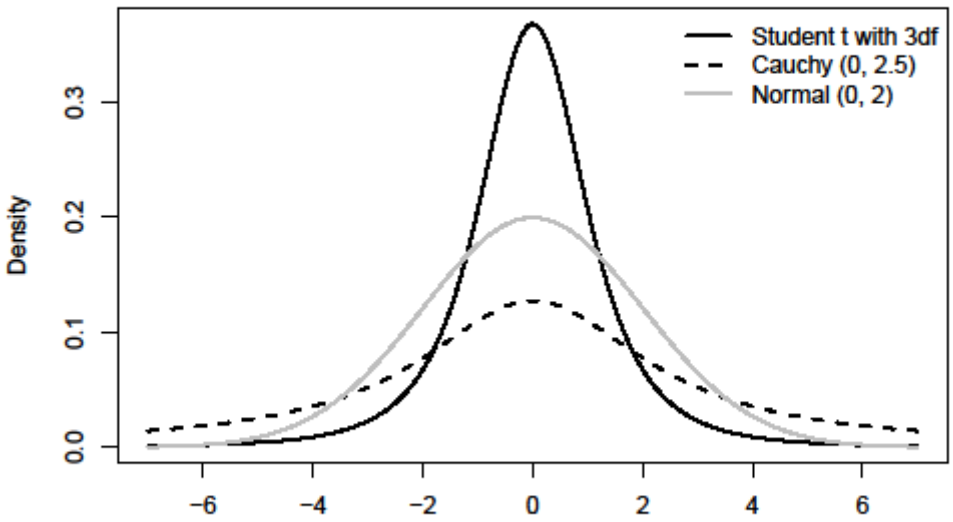

Figure A1. The weakly informative prior distributions used in this study.

Let us fit models with these priors and see if the results differ. This is called sensitivity analysis. Each of the models contained all terms and interactions that were tested in the ML model. Four Markov chains were created, with 2000 steps in each. The burn-in (or warm-up) period was 1000 first observations. Therefore, we have in total (2000 – 1000) x 4 = 4000 posterior samples in each model.

Figure A2 displays the mean posteriors of the fixed effects in the four Bayesian models, and their 95% and 90% credible intervals. For the purposes of comparison, I also added the estimates of the ML model and their confidence intervals. One can see that the estimates and the intervals are very similar. Therefore, when we have a large dataset, the models with weak or flat priors are very similar (cf. Gelman et al. 2008), and closely approximate the results of an ML model. At the same time, there are some differences when the intervals are wide, as in the case of the coefficients related to *horror aequi*. The intervals are wide due to the fact that there are few observations with *Horror=Yes* and the *to*-infinitive as the response. In that case, the default flat priors model and the ML model have the most extreme values, being the least constrained.

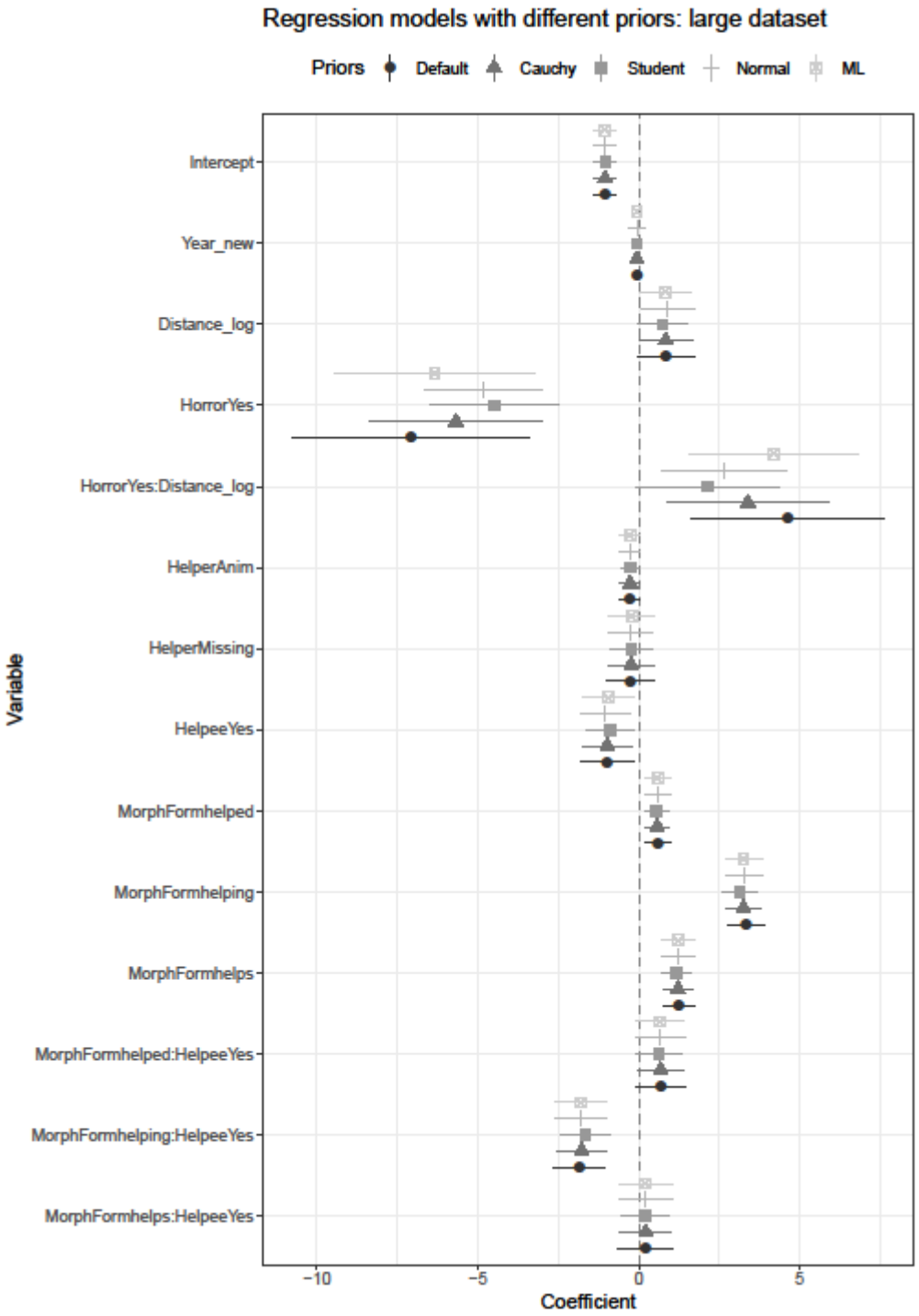

Figure A2. Effect of priors on the models: regression estimates and their 95% credible intervals. The ML model estimates and their confidence intervals are provided for comparison.

Appendix 7. Diagnostics of Markov chains

Markov chain diagnostics

When checking Markov chains, it is necessary to ensure that the chains converge to a stationary distribution around some value, and traverse the posterior distribution quickly, without getting stuck anywhere. A useful diagnostic tool is a trace plot. If a chain converges, its trace plot will look like a “fat, hairy caterpillar”, which is not bending in any direction (Lunn et al. 2013: 73–74). Figure A3 displays examples: the top plot shows convergence, and the bottom plot displays no convergence. Moreover, it is useful to create several Markov chains and to ensure that all of them converge and behave in a similar way. There is a useful diagnostic statistic, called R-hat, which should be around 1.00 (and not exceed 1.1)

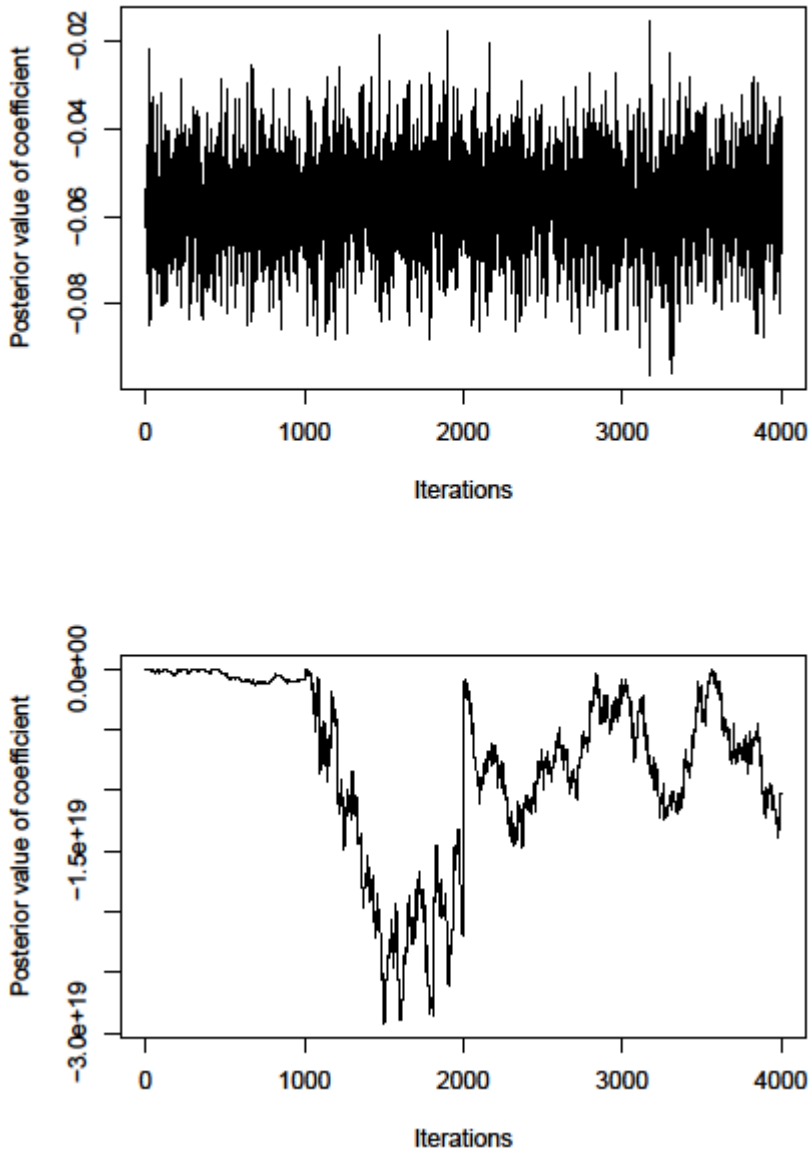

Figure A3. Examples of converging and non-converging Markov chains.

Note that the choice of the starting point can introduce a bias in the beginning of a Markov chain. This is why it is common to discard an initial fragment of this chain, for example, the first 50% of the posterior values. This part is referred to as the burn-in or warm-up period.