Glossa a journal of general linguistics

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

RESEARCH

Probabilistic grammar and constructional predictability: Bayesian generalized additive models of *help* + (*to*) Infinitive in varieties of web-based English

Natalia Levshina
Leipzig University, IPF 141199, Nikolaistraße 6-10, 04109 Leipzig, DE
natalia.levshina@uni-leipzig.de

The present study investigates the construction with *help* followed by the bare or *to*-infinitive in seven varieties of web-based English from Australia, Ghana, Great Britain, Hong Kong, India, Jamaica and the USA. In addition to various factors known from the literature, such as register, minimization of cognitive complexity and avoidance of identity (*horror aequi*), it studies the effect of predictability of the infinitive given *help* and the other way round on the language user's choice between the constructional variants. These probabilistic constraints are tested in a series of Bayesian generalized additive mixed-effects regression models. The results demonstrate that the *to*-infinitive is particularly frequent in contexts with low predictability, or, in information-theoretic terms, with high information content. This tendency is interpreted as communicatively efficient behaviour, when more predictable units of discourse get less formal marking, and less predictable ones get more formal marking. However, the strength, shape and directionality of predictability effects exhibit variation across the countries, which demonstrates the importance of the cross-lectal perspective in research on communicative efficiency and other universal functional principles.

Keywords: Bayesian regression; help; iconicity; economy; complexity; horror aequi

1 Introduction

The present paper investigates the English construction with *help* followed by the infinitive with or without *to*, as in (1):

(1) a. Mary helped John to install the program.
b. Mary helped John install the program.

The construction *help* + (*to*) Infinitive is a rare case when this choice is possible in Present-Day English. Different factors have been proposed to explain when one or the other variant is preferred. Some of them are related to the universal functional principles of iconicity, minimization of cognitive complexity and avoidance of identity (also known as *horror aequi*). Other factors include register, morphological form and the presence or absence of the Helpee. Lohmann's (2011) quantitative study of *help* in British English showed that the variation is multifactorial and probabilistic.

Moreover, it has been observed that American English has a particularly strong preference for the variant without *to*, although the bare infinitive is more common than the *to*-infinitive in both British and American varieties (e.g. Biber et al. 1999: 735). In addition, the bare infinitive has been gradually replacing the *to*-infinitive in the constructions with

help in both varieties, so that one can speak of a parallel diachronic development (Mair 2002). As shown in a corpus study by Rohdenburg (2009: 318–319), the infinitive marker *to* was dropped very rarely in British and American English with the authors born to the end of the 18th century, but there was a significant increase in the drop of the marker by the end of the 19th century. This tendency continued in American English also in the 20th century, with British speakers following the suit with some delay, which supports Mair's (2002) claim of Americanization *cum* grammaticalization of *help*.

The aims of the paper are twofold. First, I want to investigate whether the quantitative differences in the use of the bare and *to*-infinitive are also accompanied by the differences in probabilistic constraints. The paper compares the impact of the above-mentioned factors in the use of the bare and *to*-infinitive after *help* in seven varieties of online English from Australia, Ghana, Great Britain, Hong Kong, India, Jamaica and the USA, using the data from the Global Web-based English corpus, or GloWbE (Davies 2013).

The second goal is to investigate to what extent the use or omission of *to* before the infinitive can be explained by the universal tendency to use less coding material when the information is predictable, which can be seen as a manifestation of the speaker's bias towards efficient, or economical communication (e.g. Haiman 1983; Aylett & Turk 2004; Levy & Jaeger 2007; Haspelmath 2008). Here, I will focus on constructional predictability. The main hypothesis of this study is that the longer variant (i.e. with *to*) is preferred when a specific infinitive is less expected to appear as a part of the construction with *help*. This hypothesis will be tested when the other relevant factors, which are known from previous research, are controlled for in a multiple regression analysis. Importantly, the present study explores these effects in two opposite directions, testing both the predictability of the infinitive given HELP and the predictability of HELP given the infinitive, which are similar to Schmid's (2000) measures of *Attraction* and *Reliance*. As will be shown below, this approach yields some unexpected results, which open new opportunities for information-theoretic studies of morphosyntactic variation.

Methodologically, this paper employs Bayesian regression analysis, which is still novel in linguistics (but see one of the first attempts in Levshina 2016). More exactly, I use generalized additive mixed-effects regression in order to model non-linear effects of constructional predictability.

The paper is organized as follows. Section 2 discusses some of the recent studies of morphosyntactic variation, which suggest an inverse correlation between predictability and formal length. Section 3 introduces the main factors that have been discussed in previous research on the alternation. Section 4 describes the data source and the process of data extraction. In Section 5 one can find the variables that are tested in this study. Section 6 introduces the statistical method and reports the results of the quantitative analyses. Finally, a discussion of the findings is offered in Section 7.

2 Communicative efficiency and constructional predictability

Communicative efficiency can be achieved in different ways, from choosing an appropriate politeness marker, to omission of redundant information. Iconicity and minimization of cognitive complexity (see Section 2) can also be regarded as devices for maximization of communicative efficiency. In the centre of the present discussion, however, is a specific case when less predictable elements, which carry more information, get more formal coding, and more predictable elements, which carry less information, get less coding.

The inverse correlation between formal length and frequency has been known since Zipf's (1935) seminal work. However, recently it has been shown that word length is even more strongly correlated with the average predictability of words based on context than with their context-free frequency (Piantadosi et al. 2011). There is ample evidence that

more expected words, syllables or phonemes are more likely to undergo length reduction and loss of articulatory detail than less expected ones (e.g. Jurafsky et al. 2001; Aylett & Turk 2004; Bell et al. 2009; Mahowald et al. 2013).

This correlation inspired Aylett and Turk's (2004) smooth signal redundancy hypothesis, which says that information content should be spread evenly across the signal. A similar idea has also been expressed as the hypothesis of Uniform Information Density (see Levy & Jaeger 2007). These proposals involve concepts from Shannon's (1948) information theory. Information content, or surprisal, is based on the conditional probability of a unit given its context, e.g. *n* words on the right or left. It is the opposite of predictability. That is, the less predictable a unit is from its context, the more informative it is.

Of particular relevance for the present study are the studies of grammatical alternations with optional markers, which tend to be omitted when the structures that they introduce are predictable from the context, e.g. the relativizer *that* in English relative clauses after definite NPs (Wasow et al. 2011), the object marker in Japanese in typical agent-patient configurations (Kurumada & Jaeger 2015), or head-marking of the subject of the relative clause in Yucatec Maya after definite NP heads (Norcliffe & Jaeger 2016).

As far as *help* + (*to*) Infinitive is concerned, one can expect that the particle *to* will be more frequently used in the situations when the information content is higher. Information content is defined in the present study in two ways: a) based the predictability of the infinitive given HELP and b) based on the predictability of HELP given the infinitive.[1] These two measures have analogues in usage-based construction linguistics, which are known as *Attraction*, i.e. the conditional probability of a word given a construction, and *Reliance*, i.e. the conditional probability of a construction given a word (Schmid 2000). Although many corpus linguists find it useful to compute one bidirectional measure that represents the association between a construction and one of its collexemes (e.g. Stefanowitsch & Gries 2003), Schmid has been arguing that *Attraction* and *Reliance* represent two different types of information, each valuable on its own (e.g. Schmid & Küchenhoff 2013). To the best of my knowledge, these two types of predictability – predictability of a collexeme given the construction and the other way round – have not been previously taken into account in the previous studies of predictability effects in morphosyntactic alternations.

In speech, the use of additional coding material may give the speaker and the listener more time to plan and process the utterance. The predictability effects have been observed in writing, as well. As Wasow et al. (2015) hypothesize, this may happen because the speech habits are carried over to writing, or because of temporal pressures on readers. Still, the predictability effects found in writing are robust enough to test the main hypothesis of the present study on data from a written corpus.

3 Factors known from previous research

3.1 Principle of iconicity

Iconicity is the correspondence between linguistic form and function. There exist many types of iconic relationships at all levels of language structure, from phonology and orthography to morphology and syntax. For our case study, the most relevant type of iconicity is the correspondence between formal and conceptual distance. As formulated by Haiman (1983: 782), "[t]he linguistic distance between expressions corresponds to the conceptual distance between them." With regard to *help* + (*to*) Infinitive, one can say that the formal distance between *help* and the infinitive is greater when the latter is preceded by the particle *to*. In addition, iconicity of independence or autonomy may also be relevant (cf. Bybee

[1] HELP in small caps represents the construction with *help* + Infinitive as a whole, whereas *help* in italics stands for the lexeme.

1985). Events that are more integrated conceptually are also more integrated formally. In the case of *help*, it is possible to say that the bare infinitive, which is very restricted and occurs primarily as a complement to auxiliary and modal verbs and with supportive *do*, is more strongly integrated with *help* than the *to*-infinitive, which occurs in a wide range of constructions (Huddleston & Pullum 2002: 1174).

As for conceptual proximity or dependence, they are very difficult to define. In the literature, they are understood as a number of different phenomena, for example, spatio-temporal integration of the events, the degree of control and agentivity of the participants, etc. (Givón 1990: Section 13.2). With regard to *help*, it has been proposed that the variant with the bare infinitive designates a more active involvement of the Helper in carrying out the event expressed by the infinitival complement (Dixon 1991: 199). Consider the following examples:

(2) Dixon (1991: 199)
   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).

When *to* is omitted, as in (2a), the sentence is likely to describe a cooperative effort where Mary and John ate the pudding together; when *to* is included, as in (2b), the sentence means that John acted as a facilitator for Mary, who actually ate the pudding herself (Dixon 1991: 199; 230). Similarly, Duffley (1992: Section 2.3) suggests that the use of the *to*-infinitive evokes help as a condition that enables the Helpee to bring about the event denoted by the infinitive. It has also been argued that animate Helpers have a potentially greater involvement in the event (Lind 1983). Indeed, Lohmann (2011) finds that animate Helpers have higher odds of the bare infinitive than inanimate Helpers, which can be regarded as evidence in support of the iconicity account.

Yet, many researchers have questioned the relevance of this semantic distinction. For example, Huddleston & Pullum (2002: 1244) argue that there are numerous contexts and examples where this distinction cannot be traced. Similar claims were made by McEnery & Xiao (2005).

3.2 Principle of (minimization of) cognitive complexity

The principle of minimization of cognitive complexity says, "In 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). The more words between *help* and the infinitive, the more difficult it is to recognize the latter as part of the construction. Consider an example of a complex environment in (3), where the distance between *help* and the infinitive is six words.

(3) (Great Britain, blog, 3069710)[2]
   …it's a way for me to make a contribution, to **help** the country in a small way **to get back** on its feet.

The longer the distance, the more likely it is that the infinitive will be marked by the particle *to* (see also Lohmann 2011).

[2] This annotation means that the sentence is taken from the GloWbE corpus, subcorpus of blogs from Great Britain, website ID 3069710.

3.3 Principle of avoidance of identity, or horror aequi

Horror aequi is a widespread tendency to avoid repetition of identical elements (Rohdenburg 2003). This idea is also known as the Obligatory Contour Principle, which has been first formulated for phonology (Leben 1973), but has been used to explain different phenomena at all linguistic levels since then (e.g. omission of optional *that* in Walter & Jaeger 2008). Rohdenburg uses *horror aequi* to explain why the *to*-infinitive tends to be avoided immediately after a governing *to*-infinitive (e. g. *to try to do*). When the verb *help* is itself preceded by *to*, the following infinitive is usually without *to* (Biber et al. 1999: 737). See an example in (4):

(4) (Great Britain, general, 303502)
Sorry, but how is this supposed **to help** answer the question?

This hypothesis was confirmed by Lohmann (2011), who also finds an interaction between this factor and complexity (see Section 2). The more words there are between *help* and the infinitive, the weaker the influence of *horror aequi*.

3.4 Other factors

- Register: The shorter variant with the bare infinitive is considered to be less formal than the one with the marked infinitive (e.g. Rohdenburg 1996: 159; see also Biber et al. 1999: 736–737).
- Inflectional form: Lohmann (2011) observes that the form *helping* tends to be more frequently used with the *to*-infinitive in British English than the other inflectional forms of *help*. According to Rohdenburg (2009: 317), the effect of *helping* has an analogy with *daring* and *needing*, which differ from all forms of *dare* and *need* by being virtually always associated with marked infinitives. In addition to that, there is a weakly significant preference of the third person singular form *helps* for the *to*-infinitive in comparison with the base form (Lohmann 2011).
- Presence or absence of the Helpee: Biber et al. (1999: 735) show that the bare infinitive is particularly dominant in the pattern *help* + NP + infinitive clause. This observation is also supported by Lohmann (2011).
- Passive or active infinitive: According to McEnery & Xiao (2005), the passive infinitive should always be marked with *to*. However, this is not supported by my data. Both the bare and *to*-forms can be used, as shown in (5).

(5) a. (USA, general, 288902) If rural voices are important – the bread basket, our farmers, our miners – then an electoral approach, not a pure popular vote, **helps** them **to be heard**.
b. (USA, blog, 3177307)
Thank you so much for sharing and **helping** our Vets **be heard**!

One should also mention phonological factors. There is some evidence that the use of *to* in different constructions depends on prosody. Wasow et al. (2015), in particular, found an effect of prosody on the use of the bare or *to*-infinitive in their investigation of the DO-BE construction, e.g. *All we want to do is (to) celebrate*. Namely, they discovered that *to* was used to eliminate stress clash when both the copula and the first syllable of the infinitive after *be* were stressed. I'm not aware of any studies of *help* that focused directly on the effect of stress clash. However, Lohmann (2011) tested two other phonetic variables, namely, if the infinitive begins with the 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.

4 Corpus and the procedure of data extraction

The data used in the present study come from the Corpus of Global Web-based English (GloWbE) created by Davies (2013). This large corpus contains 1.9 billion words and represents online English from twenty countries. For this case study, seven geographic varieties were chosen from different parts of the world: Australia, Ghana, Great Britain, Hong Kong, India, Jamaica and the USA. The choice for this corpus was motivated primarily by its size. One needs large corpora in order to compute reliable information-theoretic measures, especially if the construction of interest is not very frequent. I used a part of the corpus with eighteen million words per country, nine million from the General subcorpus and nine million from the Blog subcorpus.

The data extraction procedure was as follows. First, I used a Python script to collect all instances of *help* in any inflectional form followed by an infinitive somewhere in the sentence. If there were finite verb forms, clause-combining conjunctions like *because*, or subject pronouns like *I*, *he* and *she* between *help* and the infinitive, the instance was discarded. A quality check based on one hundred manually extracted examples from five subcorpora revealed that this approach was quite successful in recognizing the instances of the construction: The recall was 86%, and the precision was 93%. Only active uses were collected because the bare infinitive can be used only in active sentences (Huddleston & Pullum 2002: 1244), as shown in (6):

(6) a. John was helped to cook the dinner.
b. ??John was helped cook the dinner.

The spelling variants of the verbs were normalised, so that the pairs like *maximize* and *maximise*, *fulfil* and *fulfill* were treated as one word.

In spite of the fact that the corpus compilers performed some cleaning, there were still quite a few duplicate sentences in the data. They were removed with the help of a script. Another problem were nonsense sentences, which were probably machine-generated or contained advertising information (cf. similar problems reported in Mair 2015: 31–32). However, they were not numerous and were removed during the process of variable coding.

Finally, I cleaned the data manually from the instances of a formally similar but functionally different construction with the dummy *it*-subject, where the *to*-infinitive is always used (McEnery & Xiao 2005). An example is shown in (7).

(7) Ruth Bader Ginsburg's Relationship Advice: "It helps to be a little deaf".[3]

After the data collection and cleaning, I obtained the frequencies shown in Table 1. Since the sizes of the subcorpora were identical (18 million words), the "raw" frequencies are directly comparable between the varieties. One can see that Hong Kong has the highest total frequency of the constructions, and Jamaica the lowest. However, the differences are not very large. As for the relative frequencies of the variants, the variant with the bare infinitive is the more frequent one in all countries. The USA subcorpus displays the highest relative frequency of *help* followed by the bare infinitive (84.9%), whereas the Jamaican subcorpus has the lowest one (60.8%), followed by Great Britain (70.3%) and the other countries.

The next section describes the predictor variables, which represent the factors mentioned in Sections 2 and 3. The Helper's animacy is not taken into account because it was

[3] https://www.huffingtonpost.com/entry/ruth-bader-ginsburgs-relationship-advice-it-helps-to-be-a-little-deaf_us_58a5db8ae4b045cd34bf740b (last access 25.12.2017).

Table 1: Absolute and relative frequencies of *help* + (*to*) Infinitive in seven countries.

Country	*help* + bare Inf (%)	*help* + *to* Inf (%)	Total
Australia	4556 (76.5%)	1398 (23.5%)	5954 (100%)
Ghana	4976 (71.8%)	1957 (28.2%)	6933 (100%)
Great Britain	4151 (70.3%)	1750 (29.7%)	5901 (100%)
Hong Kong	5522 (72.8%)	2058 (27.2%)	7580 (100%)
India	5228 (72.9%)	1941 (27.1%)	7169 (100%)
Jamaica	3498 (60.8%)	2256 (39.2%)	5754 (100%)
USA	5134 (84.9%)	910 (15.1%)	6044 (100%)

very difficult to automate the annotation procedure. The parser returned very poor results due to highly complex syntactic structures, e.g. when *help* was itself part of an infinitival clause. Note that the effect of animacy in Lohmann's (2011) study was rather weak. Prosodic factors (in particular, stress clash) are not tested, either, due to the practical difficulties in obtaining the stress patterns from the written data of such a large size. I added one new variable, the valency of the infinitive.

5 Predictors for regression analysis

5.1 Constructional predictability

To test the effects of constructional predictability, I computed two measures for each unique infinitive, which are described below.

- Information content of the infinitive given the construction, defined as the negative log-transformed conditional probability of the infinitive (with or without *to*) given the construction with *help*: –log *P*(verb |HELP). This conditional probability is computed as the number of occurrences of a given infinitive with HELP divided by the total frequency of the construction with *help* in the relevant subcorpus. In corpus-based constructional studies this probability is known as *Attraction* (Schmid 2000). The more frequently a verb is used in the construction with *help* in comparison with the other verbs, the lower the information content.[4]
- Information content of the construction given the infinitive, defined as the negative log-transformed conditional probability of the construction with *help* (with or without *to*) given the infinitive: –log *P*(HELP|verb). This conditional probability, which is also known as *Reliance* (Schmid 2000), is computed by dividing the number of occurrences of a given infinitive with HELP by the total frequency of the verb in the subcorpus in all forms. The more frequently a verb is used with HELP in comparison with the other uses of the same verb, the lower the information content.

5.2 Cognitive complexity

This principle is represented by linguistic distance, which was measured as the number of words between the wordform of *help* and the infinitive (the particle *to* was not counted). For example, the sentence in (8) has the distance of four words.

[4] Since the verb help can also be used without the infinitival complement, one might also be useful to take into account the surprisal of the complement after help. However, this quantity will be constant in each variety, so the simpler approach is preferred.

(8) (Hong Kong, blog, 3581048)
I worked at Airbus before going into private equity in 2001, **helping** a European family office **to diversify** their investment portfolio.

Although there are different ways of defining syntactic complexity, such as counting the number of syntactic nodes and quantifying the level of embeddedness, word counts serve as a good proxy for the more sophisticated measures (Szmrecsanyi 2004). This is why I also use simple word counts in this study.

5.3 Horror aequi

This factor is represented by the variable which reflects the presence of the particle *to* before *help*, as in (9):

(9) (India, blog, 3388613)
The Plate-Inversion protocol, and this post are two simple hacks **to help** you get started.

This is a binary variable with the values "Yes" and "No".

5.4 Other variables

- **Formality**, which is represented by the average word length in the website text where a given instance of *help* was attested. The greater the average word length, the more formal the text. This operationalization is based on Biber's (1988) multidimensional analysis of register variation. He found, in particular, that longer word forms, alongside the type-token ratio and the relative frequency of nouns and adjectives, contribute strongly to the negative pole of the first factor or dimension, which is interpreted as "Involved vs. informational production" and has conversations and academic texts at its extremes. The use of the mean word length is purely practical. Many texts in the corpus are very short and cannot provide reliable relative frequencies for the lexico-grammatical categories required for a full-fledged multidimensional analysis.
- **Morphological form** of the verb *help*, which can be *help*, *helps*, *helped* and *helping*.
- **The presence or absence of the Helpee**, illustrated by (10a) and (10b), respectively:

(10) a. (Great Britain, blog, 3058500)
It provides a systematic approach to helping **people** defeat dyslexia and related reading problems. [Presence]
b. (Ghana, general, 1259905)
These bumps and turns will only help contribute towards a relationship. [Absence]

According to the previous studies (see Section 3.4), the contexts with zero Helpees are expected to contain the *to*-infinitive more often than those with overt Helpees.

- **Valency of the infinitive**, which can be intransitive (including copulas), transitive (including ditransitives) or followed by a clause. Examples are shown in (11).

(11) a. (Ghana, blog, 3621705)
May God help nations **to live** together in peace. [Intransitive]
b. (Great Britain, blog, 3027910)
Grow your business by helping your clients **grow** theirs. [Transitive]

c. (Great Britain, general, 416004)
Everyone has something to offer and it's about helping people **believe** they play an integral part in the workplace. [Clause]

In order to code this variable, the sentences were first parsed syntactically with the help of Stanford Parser (Klein & Manning 2003). The contexts were then manually checked, and the category "Clause" was added manually. Examples with passive infinitives were excluded. Due to their extremely low frequencies, it was impossible to include them as a separate category in the regression models. At the same time, it did not seem reasonable to merge them with any other category, since previous studied suggested that they might behave differently from active forms (McEnery & Xiao 2005; see also Section 3.4).

6 Bayesian generalized additive mixed-effect models: Characteristics and results

6.1 Bayesian inference and characteristics of the models

To test the effect of the predictors on the use of bare and *to*-infinitives, I used Bayesian mixed-effects generalized additive models. For this purpose, I employed Stan, a programming language and platform for Bayesian inference (Stan Development Team 2015) and the package *brms* (Bürkner 2017), which provides an R interface to Stan (R Core Team 2017).

Seven Bayesian logistic regression models were fitted, one for each variety. The response variable was the use of the bare or *to*-infinitive. The predictors described in Section 5 were treated as fixed effects. The individual websites and the verbs that fill in the infinitive slot were treated as random effects (more exactly, random intercepts). Sum contrasts were used with all categorical and binary variables, so that zero represents the grand mean (i.e. the unweighted mean of means) of the categories. The numeric variables were centred around the mean. Two interaction terms were modelled after diagnostic tests. One interaction is between linguistic distance and the *horror aequi* variable, which was found to be significant by Lohmann (2011). The other is the interaction between the form of *help* and the presence or absence of the Helpee. In addition, an interaction between the two information-theoretic measures was taken into account by introducing bivariate smoothing terms (see below). The discriminating power of the models was excellent (all concordance indices *C* were greater than 0.9).

Bayesian regression allows the researcher to test directly the research hypothesis. In our case, we can obtain the probability of a predictor having a positive or negative effect on the presence or absence of the particle *to*. In Bayesian inference, such probabilities are called posterior probabilities, or posteriors, because they are computed after the data have been taken into account. They also depend on prior probabilities, or priors, which represent the researcher's prior beliefs in the probability of some parameters before the data are taken into account. If one provides non-informative priors (e.g. uniform ones, where any value is equally probable), this will result in posteriors that are influenced only by the data, as in frequentist statistics. As recommended by the Stan developers, I used the default weakly informative priors, which only help to constrain the posteriors to reasonable values, i.e. those to be normally found in logistic models. Bayesian regression is a perfect match for probabilistic grammar because posterior probabilities can be easily compared cross-lectally. They also allow us to study a continuum of credibility without forcing us to make binary decisions based on *p*-values. For more information about the technical details of Bayesian modelling, one can be referred to Kruschke (2011). In what follows, I focus on the results.

The algorithm returns 6000 posterior estimates of each regression parameter (1500 estimates in four Markov chains per each model). These probability distributions can be represented in a histogram which displays our posterior beliefs after the data have been

taken into account. An example is provided in Figure 1. It shows the effect of average word length on the chances of the bare and *to*-infinitive in the websites from Great Britain. The numeric values on the horizontal axis are the log-odds ratios. A positive log-odds ratio means that the odds of the *to*-infinitive increase with average word length, whereas a negative value means that the odds of the *to*-infinitive decrease (and, conversely, the odds of the bare infinitive increase). From the posterior distribution one can compute the posterior mean, which is displayed as a dot in Figure 1, as well as 95% credible intervals, which show the region between the 2.5% and the 97.5% percentiles, where the 95% of the posterior distribution lies. Credible intervals thus span the most believable posteriors. If one has to make a categorical judgment of the type "Does the variable increase the chances of one or the other outcome?", one can use this criterion. If a credible interval does not include zero, as in this illustration, one can say that the effect is credibly nonzero.

The posterior distribution can also help us assess the probability of observing the positive and negative effect of a given predictor on the chances of the *to*-infinitive by computing the proportions of the posteriors that are greater and less than zero. In our example, the proportion of the posteriors greater than zero is 100%. This information allows us to test directly the alternative hypothesis.

Additional diagnostic tests with polynomials suggested that some of the effects of the predictability variables are non-linear. To take that into account, I used the methods of generalized additive modelling (Wood 2006), which applies smooth functions to model non-linear relations between predictors and the response. More exactly, I used bivariate smoothing terms, which take into account possible non-additive effects of two predictors. Using the LOO criterion for model comparison, I chose isotropic smooths, which are appropriate when variables are on similar scales. As for the other continuous variables, no convincing non-linearity was detected.

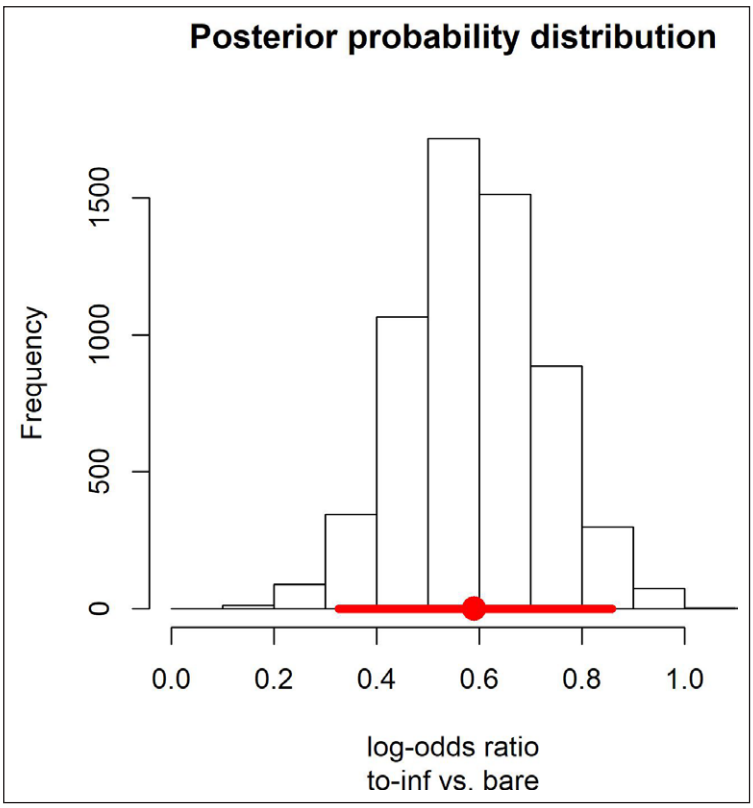

Figure 1: Posterior probability distribution of the effect of average word length in a text on the presence or absence of *to* in the subcorpus representing Great Britain.

6.2 Results of Bayesian modelling

6.2.1 Predictability-related variables

The marginal effects of the information content of a verb given HELP are displayed in Figure 2. They are based on the predicted probabilities of the *to*-infinitive. Recall that the hypothesis was as follows: The greater the information content, the higher the chances

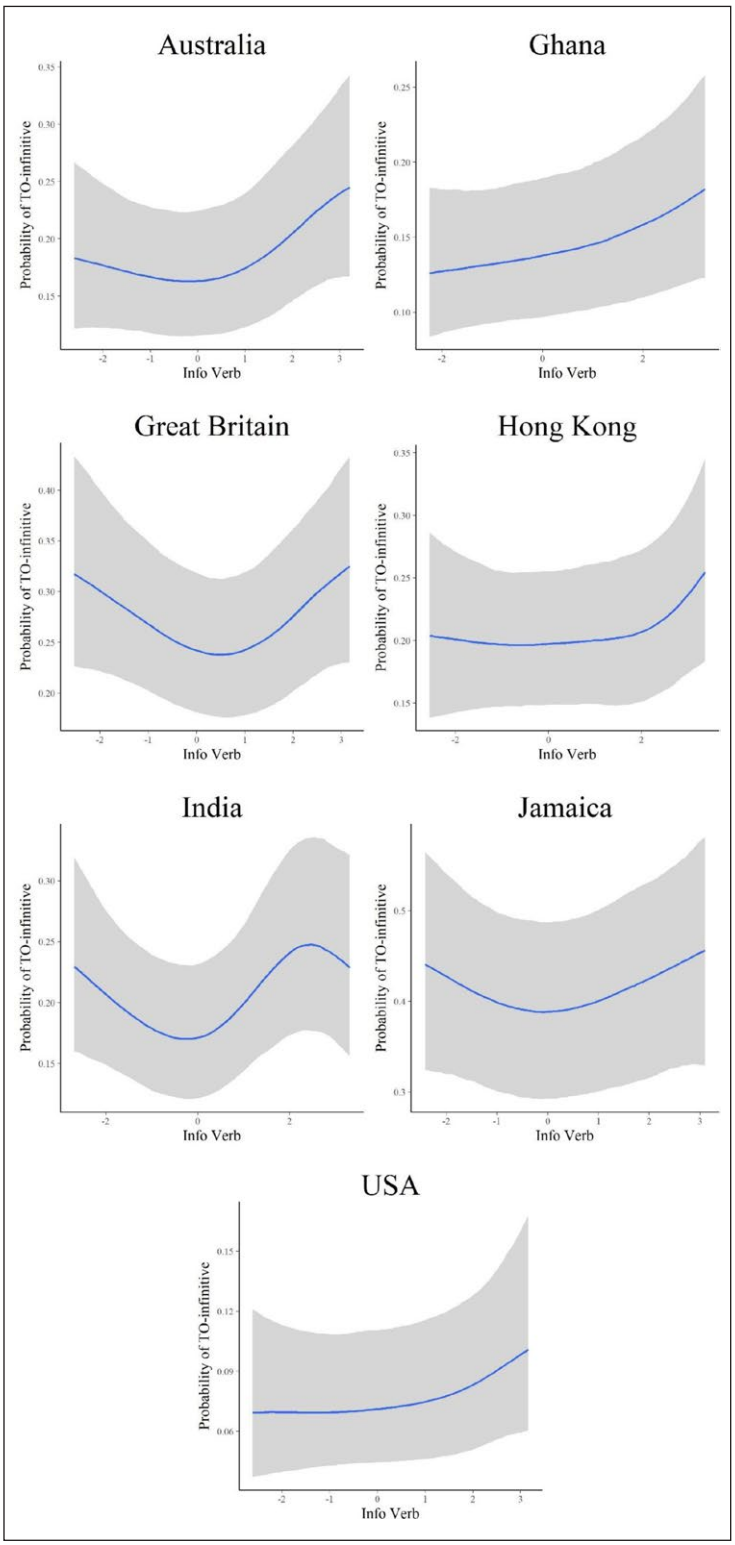

Figure 2: Marginal effects of information content of a verb given the construction with *help* (the horizontal axis) on the probability of the *to*-infinitive (the vertical axis).

of the marked form. Although some of the plots point in the right direction (e.g. the data from Ghana, Hong Kong and the USA), the 95% credible bands are very broad in comparison with the magnitude of those effects, which means that the latter are marginal at best.

In contrast, the marginal effects of the information content of the construction given a verb are more robust, as shown in Figure 3. In the USA data, the effect is the weakest. In

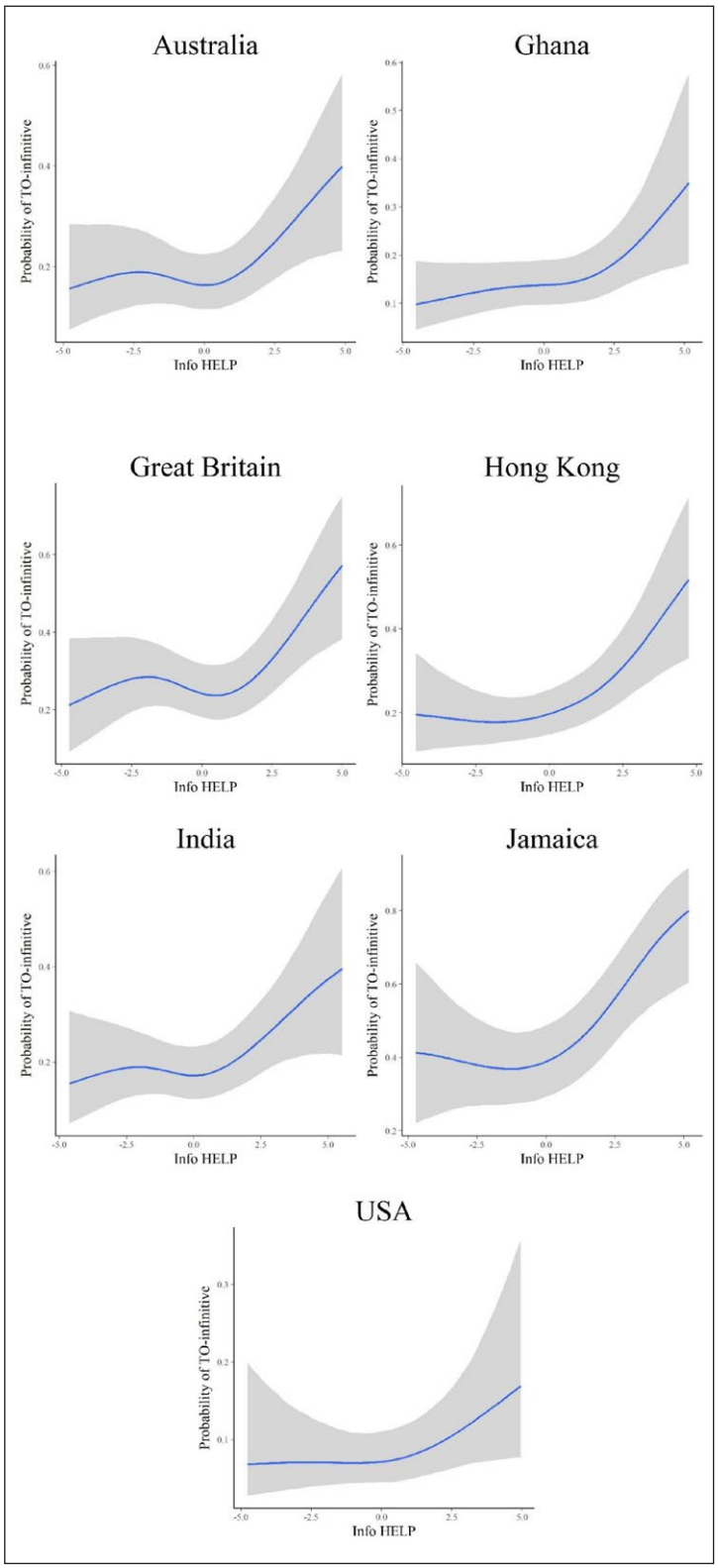

Figure 3: Marginal effects of information content of the construction with *help* given a verb (the horizontal axis) on the probability of the *to*-infinitive (the vertical axis).

Australia, Great Britain, India and Jamaica, we also observe some non-monotonicity, with a small dip in the centre.

The effects of both information-theoretic variables in interaction are displayed in Figure 4. The lighter areas (from violet to blue and then to green and yellow) indicate

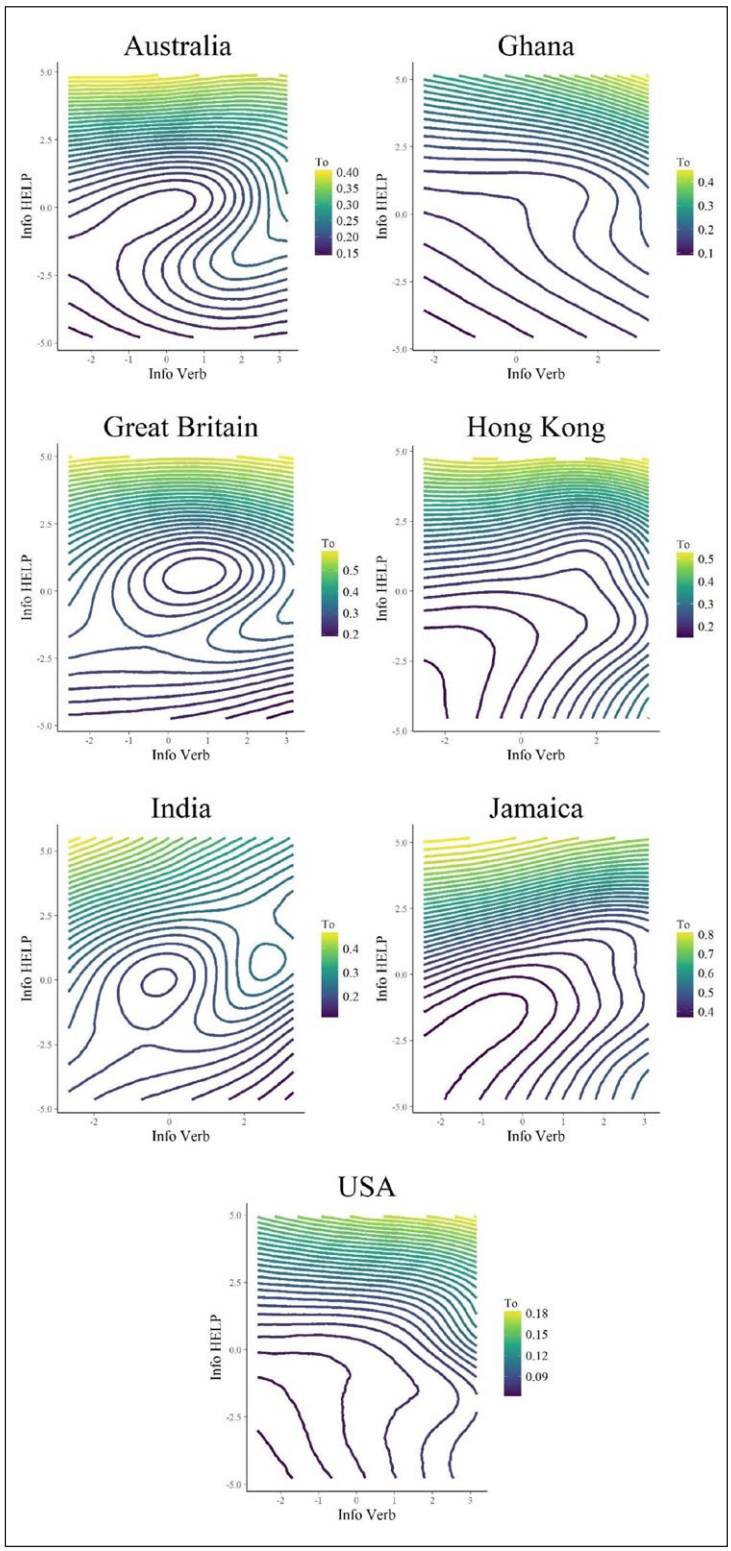

Figure 4: Non-linear effects of information-theoretic measures in seven varieties. Horizontal axis: information content of a verb given the construction; vertical axis: Information content of the construction given a verb. Lighter shades (yellow): Higher predicted probability of the *to*-infinitive. Darker shades (violet): Lower predicted probability of the *to*-infinitive.

the information content values where the chances of the *to*-infinitive increase, while the darker areas show the values with a higher preference for the bare infinitive. When the information content of HELP given a verb is very high (see the top part of the plots), the chances of the *to*-infinitive tend to increase. There is also a slight increase in the bottom right part of the plots in some of the varieties. This is a region with with high information content of a verb given HELP (the horizontal axis) and low to middle information content of HELP given a verb (the vertical axis). This increase explains the non-linear patterns discovered in Figure 3.

6.2.2 Cognitive complexity, horror aequi and their interaction

The effects of cognitive complexity and *horror acqui* are as expected in all countries. Table 2 displays the effects of linguistic distance. With each word between *help* and the infinitive, the odds of the *to*-infinitive credibly increase. There is some variation in the strength of this effect, with the American variety displaying the smallest value, and the Indian one the largest.

Table 3 shows the effects of the presence of *to* before *help* for mean linguistic distance. The chances of the *to*-infinitive decrease if there is *to* before *help*. There is some variation, again. The Hong Kong data display the weakest effect, and the Jamaican subcorpus shows the strongest effect.

The positive interaction terms (see Table 4) indicate that the odds of the *to*-infinitive become higher, as the linguistic distance between *help* and the infinitive increases. The US data display the weakest effect, while the Jamaican variety has the strongest effect, closely followed by several others.

6.2.3 The form of help, the presence or absence of the Helpee and their interaction

The results are best represented visually. Figure 5 displays the mean posteriors and the 95% credible intervals. In all varieties, the form *helping* without the Helpee has the highest

Table 2: Bayesian regression results for linguistic distance (log-odds ratios).

Country	Posterior mean	2.5%	97.5%	$P(\beta > 0)$
Australia	0.46	0.3	0.63	100%
Ghana	0.7	0.57	0.84	100%
Great Britain	0.56	0.41	0.71	100%
Hong Kong	0.69	0.56	0.83	100%
India	0.82	0.66	0.99	100%
Jamaica	0.48	0.3	0.65	100%
USA	0.38	0.22	0.54	100%

Table 3: Bayesian regression results for the presence of *to* before *help* (log-odds ratios).

Country	Posterior mean	2.5%	97.5%	$P(\beta > 0)$
Australia	–1.33	–1.51	–1.15	0%
Ghana	–1.29	–1.46	–1.14	0%
Great Britain	–1.24	–1.4	–1.09	0%
Hong Kong	–1.13	–1.27	–0.98	0%
India	–1.35	–1.54	–1.17	0%
Jamaica	–1.62	–1.82	–1.44	0%
USA	–1.3	–1.52	–1.09	0%

Table 4: Bayesian regression results for the interaction term between *linguistic distance* and the presence of *to* before *help* (log-odds ratios).

Country	Posterior mean	2.5%	97.5%	$P(\beta > 0)$
Australia	0.28	0.14	0.41	99.98%
Ghana	0.28	0.17	0.38	100%
Great Britain	0.37	0.24	0.5	100%
Hong Kong	0.38	0.27	0.49	100%
India	0.39	0.24	0.54	100%
Jamaica	0.4	0.27	0.54	100%
USA	0.22	0.08	0.35	99.9%

chances of being used with the *to*-infinitive. With the exception of the Indian variety, the base form *help* is the most likely to be followed by the bare infinitive. However, when the Helpee is present, the difference between the forms is small. Normally, the presence of the Helpee increases the chances of the bare infinitive, although its effect is quite small after the base form *help*, where the credible intervals largely overlap (see especially the US variety). In the Ghanaian variety, we even see a small increase in the chances of the *to*-infinitive.

6.2.4 Valency of the infinitive

Table 5 shows the numbers that represent the effect of transitivity of the infinitive on the presence of *to*. One can see that high probabilities (greater than 90%) are observed in the data from Hong Kong, India and Jamaica, followed by the USA (almost 87%). In the other countries, there is no strong bias in either direction. A separate check (not shown here) reveals that the presence of clause complements has no highly credible effects (close to 100%) in any of the varieties. In the USA, there is 92.3% probability that the clausal complements increase the chances of the bare form, followed by Jamaica (85.1%) and Hong Kong (84.2%).

6.2.5 Formality (average word length)

Finally, let us consider the degree of formality represented by the average word length of the text presented at an individual website. The posteriors in Table 6 show the effect of adding one letter on the log-odds of the *to*-infinitive vs. the bare infinitive. In most countries the average word length has positive effect on the chances of the *to*-infinitive, as predicted. The strongest effect is observed in Great Britain. The Ghanaian and US data display very weak positive effects. The Indian data show, surprisingly, the opposite effect: the longer the words in a text, the higher the chances of the bare infinitive.

7 Summary and discussion of the results

In general, the bare infinitive is the preferred variant in all varieties discussed here. The highest proportion of the bare infinitive is observed in the US data, whereas the lowest proportion is found in the Jamaican subcorpus, followed by the British data. The remaining countries exhibit proportions very similar to the British one.

The results of the previous studies are largely corroborated, although there are also quite a few new details.

- The variables related to *horror aequi* and the principle of cognitive complexity behave in accordance with the expectations in all varieties. They interact, such that the effect of *to* before *help* weakens with linguistic distance between *help* and the infinitive. Here, the models reveal no surprises.

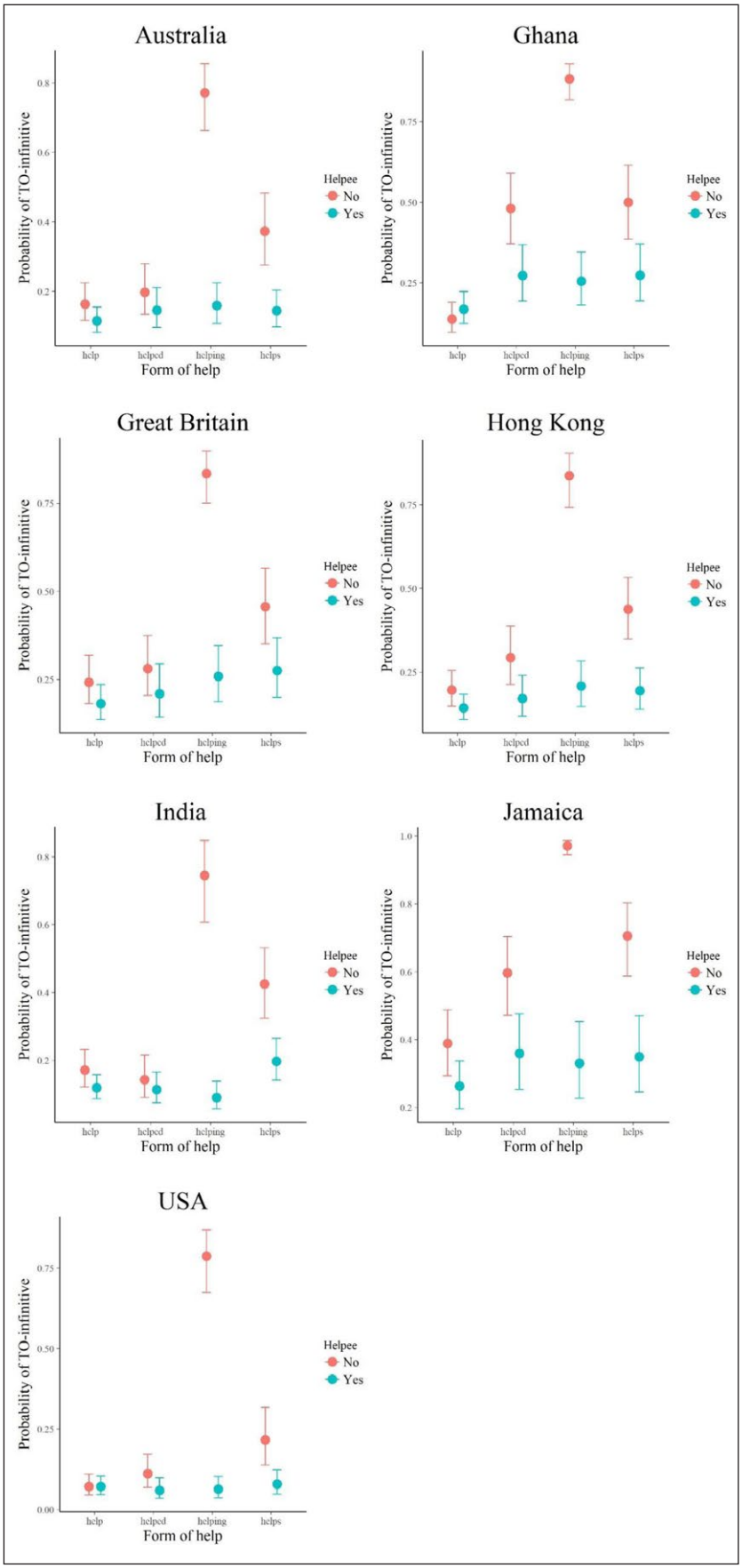

Figure 5: Posterior means and 95% credible intervals of the interaction between the form of *help* and the presence or absence of the Helpee. The vertical axis represents the probability of *to*-infinitive.

Table 5: Bayesian regression results for transitive vs. intransitive infinitive (log-odds ratios).

Country	Posterior mean	2.5%	97.5%	$P(\beta > 0)$
Australia	0	−0.16	0.17	49.5%
Ghana	−0.03	−0.21	0.16	38.7%
Great Britain	−0.01	−0.18	0.15	44%
Hong Kong	0.18	0.01	0.34	98.5%
India	0.13	−0.05	0.31	92.5%
Jamaica	0.16	−0.05	0.37	93.4%
USA	0.12	−0.09	0.33	86.6%

Table 6: Bayesian regression results for the average word length (log-odds ratios).

Country	Posterior mean	2.5%	97.5%	$P(\beta > 0)$
Australia	0.3	0.04	0.57	98.9%
Ghana	0.1	−0.17	0.36	76.1%
Great Britain	0.59	0.33	0.86	100%
Hong Kong	0.27	0.03	0.52	98.7%
India	−0.35	−0.63	−0.07	0.8%
Jamaica	0.3	−0.02	0.64	96.8%
USA	0.12	−0.2	0.44	76.6%

- The varieties also behave similarly with regard to the form *helping*, which substantially increases the chances of the *to*-infinitive. It is followed by *helps* in most varieties. However, the models demonstrate that this contrast is strong only in the absence of the Helpee. When the Helpee is explicit, the differences between the forms are small. For the base form *help*, the chances of the bare infinitive tend to be the highest, with or without the Helpee (except for the Indian variety, where *helped* is also very likely to be followed by the bare infinitive).
- As expected, the presence of the Helpee increases the chances of the bare form in all forms, with the exception of the base form *help*, when the presence of the Helpee makes little difference.
- There is a positive effect of the average word length, which serves as a proxy of formality, on the probability of the *to*-infinitive in most varieties, although it has low credibility in the Ghanaian and US subcorpora. Surprisingly, one finds a credible reverse effect in the Indian variety.
- There is also some evidence that transitive infinitives increase the chances of the *to*-infinitive in the varieties of Hong Kong, India, Jamaica and the USA, although this effect is only sufficiently credible in the data from Hong Kong. There are also some indications that the clausal complements play a role in some of the varieties, but these indications are very weak.

To summarize, there are very strong cross-lectal similarities with regard to the factors of *horror aequi* and cognitive complexity. As far as the other contextual factors (stylistic, morphological and syntactic) are concerned, most varieties behave in a similar way, but there are also exceptions. Interestingly, the US model often exhibits relatively weak effects in comparison with the other models. This may be due to the fact that the *to*-variant is the closest to extinction in that variety. The competition between the variants gradually disappears.

Let us now turn to the second question of the present study. The aim was to find out if constructional predictability determines the use of the bare or *to*-infinitive in the varieties of English and whether these effects (or lack thereof) are consistent. The generalized additive models show that there are some common effects in the expected direction in all seven varieties, but their directionality, strength and shape vary. The main conclusion one can draw is that the information content of HELP given a verb displays stronger and more systematic effects in the expected direction than the information content of a verb given the constructions. The infinitives that are associated with high information content of HELP are, as a rule, highly frequent verbs, such as *be*, *have*, *do*, *say*, *ask*, *try* and *use*.[5] These verbs appear in many diverse constructions, which explains the high information content of HELP. A few examples are provided in (12).

(12) a. (Hong Kong, blog, 3585980)
Growing plants will **help** you **to be** patient.
b. (India, general, 623003)
It will **help** your partner **to have** clear insight regarding your travelling habits.
c. (USA, general, 44601).
…if I try to **help** him **to do** it better, he gets an attitude and yells "I don't care about baseball".

To support this conclusion, Figure 6 displays the differences between the percentages of *to* in all examples and in those where HELP is highly informative given the infinitive (top 5%

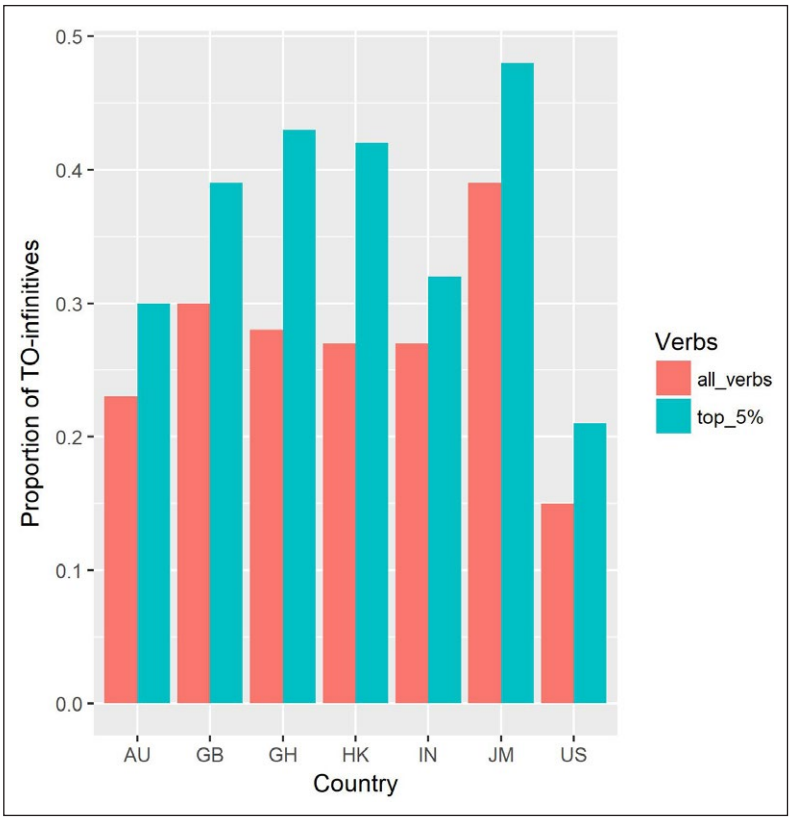

Figure 6: Percentages of *to*-infinitives in all contexts and in those where *help* given the infinitive is highly informative (top 5% of the scores).

[5] One might argue that the verbs be, have, and do are commonly used in the auxiliary function and may be special in some way. However, the effect is not limited to those verbs. Additional analyses show that a positive effect remains in all varieties even if one excludes the auxiliaries be, have and do.

of all scores in each variety). One can see that the proportion of *to*-infinitives is greater in the highly informative contexts than on average across all varieties.

These findings are rather unexpected. Most information-theoretic studies of grammatical alternations show that speakers tend to provide extra marking on unexpected units, such as the more frequent use of *that* before relative clauses (e.g. *I like the book (that) you gave me*) that are less predictable from the lexical properties of the noun phrase (Wasow et al. 2011), or the more frequent use of the case marker on semantically untypical direct objects in Japanese (Kurumada & Jaeger 2015). Unlike in those studies, the stronger predictability effect found here explains the marking on the infinitive, but the infinitive is not surprising itself. What is unexpected, is the construction in which it appears. Still, this effect can be considered a manifestation of the general tendency to maximize communicative efficiency. The speaker or writer provides an additional clue (i.e. the particle *to*) in those contexts when the infinitive is more difficult to recognize as a part of the construction with *help*. That makes the parsing of the utterance easier for the hearer or reader.

These findings open a new perspective for studies of constructional predictability in language. They show that predictability effects can go in different directions, which need to be examined separately. Moreover, it is important that the effects found in the present study are non-linear, which has not been addressed before, to the best of my knowledge. The results also demonstrate the importance of examining cross-lectal data when searching for universal functional principles of human language.

Acknowledgements

The author wants to thank the editors of this collection and the reviewers for their valuable constructive feedback, which helped to improve the quality of this paper considerably. The remaining errors are the author's responsibility. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n° 670985).

Competing Interests

The author has no competing interests to declare.

References

Aylett, Matthew & Alice Turk. 2004. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech* 47(1). 31–56. DOI: https://doi.org/10.1177/00238309040470010201

Bell, Alan, Jason Brenier, Michelle Gregory, Cynthia Girand & Dan Jurafsky. 2009. Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language* 60(1). 92–111. DOI: https://doi.org/10.1016/j.jml.2008.06.003

Biber, Douglas. 1988. *Variation across speech and writing.* Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511621024

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). DOI: https://doi.org/10.18637/jss.v080.i01

Bybee, Joan L. 1985. *Morphology. A study of the relation between meaning and form.* Amsterdam: John Benjamins. DOI: https://doi.org/10.1075/tsl.9

Davies, Mark. 2013. *Corpus of Global Web-Based English: 1.9 billion words from speakers in 20 countries*. http://corpus.byu.edu/glowbe/.

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

Duffley, Patrick J. 1992. *The English infinitive*. London: Longman.
Givón, Talmy. 1990. *Syntax: A functional-typological introduction* 2. Amsterdam: John Benjamins.
Haiman, John. 1983. Iconic and economic motivation. *Language* 59(4). 781–819. DOI: https://doi.org/10.2307/413373
Haspelmath, Martin. 2008. Creating economical morphosyntactic patterns in language change. In Jeff Good (ed.), *Linguistic universals and language change*, 185–214. Oxford: Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780199298495.003.0008
Huddleston, Rodney & Geoffrey K. Pullum. 2002. *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/9781316423530
Jurafsky, Daniel, Alan Bell, Michelle L. Gregory & William D. Raymond. 2001. Probabilistic relations between words: Evidence from reduction in lexical production. In Joan L. Bybee & Paul J. Hopper (eds.), *Frequency and the emergence of linguistic structure* (Typological Studies in Language 45), 229–254. Amsterdam: John Benjamins.
Klein, Dan & Christopher D. Manning. 2003. Accurate Unlexicalized Parsing. In *Proceedings of the 41th annual meeting of the Association for Computational Linguistics*, 423–430. http://nlp.stanford.edu/manning/papers/unlexicalized-parsing.pdf.
Kruschke, John. 2011. *Doing Bayesian data analysis. A tutorial with R and BUGS.* Amsterdam: Elsevier.
Kurumada, Chigusa & T. Florian Jaeger. 2015. Communicative efficiency in language production: Optional case-marking in Japanese. *Journal of Memory and Language* 83. 152–178. DOI: https://doi.org/10.1016/j.jml.2015.03.003
Leben, William. 1973. *Suprasegmental phonology*. Cambridge, MA: The Massachusetts Institute of Technology dissertation.
Levshina, Natalia. 2016. When variables align: A Bayesian multinomial mixed-effects model of English permissive constructions. *Cognitive Linguistics* 27(2). 235–268. DOI: https://doi.org/10.1515/cog-2015-0054
Levy, Roger & T. Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. In Bernhard Schlökopf, John Platt & Thomas Hoffman (eds.), *Advances in Neural Information Processing Systems (NIPS)* 19. 849–856. Cambridge, MA: MIT Press.
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
Mahowald, Kyle, Evelina Fedorenko, Steven T. Piantadosi & Edward Gibson. 2013. Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition* 126. 313–318. DOI: https://doi.org/10.1016/j.cognition.2012.09.010
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
Mair, Christian. 2015. Responses to Davies and Fuchs. *English World-Wide* 36(1). 29–33. DOI: https://doi.org/10.1075/eww.36.1.02mai
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
Norcliffe, Elisabeth & T. Florian Jaeger. 2016. Predicting head-marking variability in Yucatec Maya relative clause production. *Language and Cognition* 8(2). 167–205. DOI: https://doi.org/10.1017/langcog.2014.39

Piantadosi, Steven T., Harry Tily & Edward Gibson. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences* 108(9). 3526. DOI: https://doi.org/10.1073/pnas.1012551108

R Core Team. 2017. R: *A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

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.

Schmid, Hans-Jörg. 2000. *English abstract nouns as conceptual shells. From corpus to cognition*. Berlin: Mouton de Gruyter. DOI: https://doi.org/10.1515/9783110808704

Schmid, Hans-Jörg & Helmut Küchenhoff. 2013. Collostructional analysis and other ways of measuring lexicogrammatical attraction: Theoretical premises, practical problems and cognitive underpinnings. *Cognitive Linguistics* 24(3). 531–577. DOI: https://doi.org/10.1515/cog-2013-0018

Shannon, Claude E. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27. 379–423 & 623–656.

Stan Development Team. 2015. Stan: A C++ library for probability and sampling, version 2.8.0. http://mc-stan.org/.

Stefanowitsch, Anatol & Stefan Th. Gries. 2003. Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8(2). 209–243. DOI: https://doi.org/10.1075/ijcl.8.2.03ste

Szmrecsanyi, Benedikt. 2004. On operationalizing syntactic complexity. In Gérard Purnelle, Cédrick Fairon & Anne Dister (eds.), *Le poids des mots. Proceedings of the 7th International Conference on Textual Data Statistical Analysis* 2. 1032–1039. Louvain-la-Neuve: Presses universitaires de Louvain.

Walter, Mary Ann, & T. Florian Jaeger. 2008. Constraints on optional *that*: A strong word form OCP effect. In Rodney L. Edwards, Patrick J. Midtlyng, Colin L. Sprague & Kjersti G. Stensrud (eds.), *Proceedings from the Annual Meeting of the Chicago Linguistic Society*, 505–519. Chicago, IL: CLS.

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

Wasow, Thomas, T. Florian Jaeger & David M. Orr. 2011. Lexical variation in relativizer frequency. In Horst J. Simon & Heike Wiese (eds.), *Expecting the unexpected: Exceptions in grammar*, 175–195. Berlin: De Gruyter Mouton. DOI: https://doi.org/10.1515/9783110219098.175

Wood, Simon N. 2006. *Generalized additive models: An introduction with R*. Boca Raton, FL: Chapman and Hall/CRC.

Zipf, George. 1935. *The psychobiology of language: An introduction to dynamic philology*. Cambridge, MA: MIT Press.

How to cite this article: Levshina, Natalia. 2018. Probabilistic grammar and constructional predictability: Bayesian generalized additive models of *help* + (*to*) Infinitive in varieties of web-based English. *Glossa: a journal of general linguistics* 3(1): 55.1–22, DOI: https://doi.org/10.5334/gjgl.294

Submitted: 09 November 2016 **Accepted:** 26 January 2018 **Published:** 02 May 2018

]u[*Glossa: a journal of general linguistics* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS