

# Research Paper I: Determinants of variation in the needs doing alternation

Gabriel Doyle

November 14, 2008

## 1 Introduction

This paper investigates speaker choice in the *needs to be done*  $\sim$  *needs doing* alternation (which we will generally refer to as the “*needs doing* alternation”). We show that speaker choice in this alternation is not deterministic but probabilistic, conditioned on a variety of syntactic and pragmatic factors in the sentence.

This result is somewhat unsurprising; it serves to strengthen recent findings in other alternations that speaker choice is primarily probabilistic. However, the specific factors that affect this alternation are surprising. Unlike many other alternations, the alternants in the *needs doing* alternation have the same linear order of their elements. Many of the factors previously found to be relevant in speaker choice, such as accessibility and weight effects, have been explained as results of linear order changes. However, we see evidence for some of these “ordering” effects in this alternation as well, casting doubt on whether ordering is the true cause underlying these effects. We also encounter surprisingly strong lexical effects from the verb in the alternation.

We also find evidence supporting two general hypotheses motivating speaker choice. The first is that different syntactic categories in the alternants affect speaker choice through an environment prototypicality effect. This hypothesis predicts that, all else being equal, speakers prefer to use a syntactic category that is more typical given the rest of the sentence. The second hypothesis is that the *to be* in the *needs to be done* alternant functions like a syntactic disfluency (similar to *uh* or *um*) and indicates upcoming production difficulty for the speaker.

The paper is laid out as follows. Section 2 states the general problem of speaker choice and explains how syntactic alternations can be used to investigate the mechanisms that underlie it. This section also lays out the four questions about speaker choice that will be answered in this paper. Section 3 examines the use and syntax of the *needs doing* alternation and establishes the truth-conditional equivalence of the alternants. Section 4 discusses the nature of the gerund and past participle, the key difference between the two alternants, and explains why this alternation can be used to examine the effects of different syntactic categories on speaker choice. Section 5 introduces the mixed-effects logistic regression model we will use to investigate these constraints. Section 6 implements the model and identifies relevant factors in speaker choice. This section also evaluates the aforementioned hypotheses about speaker choice in light of this data. Section 7 poses follow-up questions from this experiment and discusses ways of strengthening the answers to the questions from Section 2. Section 8 concludes.

## 2 Background

### 2.1 Levels of Speaker Choice

Speaker choice takes place at a number of different levels. Consider the following scenarios:

---

<sup>0</sup>I would be remiss not to thank Joan Bresnan, Grant Goodall, T. Florian Jaeger, Andy Kehler, Roger Levy, Tatiana Nikitina, Doug Roland, Tom Wasow, Arnold Zwicky, and the members of the UCSD Computational Psycholinguistics Lab and Center for Research in Language for their comments and advice on this project and paper.

**High-level Semantic Choice** Bob and Mark go to a movie. As the credits roll and they file out of the theater, Bob turns to Mark and asks what he thought of the movie. Mark is a movie buff, and debates whether he should mention that he found striking visual parallels between this movie and an old Ingmar Bergman film or just state that he found the characters too contrived.

**Low-level Semantic Choice** Candace sees Paul at his desk around noon. She invites him to lunch, and he wishes to accept. But should he reply by saying “Yes, I’m interested in going to lunch with you,” or by saying “Yes, I’m starved”?

**Syntactic Choice** It’s the day before a big football game at Judy’s college, and someone has painted their rival’s colors on a statue on the main quad. Does she shout “Someone’s defaced the statue” or “The statue’s been defaced”?

**Phonological Choice** Jan is delivering a lecture on Paul Gauguin and wants to mention his quote “Art is either plagiarism or revolution.” How does she pronounce “either” in this context?

Speaker choice is an essential part of linguistics. Every time a person speaks, she must pass through a gauntlet of choices, through many levels. There are semantic and pragmatic levels, where the speaker chooses the idea to convey, and the specific semantic proposition she will use to convey it. These are the levels that Mark and Paul are working through as they debate what proposition to say. There are syntactic and lexical levels, where the speaker chooses words and constructs a sentence to state the intended proposition. These are the levels Judy is in as she decides between the active and passive voices. There are phonological and phonetic levels as the speaker pronounces the sentence she is composing. These are the levels Jan is in as she decides what vowel to use in *either*. And there are still other choices affecting the final delivery, such as when to breathe. Choices in one level affect choices in other levels, and in the end we are faced with a devilishly complex question, one central to psycholinguistics: How do speakers choose among the utterances that can express their conversational goals?

Unfortunately, we cannot look at sentence-level variation and expect to get much insight on this question. The question is simply too complex to attack head-on with what we currently know about speaker choice. There are too many factors to consider, and the space of possible sentences to express an idea is too ill-defined. Instead, we must start by breaking down the question into more manageable chunks. One possibility is to examine how speakers choose between alternants in a syntactic alternation. A specific definition of a syntactic alternation will be addressed in the next section, but for now it will be enough to think of a syntactic alternation as a situation in which there are two phrases that express truth-conditionally equivalent propositions. The dative and passive alternations are two common examples of this. In this paper, we will use the *needs to be done* ~ *needs doing* alternation (1a,b) as our window into speaker choice.

- (1) a. The couch needs to be cleaned.  
b. The couch needs cleaning.

Looking at speaker choice in syntactic alternations rather than at the sentence level simplifies the problem in three important ways:

- The alternants are truth-conditionally equivalent
- A smaller region of variation allows for better experimental control
- A limited set of alternants (usually two) makes comparison of options simpler

The first two points serve to limit the number of interacting levels of speaker choice we must consider. The truth-conditional equivalence of syntactic alternants minimizes semantic and pragmatic effects on speaker choice. The small region of variation in the alternation (usually consisting of at most a few words) limits the number of syntactic issues we must consider; we need to analyze the syntax of the sentence only in so far as it affects the alternant choice. The alternants themselves also represent a simplified syntactic choice, because when a speaker chooses an alternant, its internal structure is specified as well. For example, speaker

choice in the (agentless) passive alternation (Weiner & Labov, 1983) can be viewed as occurring with the patient and verb already chosen, so that the only question is whether the subject of the sentence should be the patient (the passive form) or a non-specific NP (the active form).

Syntactic alternations also simplify the analysis of speaker choice by limiting the number of choices to be considered. Once the speaker has decided to use an alternation, there are usually only two alternants available. If the speaker is considering only two options, we can use logistic regression to model the decision, rather than more exotic and less cognitively plausible statistical models. Logistic regression is also useful because its effect sizes intuitively relate to speaker choice. Additionally, logistic regression does not require any strong assumptions about the distribution of the data that the model is trained on (Agresti, 2002).

Because switching from sentence level variation to syntactic alternation simplifies the domain of speaker choice, it is likely that some factors that affect sentence level speaker choice will not manifest themselves in alternation-level speaker choice. Thus the present investigation is predominantly looking for positive evidence that a factor affects speaker choice. Showing that a factor does not affect choice in an alternation does not imply that the factor has no effect on speaker choice in other alternations or other levels of speaker choice, but we expect that the effects influencing speaker choice in syntactic alternations will also affect higher-level speaker choice.

Using syntactic alternations to investigate speaker choice is not a new idea; Weiner and Labov (1983) and Bock (1986) both investigated factors affecting speaker choice in the passive alternation, and in the past few decades, a variety of studies have investigated what syntactic alternations can tell us about the factors that influence speaker choice. However, there is still much to learn about speaker choice from alternation studies.

## 2.2 Defining Syntactic Alternations

Although alternations have been previously studied, there is no established definition of what does and does not constitute a syntactic alternation. This section uses previously studied alternations to clarify the issue. We propose that a syntactic alternation is a situation where two (or more) constructions satisfy the following conditions:

- (2)
  - i. there is significant commonality between the constructions
  - ii. the constructions are generalizable
  - iii. the constructions have the same truth conditions
  - iv. there is no obvious Gricean basis to choose between them
  - v. the speaker finds both of the constructions acceptable
  - vi. the constructions are common enough to be salient to the speaker

This definition can be motivated by looking at canonical examples of syntactic alternations, such as the following:

- (3) Passive Alternation (e.g., Weiner & Labov, 1983; Bock, 1986)
  - a. They broke into the closet. [active]
  - b. The closet was broken into. [passive]
  
- (4) Dative Alternation (e.g., Bresnan, Cueni, Nikitina, & Baayen, 2007; Bresnan & Nikitina, To appear)
  - a. Paul gave the man a present. [NP NP]
  - b. Paul gave a present to the man. [NP PP]
  
- (5) Genitive Alternation (e.g., Rosenbach, 2003)
  - a. The man's picture is overexposed. [s-genitive]
  - b. The picture of the man is overexposed. [of-genitive]

- (6) Left-dislocation Construction (e.g., Snider, 2005)
- a. Burlington’s crime doesn’t involve children. [normal]
  - b. Burlington’s crime, it doesn’t involve children. [left-dislocated]
- (7) *that*-omission (e.g., Jaeger, 2005, 2006)
- a. The car that I drove here is black. [included]
  - b. The car I drove here is black. [omitted]

In each alternation, there is substantial commonality<sup>1</sup> between the two options, as per condition (2)i. Also, in each case, the options are general constructions that are not specific to a single context. For instance, the genitive alternation can arise with most possessed NPs, and the passive alternation with essentially any transitive verb. This satisfies condition (2)ii, and stands in contrast to (8), where the two options are specific to the context of a object crossing something:

- (8) a. The mouse moved from one side of the table to the other.  
 b. The moving mouse crossed the table.

Turning to the semantics, a pair of alternants must have the same truth conditions. There is, for instance, no situation where (7a) is true but (7b) is false, nor vice versa. Weiner and Labov (1983) establish this as a necessary condition for a syntactic alternation, showing the truth-conditional equivalence of the agentless passive and non-specific active. However, we must go one step further into the semantics and also require that there is no Gricean basis for choosing between alternants. This Gricean concern is largely obviated by the truth-conditional equivalence; truth-conditional equivalence ensures the alternants are equally true (the Maxim of Quality), equally informative (the Maxim of Information), and equally relevant (the Maxim of Relevance). The only Gricean maxim that remains to be checked is the Maxim of Manner. This requires that neither alternant is both much briefer and clearer than the other. This rules out other possible truth-conditionally equivalent constructions; the dative alternation (4) does not include the option (9) because it violates the brevity condition while also reducing clarity.

- (9) Paul gave a present to the man and gave the man a present.

However, the Gricean condition does not exclude *that*-omission from the ranks of syntactic alternations. While the inclusion of *that* decreases the brevity of the alternant, it also often improves clarity. As such, the Maxim of Manner will not rule it out. Finally, speakers must consider both constructions valid. For instance, (10) is not an alternation:

- (10) a. The couch needs cleaning.  
 b. The couch needs cleaned.

Speakers of a dialect that permit (10b) almost certainly do not use (10a), and vice versa. As such, this is not an alternation because the choice between (10a) and (10b) is almost entirely dictated by the dialect of a speaker, and thus there is no actual choice between the two in most speakers’ minds. Because we are interested in language-internal effects on speaker choice in alternations, any effects we would find in a pair like (10) would be spurious. This raises an important concern in alternation studies; when studying an alternant, the researcher must be careful to ensure that the uses of the alternation come from speakers whose dialects allow both alternants. Furthermore, both alternants must be fairly common; if one of the alternants is much less common than the other, then it is unlikely that the alternants are in salient opposition as the speaker composes the sentence.

---

<sup>1</sup>The definition of commonality is left deliberately vague, as different constructions are similar in different ways. *That*-omission, for instance, has two essentially syntactically identical alternants, while the passive’s alternants similarity is entirely non-syntactic. We include commonality in the definition of a syntactic alternation primarily to assure that two massively different ways of saying the same thing are not unjustifiedly regarded as a syntactic alternation.

The conditions in (2) fall neatly into three pairs. The first two conditions define the range of construction pairs that could be viewed as an alternation. The second pair establishes that there is no clear semantic motivation determining speaker choice in the alternation. The last pair makes sure that speakers themselves view the constructions as alternants. Some of these conditions may prove to be relaxable; there may be an alternation whose alternants do not have an obvious commonality, for example. Satisfying these conditions is certainly sufficient to analyze two constructions as an alternation, but it is possible that satisfying all of them is not necessary.

## 2.3 Determinants of Speaker Choice in Syntactic Alternations

Alternation studies stretch back at least to the early days of the early days of the generative framework. Since that time, various linguists have analyzed alternations and created substantially different frameworks to explain the variability in an alternation. We review some prominent paradigms in the field, and focus on the gradient framework that this paper adopts.

### 2.3.1 Bolinger and the Need for Differences

The first view follows in the tradition of early generative works, and is best articulated by Bolinger (1968). Bolinger focuses on dispelling the notion that any two distinct constructions could be so similar that their differences could not be studied. His attack was levelled at structuralists, who had argued that distinct constructions could share the same meaning. This situation was impossible under the generativist assumption that two structures with identical meanings must have identical deep structures.

Defending the generative assumption, Bolinger argues that we should be strongly biased toward assuming that a difference in surface structure indicates a difference in deep structure. This deep-structure difference would then have to manifest itself as a semantic difference between the constructions. Bolinger notes that too often the semantic difference had been written off as a mere stylistic matter. As such, many constructions were speciously viewed as having identical meanings:

[Some claim that] active and passive are manifestly different, but the difference is ‘stylistic’, and can be ignored. This distinction is untenable. Since the speaker has as much of an option to choose either active or passive as he has to choose affirmative or negative, the use of ‘optional’ with one and not the other in the definition given above is meaningless. With no guide to any qualitative difference between style and meaning, all we can do is fall back on the quantitative one: a stylistic difference is one that is too small to count. But then how big must it be before it starts to count? It is asking too much for linguists to have to quantify meanings at this stage. Some day, perhaps, when we know more about them, but not now. (Bolinger, 1968, pg. 121)

Two examples are offered to illustrate the point. The first is a passive alternation:

- (11) a. Johnson has been misinformed by the Cabinet.  
b. The Cabinet has misinformed Johnson.

The passive form would be preferred, writes Bolinger, if one is interested in Johnson, rather than the Cabinet. This, he argues, is surely a difference in meaning. For an even subtler meaning distinction, Bolinger examines infinitival and gerundive complementizers, where he finds a distinction between potentiality and performance:

- (12) a. I like him to be nice to you. [infinitival]  
b. I like his being nice to you. [gerundive]

In the examples above, Bolinger claims that speakers prefer (12a) to express the wish that someone *will be* nice, while (12b) is used to refer to observed niceness. Citing other similar cases, he argues that infinitival clauses refer to hypothesized or potential behaviors and events, while gerundives refer to actual behaviors and events. Although this is a semantic difference between the two alternants, it still allows for certain situations where either alternant would be appropriate:

- (13) a. I like skiing.  
b. I like to ski.

The take-home message from Bolinger is thus: We should assume that any alternation reflects some true difference in meaning, even if incredibly slight. Note that the difference in meaning will not necessarily appear in the truth conditions of the sentence; for the passives, the two alternants have identical truth conditions.

Later alternation studies took Bolinger's result in two directions. Some adhere to Bolinger's hypothesis that the difference between constructions was categorical for semantic differences between alternants, while others look for gradient semantic differences and rejected the idea that alternations are driven by strict categorical semantic constraints.

### 2.3.2 The Meaning-to-Structure Mapping Hypothesis

The Meaning-to-Structure Mapping Hypothesis (MSMH) is a blanket term used by Bresnan et al. (2007) to refer to a set of proposed constraints on the dative and other alternations. The key observation underlying the MSMH is that the alternants in an alternation represent fundamentally different ways of characterizing an event. Gropen, Pinker, Hollander, Goldberg, and Wilson (1989) and Pinker (1989), for instance, propose that the NP NP dative alternant represented a change in possession, while the NP PP alternant represented movement to a goal:

- (14) a. Susan gave the children toys. [NP NP]  
b. Susan gave toys to the children. [NP PP]

For many situations, such as that of (14), both the change-in-possession and the movement-to-goal characterizations are acceptable; usually the children both claim and physically possess toys that are given to them. However, in some situations, one of the characterizations is inaccurate:

- (15) a. The lighting here gives me a headache. [NP NP; possession]  
b. \*The lighting here gives a headache to me. [NP PP; movement]

In this metaphorical use of *give*, it still makes some amount of sense to say that one possesses a headache, but it makes little sense to say that the headache has been transferred to the headache sufferer. Thus, MSMH adherents argue, only the possession alternant (NP NP) is available for this sentence; the movement-to-goal alternant (NP PP) is ruled out.

This is an extension of Bolinger's results on the *for-to* and *ing* complementizers. The different alternants have the same truth conditions, but have categorical semantic constraints on their usage. "Categorical" here means that either a situation satisfies the constraint, and thus the alternant is available, or it does not satisfy the constraint and is unacceptable. Under this view, the constructions alternate only where both characterizations are valid, and even in these situations, the alternation is determined primarily by the characterization that the speaker prefers. Speaker choice reduces to an essentially one-dimensional problem under the MSMH, although there could be some room for other factors when both characterizations are acceptable.

### 2.3.3 The Rise of Gradience

Although the MSMH is intuitively pleasing, since it reinforces the primacy of semantic concerns in speaker choice. This fits nicely with a "say what you mean" principle. However, the MSMH is too stringent. Early evidence against the MSMH came from Green (1971), who showed that syntactic factors such as pronominality can override the MSMH predictions; in some situations, people are willing to use a structure that improperly characterizes the situation. Bresnan et al. (2007) finds further evidence of this in web attestations:

- (16) From the heads, offal and the accumulation of shy, slimy matter, a stench or smell is diffused over the ship that would *give a headache to the most athletic constitution*.<sup>2</sup>

<sup>2</sup><http://www.downhomer.com/Webmag/2000/0007/page36.html>

Here, the desire to place the heavy goal NP *the most athletic constitution* at the end of the clause overrules the MSMH injunction against the NP PP option for situations that do not involve transfer to a goal. These results suggest that meaning-to-structure mappings are not categorical, but rather gradient. “Gradience” here means that each interpretation of a situation can have a preference for one structure over the other, as under the MSMH, but this preference may be overruled by other considerations, such as the heavy-NP shift in (16). Speakers will favor the structure that best fits their characterization of the situation, but are not beholden to it.

This gradient view is actually an extension of the MSMH’s categorical viewpoint. A gradient constraint decreases the likelihood of using an alternant in certain situations, but does not rule it out entirely. However, as the strength of the gradient constraint increases, the likelihood of using an alternant will tend to zero. In the limiting condition of an infinitely strong gradient constraint, categoricity results. A constraint can still be effectively categorical if it is stronger than all of the other constraints on an alternation combined, so the MSMH’s categorical version of speaker choice is a subspace of the gradient version.

Let’s look at an example of speaker choice in the gradient framework. For example, Pinker (1989) and others proposed a categorical constraint under the MSMH that “verbs of manner of speaking” (e.g., *whispered*, *muttered*) could not occur in the NP NP dative alternant:

- (17) a. \*Susan mumbled Rachel the news. [NP NP]  
 b. Susan mumbled the news to Rachel. [NP PP]

However, NP NP alternants can occur with these verbs, given the right situation (Bresnan & Nikitina, To appear, (14)):

- (18) a. You just mumble him an answer.  
 b. [...] she muttered him a hurried apology as well before skirting down the hall.

Moreover, these sentences sound better than their NP PP alternants:

- (19) a. (?)You just mumble an answer to him.  
 b. (?)She muttered a hurried apology to him as well before skirting down the hall.

The key to these examples is that the recipient in both (18a) and (18b) is *him*, a short, pronominal, concrete NP, while the theme in each sentence (*an answer* in (18a), *a hurried apology* in (18b)) is a longer, indefinite, abstract NP. As shown in (Bresnan et al., 2007), gradient constraints favor short, pronominal, concrete NPs preceding long, indefinite, abstract NPs in this alternation. Verbs of manner of speaking exert their own gradient preference for the NP PP form, which explains the preference for (17)a over (17)b. In (18a) and (19a), though, we see that the preference for *him* to precede *an answer* is strong enough to overcome *mumble*’s preference for the NP PP form. These preferences cannot be accounted for in the categorical framework.

### 2.3.4 Modelling Speaker Choice in the Gradient Framework

Assuming the gradient framework is valid, how do we go about assessing the strength of the various gradient factors? In the gradient framework, speaker choice is driven by a set of gradient constraints that somehow influence which alternant will be chosen. Cognitively, this could be realized in a variety of ways. We will stay intentionally agnostic about the specific mechanism of speaker choice, and simply assume that speaker choice is probabilistic, with the constraints exerting gradient influences on the probability of the speaker choosing a given alternant. In this framework, we presume that, given a sentence context  $S$ , the gradient constraints combine to yield an overall probability  $P(\text{needs to be done}|S)$  of choosing the *needs to be done* alternant for this sentence. We include a simplifying assumption that  $P(\text{needs to be done}|S) + P(\text{needs doing}|S) = 1$  for every  $S$  — that is, the speaker considers only the two alternants we have identified in each context. This framework implies that the same speaker can choose a different alternant each time he sees an environment, because the choice of an alternant is probabilistic. This fits intuitively with what we see in actual usage of alternations.

Using this probabilistic framework for speaker choice, we can model the probability  $P(\text{needs to be done}|S)$  with a probabilistic model. Specifically, we use a mixed-effects logistic regression model of speaker choice. In logistic regression, each of the gradient constraints on speaker choice is a weighted factor on the odds of choosing one alternant over the other. The mixed-effects regression accounts for the verb-specific effects that will be discussed in Section 2.4. Logistic regression is standard for the analysis of syntactic alternations, uncovering the previously observed gradient effects. We will discuss the will be covered in depth in Section 5.1.

### 2.3.5 Known Constraints on Speaker Choice

Previous work has identified a variety of gradient effects on speaker choice. Many of these factors appear to share a common motivation to put more accessible elements in certain positions in the sentence. These include animacy, concreteness, discourse status, topicality, definiteness, pronominality, and weight effects. In addition, there are gradient effects with other motivations, such as structural parallelism and information-theoretic constraints. This section offers a quick overview of these known gradient effects, many of which also affect the *needs doing* alternation.

**Animacy Hierarchy** Animacy and concreteness have been identified as factors driving speaker choice in many alternations (Prat-Sala, 1997 for passives, Bresnan et al., 2007 for datives, Rosenbach, 2003 for genitives, Snider, 2005 for left-dislocation). In each of these alternations, speakers prefer the alternant that places an animate NP earlier in the sentence. For the genitive alternation, Rosenbach explains this as an result of a cross-linguistic preference for animate NPs to occur earlier in a sentence (Siewierska, 1988; Ortman, 1998; Yamamoto, 1999), which has been argued to be a cognitive universal based on increased cognitive accessibility of animate and concrete NPs (Dahl & Fraurud, 1996; Haspelmath, 1999). This argument easily extends to the other alternations as well. Concreteness has not shown as clear of results as animacy, largely because many of the studies consider animacy and concreteness together. When the two factors are separated (e.g., Bresnan et al., 2007) concreteness shows a preference to occur earlier in the sentence, as with animacy. This fits with Silverstein’s animacy hierarchy (Silverstein, 1976), where animate concrete NPs are most accessible, followed by inanimate concrete NPs, followed by inanimate abstract NPs. Since we observe the strongest position effects for animate concrete NPs and somewhat weaker position effects for inanimate concrete NPs, it appears that the generalization to draw from these alternations is that speakers tend to choose the alternant that places the NPs that are further up the animacy hierarchy earlier in the sentence. It should be noted that the animacy hierarchy does not always push animates forward. (Snider, 2005) discusses topicalization, where inanimates are more likely to topicalize (and thus move forward in the sentence) than animates. However, this may be due to a preference for animates to already occur earlier in the sentence and thus to have no need to topicalize.

**Discourse Status** Discourse status also affects speaker choice in these alternations, and as with animacy and concreteness, this appears to be due to a tendency for previously mentioned (discourse-old) elements to occur earlier in the sentence. Rosenbach (2003) notes a preference for topical possessors to take the s-genitive, placing them earlier in the sentence. Similarly, Bresnan et al. (2007) note that given NPs occur earlier in the dative alternation than non-given NPs do. Again, it appears that speakers use accessibility to decide what should come first in an alternation. This carries over to pronominality and definiteness as well; pronouns are more accessible in discourse than names and definite NPs, which are themselves more accessible than indefinite NPs. This is reflected in the ordering of NPs as well: pronominals tend to come first, followed by definite NPs, and lastly indefinite NPs.

**Weight Effects** Weight effects are common in these alternations as well. Zaenen, O’Connor, and Wasow (2007) note that speakers prefer the of-genitive with heavier possessors (measured by length in words), so that the heavy-NP occurs later in the sentence. Similarly, heavier NPs occur later in the dative alternation (Bresnan et al., 2007). This is a general result that holds outside of syntactic alternations as well; heavy-NP

shift has been noted across syntactic structures and across languages. As with the previous effects, speaker choice here is influenced by a desire to move an NP to a different part of the sentence.

All of the aforementioned effects could be motivated by ordering preferences, with easier, shorter, and more salient things coming earlier in the sentence. However, we will see that some of these factors are also active in the *needs doing* alternation, which has no ordering changes. This suggests that the ordering-preference motivation alone is insufficient to explain these effects.

**Parallelism** Moving beyond the “ordering-preference” factors, we encounter structural parallelism. Speaker choice is influenced by the recent occurrence of a similar structure; if, for instance, a speaker has just heard a passive sentence, he is more likely to use a passive sentence in the near future (Bock, 1986). Likewise, Bresnan et al. (2007) show that syntactic parallelism influences speaker choice in the dative alternation. Unlike the previous factors, parallelism influences speaker choice independent of ordering; instead, it influences the speaker to repeat the same structure that the speaker has been primed with. Parallelism effects are generally outside the scope of this project. We investigate one parallelism-like factor on the *needs doing* alternation, but find it to be insignificant.

**Information-Theoretic Effects** Finally, some alternations are also subject to information-theoretic effects. Jaeger (2005) argues that *that*-omission is driven in part by audience design, as a signal to upcoming processing difficulty in the sentence. Also in the *that*-omission alternation, Levy and Jaeger (2007) find that omission is dependent on predictability in a sentence; when the presence of a relative clause is more predictable, the *that* is more likely to be omitted. Although these information-theoretic effects presumably also occur in other alternations, this has not been extensively investigated. Formal investigation of such effects are outside this project’s scope, but we will review some informal evidence for information-theoretic effects in Section 6.4.1.

An interesting result from these studies is that each alternation behaves differently. A factor can affect one alternation but not another, and the same factor can affect different alternations differently, as seen with topicalization and left-dislocation. With each new alternation that is studied, we stand to gain additional data about how various factors affect speaker choice and how the factors interact. This in turn gives us insight into the underlying sources that lead to the observed gradient effects.

## 2.4 Open Questions

What do we stand to gain from studying the *needs doing* alternation? We answer four specific questions about speaker choice in the present study.

**Does environment prototypicality affect speaker choice?** This study gives evidence for environment prototypicality effects, where the use of different syntactic categories in different alternants affects speaker choice. Many alternations have different structures for their alternants. The dative alternation switches between an NP and a PP in the predicate or a pair of predicate NPs; the genitive switches between a determiner in the s-genitive and a PP in the of-genitive. However, the possible effects of such category change have not been studied because the changes to the syntactic category are inseparable from changes to the word order. In the *needs doing* alternation, the word order remains fixed, so the effect of syntactic category changes can be disentangled from the effect of word order, as we will see in Section 4.1<sup>3</sup>.

**Can accessibility effects exist without word-order variation?** The *needs doing* alternation also offers an interesting testing ground for the accessibility effects that have been noticed in the other alternations. These effects have been explained as the result of a general preference for more accessible NPs to occur earlier in a sentence. In the *needs doing* alternation there is no difference in the order of words; only the syntactic category and alternant length change. Nevertheless, accessibility effects are observed for the *needs*

---

<sup>3</sup>Although the *needs doing* alternation allows for a cleaner contrast than the other alternations by removing effects of word order, it is still not a perfectly clean contrast, as *to be* intervenes between *needs* and the verb in one alternant but not the other.

*doing* alternation, as detailed in Section 6.4. The presence of these effects in a constant-order alternation shows that accessibility effects are more than just simple linear ordering preferences.

**Are there lexical effects of verb choice?** Because the *needs doing* alternation can occur with different verbs (e.g., *needs cleaning*, *needs reminding*), we are able to study the effect of different lexical items in an alternation. This follows Bresnan et al’s (2007) work on the dative alternation, which included models that accounted for effects of different verbs and different verb senses. The dative alternation only occurs with a limited set of verbs compared to the *needs doing* alternation; the Bresnan et al dataset contained 55 verb senses, while we observe over 300 different verbs in our dataset. In our model of the *needs doing* alternation, some common verbs exert influences as strong as those of the non-lexical factors we consider. This is interesting on two fronts. First, it is interesting that strong lexical effects appear in a wide range of verbs, given the relatively rare usage of this alternation compared to the ubiquity of the dative alternation. Second, it suggests that there may be some undetected gradient semantic issues, perhaps similar to the categorical effects proposed by Pinker, 1989; Gropen et al., 1989, affecting speaker choice.

**Is *to be* a disfluency?** Lastly, this alternation may show evidence of a disfluency effect with a semantically-meaningful element. The *needs to be done* alternant has two words between *needs* and the past participle, whereas the *needs doing* alternant has the gerund abutting *needs*. It is possible that *to be* serves a similar purpose to a disfluency, either by buying time for the speaker to alleviate production difficulties (Ferreira & Dell, 2000) or by signaling to the audience that the speaker anticipates production difficulties (Clark & Fox Tree, 2002). This would resemble the disfluency analysis of *that*-omission in Ferreira and Dell (2000); Race and MacDonald (2003); and Jaeger (2005). In *that*-omission, though, *that* is a function word that carries no semantic information and can be omitted without affecting the sentence’s meaning. In the *needs doing* alternation, the “disfluency” *to be* carries the semantic information that the upcoming verb is passivized. Its omission leads to a meaningless sentence.<sup>4</sup> The observation of a disfluency effect in the *to be* form suggests that a disfluency need not be semantically empty; possible evidence for such an effect is discussed in Section 6.4.1.

To explain how these effects manifest themselves in the *needs doing* alternation, we must first examine the structure and distribution of the alternation. The next section discusses the current state of knowledge regarding the *needs doing* alternation.

### 3 *Needs doing* and *needs to be done* as a Syntactic Alternation

In this section, we establish that the choice between *needs doing* and *needs done* is indeed a syntactic alternation, based on the definition developed in Section 2.2. It is obvious that the two constructions have a great deal of syntactic commonality. Likewise, it is clear that the alternation is general, as almost any transitive embedded clause can occur with this alternation. There is no Gricean basis for choosing between the two, either, as both forms are approximately equally concise. The two outstanding questions are whether the alternants are saliently opposed for speakers and whether the alternants are truth-conditionally equivalent and acceptable in the same contexts. Both of these conditions are met, at least in British English. We begin with analysis of how the constructions are used, demonstrating that they form a salient opposition. We then examine the semantics of the constructions, and conclude that there is no difference in their truth conditions. In the process, we will also examine the syntactic structure of the alternants.

#### 3.1 Use of the Constructions

Not all speakers have the *needs to be done* ~ *needs doing* alternation in their dialect. Instead, these speakers use the *needs done* construction (20) in alternation with *needs to be done*.

(20) The couch needs cleaned.

---

<sup>4</sup>Unless the speaker comes from a *needs done* dialect region, that is.

The distribution of this non-standard construction has been studied at length (Murray, Frazer, & Simon, 1996; Murray & Simon, 1999, 2002). It is generally limited to Appalachian and North Midland American English speakers in the United States, and to Ulster and Scottish Lowland English speakers in the United Kingdom. Furthermore, even in areas where *needs done* is in use, it does not dominate; there is almost always a sizable portion of the population that does not use *needs done*. This limited distribution is essential to our analysis, as it means that almost all of our speakers have the standard *needs to be done* ~ *needs doing* alternation.

Unfortunately for our purposes, many American speakers of English exhibit a strong preference for the *needs to be done* form over *needs doing*. In the Corpus of Contemporary American English (Davies, 2008–), there are only 56 instances of the string “needs doing”, compared to 1199 instances of the string “needstobedone”.<sup>5</sup> As such, it is debatable whether both constructions are really salient to speakers of Standard American English.

Luckily, British speakers are substantially fonder of the *needs doing* construction. In the British National Corpus (BNC), there are only three times as many *needs to be done* constructions as *needs doing* constructions. This suggests that, while *needs to be done* is still preferred, both alternants are salient to British speakers as they construct sentences.

Further evidence that Britons are actively considering the two alternants during sentence construction comes from the fact that often both alternants are used in a single discourse by the same speaker. On occasion, the same speaker will even use both alternants in the same sentence, as in these examples from the BNC:

- (21) a. Some people believe that they only *need to be turned* once more, while others say that they *need scrapping*.  
 b. [...] in terms of the time that *needs spending* doing the job even if it’s relatively simple in terms of what *needs to be done* [...]

Further examples are attested on British websites:

- (22) a. Life is pretty much like the UK on a day to day basis; kids go to school, shopping *needs to be done*, housework *needs doing* [...]<sup>6</sup>  
 b. Whatever *needs doing*, and whenever it *needs to be done*, eZ PC can deliver!<sup>7</sup>  
 c. There is so much more that *needs doing* — a whole global system which *needs to be changed*.<sup>8</sup>

In these sentences, the contexts where the two alternants are used are extremely similar. The change in alternant choice across similar contexts shows that the speaker considered both alternants in the second context, and likely considered both in the first context as well. While this is no guarantee that all British speakers actively consider both alternants in all situations, it shows that often both alternants are salient for a single speaker. This salience suggests that it is indeed fair to examine *needs to be done* ~ *needs doing* as an alternation — so long as we restrict our sample to British English. We will train the logistic regression model on data from the BNC to take advantage of this.

One task remains before we can confidently declare this an alternation: demonstrating that *needs to be done* and *needs doing* are truth-conditionally equivalent. We start by examining the syntax of the constructions. First we show that the syntactic differences between the constructions do not induce different truth conditions. We then delve further and show that there is a syntactic difference that in related constructions induces meaning differences, albeit below the level of truth conditions. We show evidence that this difference does not induce any categorical meaning differences in our alternation.

<sup>5</sup>There are only six instances of “needs done”, so *needs done* users do not account for this discrepancy.

<sup>6</sup><http://britishexpats.com/articles/australia/first-year-in-perth/>

<sup>7</sup><http://www.ezpcrepair.co.uk/>

<sup>8</sup><http://www.sanctus1.co.uk/blog/2006/02/fairtrade-high-street.html>

## 3.2 Syntactic Structure and Truth-Conditional Equivalence

Following Weiner & Labov, 1983, we need to establish truth-conditional equivalence between the constructions. There are two syntactic concerns that could lead to semantic differences, but we find that neither affects the truth conditions.

### 3.2.1 Raising in the Constructions

The first issue in each construction is the nature of the embedded subjectless clause (e.g. *to be repaired* in (23a)).

- (23) a. The drainspout needs [to be repaired].  
b. The child needs [reprimanding by a teacher].

The Extended Projection Principle requires these clauses to have a subject, and the understood subject of the embedded clause is the matrix subject (*the drainspout* in (23a)). Thus the question reduces to whether the subjectless embedded clause is the result of raising or subject control. If these are raising constructions (25), then the matrix subject forms a chain with the embedded subject. If these are control constructions (24), then the matrix subject is base-generated in the matrix clause but controls a PRO in the embedded clause:

- (24) **Raising:** [The drainspout]<sub>i</sub> needs [e<sub>i</sub> to be repaired].  
(25) **Control:** [The drainspout]<sub>i</sub> needs [PRO<sub>i</sub> to be repaired].<sup>9</sup>

The difference between (24) and (25) is of more than theoretical importance. A single NP chain can only receive one  $\theta$ -role. Clearly the embedded subject receives a  $\theta$ -role from the embedded VP. If the matrix subject is raised from the embedded subject position, then it cannot receive a second  $\theta$ -role from *need* in the matrix clause. In that case, the matrix subject would not be an argument of *need*, and *need* will only assign a  $\theta$ -role to the embedded proposition. Thus (24) would be paraphrased as “What needs to happen is the drainspout getting repaired,” with no explicit statement of who needs the drainspout repaired. If, on the other hand, there is a PRO in the embedded subject position, then the matrix subject must receive its  $\theta$ -role from *need*. Then *need* would assign two  $\theta$ -roles, one to the embedded proposition (what needs to be done) and one to the matrix subject (who needs to have the proposition done). Thus (25) would be paraphrased as “The drainspout needs the drainspout to get repaired,” where the drainspout explicitly needs itself repaired.

Note that the constructions do not need to have the same structure; one could exhibit raising and the other could exhibit subject control. If that were the case, there would be a clear categorical semantic distinction between them. However, standard raising tests reveal that both are raising constructions.

**Selectional restriction test** The discussion above already gives the first piece of evidence that these are raising constructions. If the matrix subject were assigned a  $\theta$ -role by *need* — as in the control structure — then the matrix subject would almost certainly have to be sentient. Inanimate objects are generally incapable of needing anything. Sentences with inanimate matrix subjects, such as (23)a and (26), are acceptable, presenting evidence for the raising structure.

- (26) a. [The needs of the patient, the department and the ENP] thus need to be balanced.  
b. [This traditional mould] needs oiling and flouring before baking.<sup>10</sup>

---

<sup>9</sup>For the *needs doing* construction, the matrix subject would raise from the embedded clause’s object position instead. It is not immediately clear how a control structure could be obtained in the gerundive clause. Intuitively, the controlled PRO would be the object of the gerund, but then PRO would be governed by the gerund, violating the PRO Theorem. Since the raising tests reveal that *needs doing* is a raising construction, it is not important exactly how a control structure could be realized. We perform the raising tests for *needs doing* to give a stronger argument for the raising structure than simply that there is no plausible control structure.

<sup>10</sup>Both examples are from the BNC.

**Dummy test** Similarly, a control structure would not allow for expletives as the matrix subject, since expletives could not receive *need*'s  $\theta$ -role. But expletive subjects are attested in both constructions in the BNC:

- (27) a. [...] it needs to be noted that the absolute value of the index is higher than the old index.  
 b. It hardly needs remarking that the experimental procedure of varying the circumstances fits exactly the account [...]

**Idiomatic chunk test** Lastly, the idiomatic interpretation of an idiom is lost if a  $\theta$ -role is assigned to an idiom chunk. Thus, if these were subject control structures, it would be impossible to get an idiomatic reading in the construction. Of the three tests, this one is the hardest to find examples for. There do not seem to be any in the BNC, but a few examples come up in web searches.

- (28) a. Apparently enterprising fans decided the cat needed to be let out of the bag sooner than the official release date, so the cd has been leaked on the net.<sup>11</sup>  
 b. Pass the buck when the buck needs passing.<sup>12</sup>

Despite their rarity, such sentences seem to be grammatically acceptable and definitely retain the idiomatic meaning, which is the final piece of evidence that both constructions have raising structures. Both constructions having the same structure is a necessary condition for them to be truth-conditionally equivalent, so this is a reassuring result.

### 3.2.2 The Embedded Passive

Another potential difference between the *needs to be done* and *needs doing* constructions is that the former contains a passive clause. Passivizing a verb absorbs its external  $\theta$ -role (Jaeggli, 1986; Baker, Johnson, & Roberts, 1989). This explains why the logical subject argument, usually the agent, is not realized as an NP. Semantically the verb retains the thematic role, as there is still some unstated agent. Baker, Johnson, and Roberts argue that this unstated agent is an “implicit argument” of the verb. The centerpiece of Baker, Johnson, and Roberts’s justification for the implicit argument is the fact that rationale clauses (RatCs) are acceptable in passives but not in other agentless voices, such as the middle:

- (29) a. The bureaucrat was bribed [<sub>RatC</sub>PRO to avoid the draft].  
 b. \*The bureaucrat bribes easily [<sub>RatC</sub>PRO to avoid the draft].

(29a) is grammatical and means that some specific person avoided the draft by bribing the bureaucrat. This specific but unstated person is represented by the implicit argument of *was bribed* in Baker, Johnson, and Roberts’s analysis, and thus can control PRO. In (29b), there is no referent to control PRO because there is no implicit argument. It is unclear whether gerunds have an implicit argument, although it seems likely as it is possible to use RatCs with gerunds in some situations:

- (30) a. Bribing the bureaucrat [<sub>RatC</sub>PRO to avoid the draft] was easy.  
 b. Boiling water [<sub>RatC</sub>PRO to purify it] is not the solution, as it does not remove contaminants.<sup>13</sup>

However, we are less concerned with whether the gerund has an implicit argument than we are with whether there is a difference in the behavior of the *needs to be done* and *needs doing* constructions due to the presence or absence of an implicit agent. If there were such a difference, we would expect that RatCs would be ungrammatical with the *needs doing* construction. This is not the case, though.<sup>14</sup>

<sup>11</sup><http://timewarpandnewrocknews.blogspot.com/2006/03/moz-wait-is-over-for-some.html>

<sup>12</sup><http://www.articleview.net/Advice/Stress-Prevention-in-a-Word-No.html>

<sup>13</sup><http://www.purwater.com/food>

<sup>14</sup>“In order to” RatCs are used here because it is easier to search for such RatCs than it is to search for “to” RatCs. In either type of RatC, the PRO is controlled by the implicit agent, so there is no difference between these RatCs for our purposes.

- (31) a. If our staff identifies a hazardous condition (such as a cable that needs replacing [<sub>RatC</sub>in order PRO to safely lift your dock or track]) we will remedy such condition at our discretion [...]<sup>15</sup>  
 b. [<sub>RatC</sub>In order PRO to fix this problem], the chip needs replacing with a Revision B 35-PAC-AUB.<sup>16</sup>

Whether or not the gerund has an implicit argument, it behaves in the same way as the passive; the implicit agent of the gerund can control the PRO in a rationale clause, and thus maintains the truth-conditional equivalence of the constructions.

### 3.3 Potential Lower-Level Semantic Differences

We have shown that the syntax of the constructions lead to no apparent truth conditional differences. However, the door is still open for lower-level semantic constraints. We examine some potential meaning differences that would operate below the level of truth conditions and show that if these effects exist, they are minor gradient factors in speaker choice, and not categorical constraints. Only the first two of these differences are syntactically motivated; the others are non-syntactic and are quickly shown to be non-categorical.

#### 3.3.1 Infinitives and Gerunds: Potentiality and Performance

Note that *need* is different from other verbs that can take either infinitival or gerundive phrases as objects:

- (32) a. The duke *needs to be appeased*.  
 b. The duke *needs appeasing*.  
 c. The duke *needs to appease* the king.  
 (33) a. The duke *loves to be appeased*.  
 b. The duke *loves appeasing* the king.  
 c. The duke *loves to appease* the king.

For *need*, the infinitival-passive form (32a) and gerundive form (32b) are paraphrases of each other, while the infinitival-active form (32c) means something else entirely. For other verbs, it is the infinitival-active form (33c) that is a paraphrase of the gerundive (33b), while the infinitival-passive form (33a) means something else. Despite this difference, there is a clear resemblance between the forms; in each case, there is an infinitival paraphrase of the gerundive.

This is a useful resemblance, because the choice of infinitival and gerundive clauses with the *love*-like verbs have been more examined more extensively than the choice with *need*. Quirk, Greenbaum, Leech, and Svartvik (1985, pg. 1191) note that “where both constructions [(33b) and (33c)] are admitted, there is usually felt to be a difference of aspect or mood which influences the choice. As a rule, the infinitive gives a sense of mere ‘potentiality’ for action ... while the participle gives a sense of the actual ‘performance’ of the action itself.” This resembles Bolinger’s distinction between the potential actions of infinitival clauses and the realized actions of gerundives from Section 2.3.1. As with Bolinger’s, this distinction manifests itself both in the interpretation of different sentences (e.g., (12)) and in the differential acceptability of some minimal pairs:

- (34) a. I would like to purchase some nicer shoes.  
 b. ??I would like purchasing some nicer shoes.  
 (35) a. ??I hated to live in that cramped dorm room.  
 b. I hated living in that cramped dorm room.

The conditional in (34) makes it clear that the event of purchasing shoes is still hypothetical. Thus the potentiality of the infinitive (34a) is a good fit, while the performance aspect of the gerund (34b) is a bad

<sup>15</sup><http://www.eagledocks.com/public/service/Fall%202008%20Service%20Instructions.pdf>

<sup>16</sup><http://www.amiga-hardware.com/showhardware.cgi?HARDID=284>

fit. Similarly, the past tense in (35) implies that living in the room has happened, the potentiality is now a bad fit (35a), while the performance of the gerund (35b) is a good fit.

However, for some verbs, Quirk et al note that “the difference is more subtle, and may be overruled or neutralized by the meaning of the verb” (pg. 1191). They cite *escape* as an example; its inherent negativity cancels out the ‘performance’ sense, so that in *Oppenheimer escaped being convicted of treason*, there is no implication that Oppenheimer was actually convicted of treason, despite the use of the gerundive.

Is *need* more like *love* or *escape*; does it manifest this potentiality/performance dichotomy or not? It seems that, like the inherent negativity of *escape*, the inherent potentiality of *need* cancels out the sense of performance in the gerundive form. If the potentiality/performance distinction exists with *need*, then we expect to see that the infinitival form is preferred in situations where the event is explicitly denied, and the gerundive form to be preferred when the event is explicitly completed. This can be tested using contextualized stimuli similar to those in So, 1973. So tested the potentiality/performance distinction for verbs like *try* by setting up two possible contexts for the constructions. Both contexts use the past tense. One context implies that the action was successfully completed (36a), biasing the respondent toward the gerund’s performance aspect. The other context implies that the action failed to occur (36b), biasing the respondent toward the infinitive’s potentiality aspect.

- (36) a. **Success:** I tried  $\left\{ \begin{array}{l} \text{closing} \\ \text{to close} \end{array} \right\}$  the window, but that didn’t help. I still felt cold.  
 b. **Failure:** I tried  $\left\{ \begin{array}{l} \text{closing} \\ \text{to close} \end{array} \right\}$  the window, but I couldn’t. It was stuck.

So found that for six different verbs, including *try*, native speakers preferred the gerund in the successful condition and preferred the infinitive in the failure condition, as predicted by Bolinger. Similar stimuli can be used to see if the *need* constructions exhibit the potentiality/performance distinction:

- (37) a. **Success:** The window needed  $\left\{ \begin{array}{l} \text{closing} \\ \text{to be closed} \end{array} \right\}$ , but that didn’t help. I still felt cold.  
 b. **Failure:** The window needed  $\left\{ \begin{array}{l} \text{closing} \\ \text{to be closed} \end{array} \right\}$ , but I couldn’t. It was stuck.

A small run of these stimuli revealed no such preference for *need*. Eleven native English speakers were given four stimuli sets like those in (37), and were asked to rate the acceptability of each sentence on a scale from 1 (wholly unacceptable) to 7 (perfectly acceptable). The average ratings were:

	gerund	infinitive
success	5.32	6.98
failure	5.84	6.48

All of these ratings are high, indicating that the subjects found both alternants acceptable in both circumstances, and there are no significant differences in the ratings. If anything, this data might suggest a slight preference for the infinitive in the success (performance) condition, exactly the opposite of what the potentiality/performance distinction would have predicted.

### 3.3.2 Aspectual Differences

The verb within the embedded clause could interact with alternant choice. In a similar vein to the performance and potentiality distinction, it could be the case that the VP in the embedded clause is inherently better-suited to an infinitive or to a gerund. We will use aspect as our window into such effects. Why might aspect matter? The aspect of a verb phrase determines the conceptualization of the action it describes:

- (38) a. I finished drawing a picture.  
 b. ?I finished drawing.

It is somewhat awkward to use the verb *finish* to refer to drawing, as drawing does not have any inherent endpoint. In this respect, the VP *drawing* is different from the VP *drawing a picture*; finishing drawing a picture makes sense because there is a endpoint at which the picture is completed. These two verb phrases are said to differ in aspect.

We use Vendler’s ((1957)) aspectual terminology in this paper. Vendler describes four aspectual categories for verb phrases: states, activities, accomplishments, and achievements. We will look at these categories in terms of two characteristics: telicity and durativity. A telic VP (38a) has an inherent endpoint, while an atelic VP (38b) does not. Durativity, on the other hand, is whether the action of a verb phrase is conceptualized as taking place over time:

- (39) a. I shopped for a new shirt for an hour.  
 b. ?I found a new shirt for an hour.

*Shopping for a shirt* takes time, but *finding a shirt* is (approximately) instantaneous. We refer to time-consuming VPs (39a) as durative, and instantaneous VPs (39b) as non-durative. We will discuss Vendler’s aspectual categories in more depth in Section 5.3. For now we will look only at durativity. Certain aspects sound awkward in the present progressive:

- (40) a. \*?John is loving Mary. [state]  
 b. John is flying. [activity]  
 c. John is climbing a mountain. [accomplishment]  
 d. \*?John is noticing the mistake. [achievement]

State and achievement VPs can not occur in the present progressive with their standard meanings<sup>17</sup>, though activity and accomplishment VPs can. This could lead to a categorical constraint on the *needs doing* alternation, because the gerund that occurs in the *needs doing* alternant is syncretic with the present progressive form. There are two reasons that this could prevent states and achievements from occurring in the *needs doing* form. One is that the syncretism of the gerund and the present progressive implies some amount of progressivity in the gerund that states and achievements cannot take. The other is that speakers’ avoidance of the present progressive form with certain verbs makes them generally avoid the *-ing* ending with those verbs. This could explain the awkwardness of sentences such as (41a,b).

- (41) a. (\*?) The form needs finding.  
 b. (\*?) The facts need facing.

Note that this would not affect the truth conditions, but could be a categorical semantic distinction that could confound our model. However, as it turns out, achievement verbs are regularly found in the *ing* form in the BNC:

- (42) a. [...] a purity which *needs finding*, indeed which must be searched for, struggled for, hunted down  
 [...]
 b. She didn’t *need prompting*.

The restriction against state verbs in the gerundive alternant is also non-categorical, although such occurrences are rare in the BNC sound a bit awkward to American ears:

- (43) a. Products *need believing* [in]  
 b. The question of the absolute right of a woman to bear a child *needs offsetting* by the right of a child to a balanced upbringing.

<sup>17</sup>Some achievement verbs can be progressivized if “they are understood as having a preparatory process” (Partee, 2007), but then the action of the verb phrase is the preparatory process, rather than the action itself. Certain state verbs can also occur in the present progressive, as in McDonald’s “I’m loving it” slogan. However, this is a sense of *love* that is more akin to *enjoy* than to the standard meaning of *love*.

However, as with idioms, state verbs are relatively rare in either alternant, so this is not entirely unexpected. Moving to a web search, we find a variety of grammatically reasonable uses of state verbs in the *ing* form:

- (44) a. face things that *need facing*<sup>18</sup>  
b. Old school anime like Urotsukidoji or Wicked City seemed to require a story, and characters who *needed believing* in, or at least some form of credibility.<sup>19</sup>

Clearly, although state verbs are rare in the *ing* form in the BNC, there is no categorical restriction against them. In fact, as we will see, durativity is not a significant factor in the final regression model for the alternation, although telicity is.

### 3.3.3 Murphy's Proposals

We round out the search for categorical semantic differences with evidence against the three distinctions proposed by Lynne Murphy (cited in Murray et al., 1996). Murphy herself acknowledged that these were unlikely to be borne out, and we include them because they are the only distinctions proposed specifically for this alternation in the literature.

**Possessor as agent in *to be* form** Murphy's first proposal is that if the *to be* alternant is used, and the subject of the sentence has a stated possessor, then the implied agent for the needed action is the possessor of the subject.

- (45) a. John's car needs to be washed  $\Rightarrow$  *John* will be the one to wash the car  
b. John's car needs washing  $\nRightarrow$  John will be the one to wash the car

Under this proposal, if one says that *John's car needs to be washed*, it is implied that it is John who needs to wash the car. If one instead says *John's car needs washing*, anyone could be the intended washer, or there could be no specific intended washer at all. However, the *to be* alternant can be used with a possessed subject without implying that the agent is the subject's possessor, as can be seen in examples from the BNC:

- (46) a. He said Prost would almost certainly be granted his super licence, but said *his behaviour in using "insulting terms" in his criticism of FISA would still need to be considered by the world council next month*.  
b. For example, the man may travel to work each day by car and so carry an extra \$20 with him just in case *his vehicle breaks down and needs to be repaired at once*.

In (46a), the sentence explicitly states that the world council, rather than Prost himself, will consider Prost's behavior. In (46b), the agent is unspecified, but it is strongly suggested that it is not the man himself — presumably the \$20 is to pay someone to fix his car, not for him to buy the necessary parts to fix it himself. Thus we see that there is no categorical restriction that the subject's possessor must be implied as the agent for the needed action in the *to be* alternant.

***ing* form implies benefit to subject** Murphy's second proposal is that if the *ing* alternant is used, then the subject must somehow benefit from the needed action. This accounts for the claimed awkwardness of sentences like (47).

- (47) (\*?)My old books need selling.

However, we have already give substantial evidence that inanimate objects can be subjects in this alternation, as part of the raising tests in Section 3.2.1. In fact, as we will see in Section 6.1, subject animacy does not even have a significant gradient effect on speaker choice.

<sup>18</sup><http://www.43things.com/things/view/324008/face-things-that-need-facing>

<sup>19</sup><http://www.digital-retribution.com/reviews/dvd/0776.php>

***ing* form requires a pre-existing subject** The last of Murphy’s proposals is that the *ing* alternant can only be used if the subject already exists. This would account for the awkwardness of sentences like (48a).

- (48) a. (\*?)My paper needs writing.  
b. (?)The lecture needs preparing.

Murphy herself suspects that this is non-categorical because (49b) is only slightly bad. In the BNC, this suspicion is borne out:

- (49) a. I do not believe that the current management at British Rail is capable of building the project, although I believe that *it needs building*.  
b. *The libretto now needs writing*, not in terms of expectation of aid from elsewhere — little will come — but expressing the room for the innate, self-help strengths of local people and communities.

### 3.4 Summary

We find that *needs to be done* ~ *needs doing* is indeed a syntactic alternation, at least in British English. Both constructions appear to be salient in the minds of British speakers, and they are truth-conditionally equivalent. Furthermore, we find that there are no apparent major low-level semantic differences lurking about. This clears the way to analyze the effects of other factors on speaker choice in the alternation. Notably, it opens the door to the Environment Prototypicality Hypothesis, which we discuss in the next section.

## 4 The Gerund as a Mixed Category

In the previous section, we found that the syntactic difference between the infinitive and gerundive forms did not lead to a noticeable semantic difference between *needs to be done* and *needs doing*. In this section, we return to the syntactic difference and explain how it could affect speaker choice nevertheless.

Syntactic alternations often involve a change in syntactic category, but usually this change is bound with a substantial change in word order and overall syntactic structure. The *needs doing* alternation involves a change of syntactic categories without substantial syntactic reorganization and thus presents a testing ground for the effect of category change. The verb in the two alternants is in different syntactic categories; the past participle of the *needs to be done* alternant is a verb with no nominal properties<sup>20</sup>, while the gerund of the *needs doing* alternant is a mixed category with both verbal and nominal properties (Malouf, 2000). The two constructions are otherwise syntactically similar; the word order is unchanged, the same arguments are the syntactic subjects, and the syntax of the rest of the sentence remains the same regardless of which alternant is used. The key difference is between the gerund and the participle.

Before we look for effects of the syntactic difference on speaker choice, we must establish the syntactic difference. The past participle has no nominal properties. But what is the nature of the gerund? This section details Malouf’s (2000) argument for the gerund being a mixed-category item.

Quirk et al., 1985 describe a continuum of nominalization strategies in English, each using a form that ends in *ing*. This continuum ranges from purely nominal forms (50a) to purely verbal forms (50c). Between the endpoints of this continuum lies the gerund (50b), which has some nominal features and some verbal features, but not the full set of either category’s features.

- (50) a. *Brown’s painting of his daughter* hangs in the town museum.  
b. *Brown’s deftly painting his daughter* is a delight to watch.  
c. Brown *is painting* his daughter.

---

<sup>20</sup>The past participle may also have some adjectival characteristics. It would be interesting to look for additional effects from such characteristics in subsequent work.

Malouf focuses on four properties of gerunds that place them between nouns and verbs:

- (51) i. Gerunds can govern NPs. [verbal]
- ii. Gerunds are modified by adverbs, not adjectives. [verbal]
- iii. A gerundive phrase has the same external distribution as an NP. [nominal]
- iv. The gerund can take an optional subject, which may be of genitive or accusative case. [mixed]

We illustrate these properties with a few examples. In (52a), the gerund *seeing* governs the direct object NP *Mike*, just as the verb *see* would — this is an example of property (51i). (52b) illustrates property (51ii), as an adverb is grammatical, but an adverb is ungrammatical.<sup>21</sup> (53a-c) show that a gerundive phrase, like an NP, can be the object of a preposition. This, along with further examples given in Malouf, demonstrates property (51iii), which states that gerundive phrases and NP have the same external distribution. Finally, gerunds can have a subject, and this subject can be either in genitive case (54a), as with a noun, or in accusative case (54b), as with a verb. Most importantly, a gerund can *simultaneously* exhibit both verbal and nominal characteristics; (52b) includes a gerund with a genitive subject (a nominal characteristic) and an adverbial modifier (a verbal characteristic).

- (52) a. Seeing Mike angered me.
- b. My quietly/\*quiet escaping the room went unnoticed.
- (53) a. I worried about Bob eating the berries.
- b. \*I worried about Bob ate the berries.
- c. I worried about Bob's digestive tract.
- (54) a. I want to hear his singing/songs/\*sing.
- b. I want to hear him singing/\*songs/sing.

Malouf analyzes previous explanations for the gerund's peculiar set of properties, but finds flaws in each. He argues instead that this mixed set of properties stem from the gerund's mixed category membership; it is simultaneously a verb and a noun.<sup>22</sup> With gerunds as a mixed category, all four of the important gerundive properties arise naturally.

That said, the nature of gerunds is still controversial. Some (e.g., Aarts, 2004) argue that gerunds are not mixed-category items, but rather category-ambiguous items. Under a category-ambiguous viewpoint, in any given situation, a gerund is either a noun or a verb; but it may be a more or less prototypical member of the category. Thus in a sentence like (52)b, Aarts argues that *escaping* is a verb, since it is modified by an adverb and can take an argument. Certainly it is not a prototypical verb, as it can take a genitive subject, but this does not make it any less of a member of the verb category. This view assumes that any given element can only be in one syntactic category at a time, although it may switch categories in different environments. We will sidestep this specific debate in the current paper, because it is not essential to our point. The fact that gerunds may behave like nouns — regardless of whether this behavior indicates membership in the noun syntactic category — is sufficient for our purposes.

Unlike the gerund, the past participle has no nominal characteristics. It cannot be modified by adjectives (55a), it cannot take a genitive subject (55b), and its external distribution is not at all like that of an NP (55c). The past participle may not be strictly verbal, as it can function as an adjective, but the necessary point for this study is that it is not in any sense noun-like.

- (55) a. The food ought to be quickly/\*quick eaten.
- b. \*The food was his eaten.
- c. \*I worried about eaten the berries.

Since gerunds have both nominal and verbal behavior, whereas the past participle has no nominal behavior,

---

<sup>21</sup>An adjective can be grammatical with an *ing* form, as in *Brown's deft painting of his daughter*. However, here *painting* has no verbal properties, as it requires a genitive subject and cannot govern *his daughter*. This is a purely nominal form, like (50a), and thus falls outside of the scope of the present discussion of mixed-category gerunds.

<sup>22</sup>More specifically, Malouf places the gerunds under both the noun and relational categories in the HPSG head hierarchy.

this could factor into speaker choice. Our specific hypothesis is that the *needs doing* alternant adds nominal characteristics to the alternation that *needs to be done* would not add. Because of these nominal characteristics, especially the NP-like external distribution, the gerund will fit better in an environment where one would expect to see a noun than the participle will.

#### 4.1 Environment Prototypicality Hypothesis

What do we mean by a “more nominal” environment? Consider sentences (56a,b).

- (56) a. \*The couch needs a to be cleaned  
b. The couch needs a cleaning.

*Clean* is in a prototypical place for a noun in each of these sentences. In both sentences, the verb follows a determiner. This is a highly prototypical place for a verb and a highly non-prototypical place for a noun; so highly non-prototypical, in fact, that it is grammatically unacceptable. Now consider sentences (57a,b).

- (57) a. The couch needs to be completely cleaned.  
b. \*?The couch needs completely cleaning.

Here the verb form in each sentence is a prototypical place for a verb rather than a noun. The verb forms follow an adverb here, which is a highly prototypical place for a verb, but a non-prototypical place for a noun. Since both forms have verbal properties, neither is grammatically unacceptable. The added nominal properties of the gerund, though, make the gerund substantially less acceptable in this context. The gerund is still attested in such contexts, though:

- (58) [...] there might be some dirt, dust or foreign matter that *needs completely cleaning* from the aperture  
[...]<sup>23</sup>

These acceptability differences are the basic idea behind the Environment Prototypicality Hypothesis, which states that environment prototypicality has a gradient effect on speaker choice. Environment prototypicality is the measure of how likely an environment is to take a given syntactic category. The environment prototypicality view can be viewed as a gradient extension of syntactic category agreement. For instance, nouns are modified by adjectives, and verbs by adverbs. This is a form of environment prototypicality; adverbs are prototypical fillers of the “modifying a verb” environment, and adjectives are prototypical for the “modifying a noun” environment. In fact, these two categories are so prototypical that there is a categorical constraint that nouns must be modified by adjectives and verbs must be modified by adverbs.

Environment prototypicality offers a gradient framework for such effects. The gradient framework has two immediate advantages. The first is that it allows a given context to exert differential preferences for categories. In the categorical view, either a category is permitted in a given context or it is not, and there is no way to indicate syntactic preferences, only requirements. The categorical view could not account for prepositional phrases that could occur modify either nouns or verbs, but have a preference for one over the other, which is the sort of syntactic category preferences we see in the alternation. (More will be said on this in Section 6.4.2.) The second advantage is that the gradient framework allows for different effect strengths depending on category membership values, and thus can readily account for Malouf’s mixed-category elements or Aarts’s category-ambiguous items. In either case, prototypical members of a category are more affected by environment prototypicality effects on that category than less prototypical members are.

In this alternation, the Environment Prototypicality Hypothesis predicts that the partially-nominal gerundive form will be preferred in environments that favor nouns, while the non-nominal past participle form will be preferred in environments that favor verbs (since the past participle is a more prototypical verb). The only remaining issue is how to quantify the prototypicality measure. In Section 6.4.2, we implement

<sup>23</sup><http://www.autoinsider.co.uk/problems/show.php?problem=670>

a method of estimating the prototypicality of an environment using the context following the alternation. Using this estimate, we find evidence for environment prototypicality effects.

## 5 Modelling the Alternation

Having established that the *needs doing* alternation is non-categorical, we want to determine what gradient factors speakers use to choose an alternant. To do this, we will model the alternation with a mixed-effects logistic regression model, trained on a set of corpus usages of the alternation. As mentioned earlier, logistic regression is an effective model for speaker choice in this alternation because it estimates the *probability* of the speaker choosing an alternant given the environment surrounding the alternation, rather than stating which alternant is preferred. Additionally, logistic regression can simultaneously control for multiple potentially correlated variables at once (Bresnan et al., 2007). This allows us to tease apart the effect of correlated variables, such as animacy and concreteness, or verb length and frequency. Also, the coefficients of a linear regression model have an intuitive interpretation as effects on the odds of choosing an alternant. We use a *mixed-effects* logistic regression so that we can investigate the effect of different verbs on speaker choice — the mixed-effects idea will receive further attention in Sections 5.1.4 and 6.3.

This section outlines the mechanics of linear models and generalized linear models, focusing on the special case of logistic regression. Building from this, we then discuss the mixed-effects version of logistic regression, where some factors are assumed to have randomly-distributed idiosyncratic baseline preferences. This is the model we will use to investigate the factors affecting speaker choice in the *needs doing* alternation. With the mechanics of the model in place, we will discuss the dataset that it is trained on. We conclude with an explanation of the set of factors that the regression considers in modelling speaker choice.

### 5.1 Mixed-Effects Logistic Regression

#### 5.1.1 Linear Models

Logistic regression is a *generalized linear model*, which, as the term suggests, is a generalization of the standard linear model. Standard linear models determine the conditional probability  $P(Y|X)$  of an outcome  $Y$  given input data  $X = (X_1, \dots, X_n)$  from the assumption that the outcomes are generated probabilistically from an equation

$$Y = \beta_0 + \sum_i \beta_i X_i + \text{noise} \quad (1)$$

where  $\{\beta_i\}$  are real-valued weights, and the noise term is a normally distributed random variable with mean 0 and a constant standard deviation  $\sigma^2$  (i.e., Gaussian noise). From the probability density function for the Gaussian distribution, the conditional probability  $P(Y|X)$  is determined by  $\sigma^2$  and the distance between  $Y$  and the linear predictor  $\eta(X) = \beta_0 + \sum_i \beta_i X_i$ , by Eqn. 2.

$$P(Y|X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(Y-\eta(X))^2}{2\sigma^2}} \quad (2)$$

The conditional probability in this equation defines a Gaussian centered at the value of the linear predictor  $\eta(X)$ , with standard deviation  $\sigma^2$ . Thus conditional probability is highest when  $Y = \eta(X)$  and drops off monotonically as  $Y$  moves away from the linear predictor. Intuitively, such models are well-suited for tasks like predicting the travel time for a given trip. There is an average trip time that depends on factors like time of day, means of transportation, weather conditions, and so on. However, the trip time is not completely deterministic – two different cars making the same trip under the same conditions can end up with significantly different travel times, due to random occurrences along the trip (such as another driver cutting one car off but not the other). The deterministic factors then specify a predicted trip time, and the noise term accounts for these random deviations from it. Most trips will be close to the average time, but occasionally things will go especially well or especially poorly and the trip time will be much higher or lower.

### 5.1.2 Logistic Regression as a GLM

A linear model can be thought of as having two main components, the linear predictor  $\eta(X) = \beta_0 + \sum_i \beta_i X_i$  and a noise distribution centered around the linear predictor. In a standard linear model, the mean  $\mu(X)$  of the predicted distribution is equal to the linear predictor  $\eta(X)$ . Logistic regression uses the logit link function to relate  $\mu(X)$  and  $\eta(X)$ , so the probability of “success” is equal to the exponential of the linear predictor (Agresti, 2002):

$$p(\text{alternant}|X) = \mu(X) = \frac{e^{\eta(X)}}{1 + e^{\eta(X)}} = \frac{e^{\beta_0 + \sum_i \beta_i X_i}}{1 + e^{\beta_0 + \sum_i \beta_i X_i}} \quad (3)$$

“Success” is a somewhat inappropriate term for logistic regression in alternations. For alternations, the two alternants are labelled “class 0” and “class 1”, and the probability of success is really the probability of choosing whichever alternant is labelled “class 1”. The choice of which alternant is labelled as which class is completely arbitrary. In the present alternation, *needs to be done* was chosen to be the “class 1” alternant, so success is the same as choosing the *to be* alternant.<sup>24</sup>

The underlying idea for logistic regression is that an example  $X$  has some probability  $P(Y = y_1|X)$  of being in class  $y_1$ . Whether  $X$  ends up in class  $y_1$  is essentially a weighted coin-flip, where the probability of heads is  $P(Y = y_1|X)$ . This fits with the probabilistic framework we have assumed for speaker choice; the speaker has some probability of choosing an alternant, and the probability is determined by the context of the rest of the sentence. In our framework, the classes for  $Y$  are the two alternants, and  $X$  is the vector of values from the factors based on the rest of the sentence.

### 5.1.3 Properties of Logistic Regression

A major advantage of using logistic regression — one that has propelled its adoption in linguistic research — is that the  $\beta_i$  coefficients have a convenient interpretation as effects on the odds of an outcome. The odds of an outcome is the proportion of outcomes in each class; odds of 2:1 in favor of heads, for instance, means that there are expected to be twice as many heads as tails. In binary classifications, the odds of success are given by  $\frac{p}{1-p}$ , where  $p$  is the probability of success.<sup>25</sup> Logistic regression uses Eqn. 3 to determine the probability of success (the *needs to be done* alternant), and plugging this equation into the odds calculation reveals that the odds are equal to the exponential of the linear predictor.

$$\frac{p}{1-p} = e^{\beta_0 + \sum_i \beta_i X_i} = e^{\beta_0} \prod_i e^{\beta_i X_i} \quad (4)$$

The odds are the product of the exponentials of each term in the linear predictor. This gives the intuitive meaning behind the coefficients for the linear predictor: their exponentials  $e^{\beta_i}$  are multiplicative effects on the odds, and the coefficients  $\beta_i$  are additive effects on the log-odds. Thus, for continuous factors, incrementing  $X_j$  by one multiplies the odds of an alternant by  $e^{\beta_j}$ . For categorical factors, switching  $X_j$  from off to on multiplies the odds of an alternant by  $e^{\beta_j}$ . If a  $\beta$ -coefficient is positive, this makes success more likely; if it is negative, it makes failure more likely.

### 5.1.4 Mixed Effects Models

With logistic regression established, we can move on to a more complicated situation. The models we have considered assume that the observations are independent of each other given the predictor variables. However, in many situations, there are underlying variables that group certain examples together. One situation in which this could occur is in a reaction-time experiment being repeated with many different subjects. Each subject may have a mean reaction time that is faster or slower than the group average. Mixed-effects models allow us to account for this variation by assuming that examples are clustered (e.g., by speaker or by item),

<sup>24</sup>Note that this means that anytime the *ing* alternant is preferred in a given sentence, the “success” alternant will actually be a worse choice than the “failure” alternant.

<sup>25</sup>The odds in the coin example, for instance, are calculated as follows:  $p(\text{heads}) = \frac{2}{3}$ , so the odds of heads is  $\frac{2/3}{1-2/3} = \frac{2/3}{1/3} = \frac{2}{1}$ .

and each cluster can determine its own values for  $\{\beta_i\}$ . This gives the model more predictive power, because variation can be explained both at the cluster level and the observation level.

Mixed-effects models get their name from the two types of factors they consider. The first type are the fixed effects, which are the components of the linear predictor in a regular logistic regression. These are deterministic. The second type of effect are the random effects, which may be based on the same or different factors from those of the fixed effects. However, whereas the fixed effects have set coefficients  $\{\beta_i\}$ , the random effects have a stochastic vector  $b = (b_0, b_1, \dots, b_m)$  that follows some joint probability distribution. In our model,  $b$  follows a multivariate normal distribution with covariance matrix  $\Sigma$ . The  $b$  vector can have two different effects; it can create different intercepts for different clusters, or different slopes for existing factors. The two effects combine to give the linear predictor for a mixed-effects logistic regression:

$$\eta(X) = \beta_0 + \sum_i \beta_i X_i + b_0 + \sum_j b_j Z_j \quad (5)$$

The random effects’ coefficients  $b$  are assumed to come from a multivariate normal distribution over the random factors  $\{Z_j\}$ . In this study, there is a single random effect of the verb in the alternation (e.g., *do* in *needs doing*), with one  $Z_j$  for each verb observed in the dataset. Therefore, our model assumes that each verb can set an idiosyncratic intercept, with odds favoring one alternant over the other. The parameters of the mixed-effects logistic regression can then be fit using a maximum likelihood method. In maximum likelihood parameter estimation, an algorithm is used to search for parameters that maximize the likelihood of the logistic regression model generating the observed training data. In the present study, we are less interested in the specific values of  $b$  (the random effect of verbs) than in the specific values of  $\beta$  (the fixed effects of sentence factors). Thus we can marginalize the likelihood over all possible values for  $b$  and estimate  $\Sigma$ , the covariance matrix, to determine the distribution of the random effects. This yields a likelihood function of:

$$L(\beta, \Sigma; X) = \int_{-\infty}^{+\infty} P(x|b, \beta, \Sigma)P(b|\Sigma)db \quad (6)$$

Training the logistic regression consists of finding  $\{\beta_i\}$  and  $\Sigma$  to maximize this likelihood then gives us the parameters for the model.  $\Sigma$  allows us to estimate the distribution of the random effects, to see how much of the variation is explained by them. Ideally, we would like to know both the distribution of the random effects and estimates for the specific values. Unfortunately, even this likelihood, which only gives us the covariance matrix, cannot be evaluated exactly. Instead, it is estimated by Laplacian approximation. Estimating the values of  $b$  at the same time is a computationally expensive task, so we will limit ourselves to estimating the distribution parameters in the maximum likelihood step. Once the model has been fit, we can then estimate the specific values of  $\{b_j\}$  from the best linear unbiased predictors (BLUPs) for the random effects.

## 5.2 Dataset

The regression model is trained on a dataset of 1004 sentences from the BNC. This training set is a subset of 5926 sentences from the BNC that were found by a **tgrep** search over a parsed version of the corpus (Doug Roland, p.c.). The **tgrep** search identified sentences containing a form of the word *need* either followed by a gerund or by the words *to be* and a past participle or past tense verb. 4587 of these sentences used the *to be* alternant, 1336 used the *ing* alternant, and three used both. This set contains a number of false positives from a variety of causes, such as non-verbal uses of *need* (59a), intervening constituent boundaries, and other parts of speech being mislabelled as verbs (59b). Details of the **tgrep** searches and the issues of false positives are detailed in the Appendix.

- (59) a. The prison managers say if inmates use their own initiative there’s no *need to be bored* and cause trouble.  
 b. [...] sophisticated administration systems are *needed[,] exercising* controls in five main areas.

Before being included in the model dataset, each sentence was manually annotated for the relevant features (animacy, concreteness, etc.). During this annotation process, false positives from the **tgrep** search were

removed from the dataset. Due to the time demands of annotation, the whole dataset could not be included in the model. Instead, a randomly retrospectively sampled (Agresti, 2002) dataset of 1004 sentences was used. Thus, unlike the full search set, in which 77% of the sentences used the *to be* alternant, the training set contained equal numbers of each alternant. Retrospective sampling was used to ensure that, despite the smaller size of the training set, there would still be enough instances of the rarer *ing* alternant to draw statistically significant conclusions about the relevant factors.<sup>26</sup>

### 5.3 Factors Considered

We include 20 factors in the regression model. These factors can be generally divided into four categories: subject variables, intra-construction variables, post-construction variables, and sentence-level variables. We consider each set in turn.

#### 5.3.1 Subject Variables

Six factors derived from the subject are considered: animacy, concreteness, pronominality, definiteness, subject length, and presence of a relative clause. (Recall that the syntactic subject for the alternation is the semantic object of the verb.)

**Animacy and Concreteness** Animacy and concreteness are based on the animacy encoding system proposed by Zaenen et al. (2004). However, their encoding system contains a number of categories (such as intelligent machines) that were unattested or only rarely attested in the dataset. Thus the NPs in the alternation dataset were collapsed into three categories: animate concrete, inanimate concrete, and inanimate abstract. Subjects that were Human, Animal, or Org under Zaenen et al’s system were marked as animate concrete, subjects that were Concrete or Place were marked as inanimate concrete, and other subjects were marked as inanimate nonconcrete. Examples from each category are given below:

animate concrete, [+animate, +concrete]: *Paul Westerberg, a baboon, his sister*  
inanimate concrete, [-animate, +concrete]: *a factory, beef stew, ancient wiring*  
inanimate abstract, [-animate, -concrete]: *bad ideas, a suggestion, thirty minutes*

By reducing these variables to two binary features, the issue of sparsely populated categories (like Org) is avoided, and the animacy encoding can be represented as two binary variables:  $[\pm\text{animate}]$  and  $[\pm\text{concrete}]$ <sup>27</sup>. We encode animacy and concreteness as these two separate binary variables to isolate the effect of each on speaker choice.

**Definiteness** Definiteness is a ternary categorical feature. Pronouns, names, possessive NPs, and NPs with a definite or discourse-linked marker (e.g., *this*, *such*) are marked as definite. NPs with an indefinite determiner and DPs without determiners are marked as indefinite. Quantified expressions and free relatives are marked as neither definite nor indefinite. Givenness is not included in this set, despite Bresnan et al’s (2007) finding of a significant givenness effect on the dative alternation, because it can only properly be determined given context, and this would require more time to annotate than was available. However, definiteness is (weakly) correlated with givenness (Zaenen et al., 2007, a.o.), so the effect of definiteness may be indicative of an effect of givenness as well.

---

<sup>26</sup>Retrospective sampling affect the magnitude of the random effects in a mixed-effects model, generally resulting in somewhat inflated effects (Neuhaus & Jewell, 1990). Ideally, the results in Section 6.3 should be verified on a non-retrospectively sampled dataset, to assure that the values are accurate. However, the random effect strengths are only estimates of the true effects, and calculating them more precisely would require a substantially larger set of annotated sentences.

<sup>27</sup>No animate abstract objects were attested in the dataset, and it seems impossible for such an object to exist. Thus animates are a subset of concrete objects, and abstract objects are a subset of inanimates.

**Relativization** Presence of a relative clause is a binary feature. This is marked only if the alternation is used in a relative clause modifying the subject (60a), not if the alternation outside the relative clause (60b).

- (60) a. [The house that *needs repainting*] is at the end of the block.  
b. [The house that I live in] *needs repainting*.

**Subject Length** Subject length is a quantitative variable. Subject length is given to the model as a smoothed log of the word count in the subject, so subject length =  $\log(\# \text{ of words in subject} + 0.5)$ . The smoothed log is used to reduce the effect of outliers, such as one sentence that had a 23-word subject.

### 5.3.2 Intra-Construction Variables

The alternation itself adds another six variables to the model, mostly related to the verb in the construction: tense, presence of a verb particle, verb length, verb frequency, verb aspect, and the verb itself.

**Tense** Tense is fairly straightforward. It is a four-way factor: past, present, future, or other. The “other” category is populated predominantly by instances of infinitival *to need doing/to need to be done* alternants. Tense is an especially interesting variable, both because different tenses could directly influence the choice of construction and because the past tense could induce a syntactic Obligatory Contour Principle (OCP) effect (Mann & MacWhinney, 1984). The potential OCP effect arises because using the past tense of the alternation often involves two consecutive verbs with *-ed* suffixes (61).

- (61) The lawnmower **needed** to be refueled.

If the OCP matters in this alternation, it is conceivable that the proximity of the *-eds* will lead to a dispreference for this form. On the other hand, it is also possible that speakers will prefer the *to be* alternant in the past tense because of perceived tense agreement in this alternant.

**Verb Particle** Verb particle is included as a binary variable that notes whether the verb in the construction has a particle (62a) or not (62b). Unclear situations (62c,d) were resolved by determining if the action of the verb with the preposition implies the action of the verb without the preposition, following Zaenen et al., 2007.

- (62) a. The toys needed to be [picked up] off the floor.  
b. The dresser needs to be carried [up the stairs].  
c. The trash needs to be thrown [out the window].  $\Rightarrow$  The trash needs to be thrown.  
d. The trash needs to be [thrown out].  $\nRightarrow$  The trash needs to be thrown.

**Verb Length** Verb length is measured in terms of the number of syllables in the uninflected form of the verb (e.g., *do* rather than *done* or *doing*). The number of syllables in each verb was determined using a Python script that scanned the CMU Pronunciation Dictionary for each verb. The syllable length of the verbs were not log-transformed or smoothed, since they were all between 1 and 5 syllables.

**Verb Frequency** Verb frequency was also included as a factor. These frequencies are based on the lemma (uninflected verbform) frequencies listed in the CELEX database for these verbs. A verb that was not in the CELEX database was given a frequency of zero. However, verb frequency is given to the model as a smoothed log measure; verb frequency =  $\log(\# \text{ of occurrences in CELEX} + 0.5)$ . Thus a verb that was not in the database is listed as having a small non-zero frequency. This follows observations in the processing literature that show log frequency to be a better predictor of processing time than frequency itself (Smith & Levy, 2008, a.o.). As with the length smoothing, this it also reduces the effects of outliers, extremely common or uncommon verbs.

**Verb Aspect** Aspect, as discussed in Section 3.3.2, has four categories: states, activities, accomplishments, and achievements. These can be represented with two binary features:  $[\pm\text{telic}]$  and  $[\pm\text{durative}]$ . Telicity refers to the presence of an intrinsic endpoint to the action of the verb, so *run* or *draw* would be atelic, as they have no inherent endpoints. *Run a mile* or *draw a circle* would be telic, because these have inherent endpoints. Durativity refers to the action of the verb being perceived as covering a block of time; thus *hitting a baseball* would not be durative, but *rounding the bases* would. Vendler’s four aspectual categories can be represented with these two binary features: states are  $[-\text{telic}, -\text{durative}]$ , activities are  $[-\text{telic}, +\text{durative}]$ , accomplishments are  $[\text{+telic}, +\text{durative}]$ , and achievements are  $[\text{+telic}, -\text{durative}]$ . Note that durativity matches up with the ability to use the progressive form of a verb; states and achievements are  $[-\text{durative}]$  and generally unable to occur in the progressive. Aspect of each verb in the dataset was determined using the tests in Partee, 2007, and telicity and durativity values assigned accordingly. When a verb did not clearly fit into one of the categories, it was assigned telicity or durativity values if possible. The regression model considered effects of telicity and durativity separately; this is the same treatment that was used with animacy and concreteness.

**Random Effect of Verb** Finally, the verb’s identity is a factor in the regression model. Verbs are treated as a random effect, so the model chooses an intercept (an idiosyncratic preference for one alternant) for each attested verb and uses this intercept every time it sees the verb. Note that it is not redundant to include information about the verb (such as frequency) in the model when the verb itself is a random factor. The idea behind including the verb as a random factor is that two verbs with the same basic statistics (frequency, length) may still have different idiosyncratic preferences in the alternation. If verb frequency and length were not in the model, the idiosyncratic preferences could actually be drowned out by frequency or length effects.

### 5.3.3 Post-Verbal Dependents and Ambiguously-Attached Phrases

There are two post-construction variables, both based on the length of phrases that follow the alternant. Constituents after the alternant can be grouped into three general categories: post-verbal dependents, ambiguously-attached phrases, and syntactically separate constituents. An example each type is given in (63).

- (63) Barrels need to be lopped [<sub>1</sub>from the oil cartel’s production] [<sub>2</sub>from time to time], [<sub>3</sub>but no member will do so without prodding].

Post-verbal dependents, such as constituent 1 in (63), are constituents that unambiguously modify the alternation. Such modifiers are arguments or adjuncts of the verb in the alternation and generally cannot be felicitously fronted because they do not modify the sentence as a whole. Ambiguously-attached phrases, such as constituent 2 in (63), could modify either the verb in the alternation or the sentence as a whole. Unlike post-verbal dependents, these generally can be felicitously fronted. Ambiguously-attached phrases often refer to time, but they also can fill other adverbial roles. The last category, syntactically separate constituents, refers to constituents that are clearly unconnected to the alternation. These are often sentence-level conjuncts, like constituent 3 in (63), or predicates of sentences with the alternation in a subject relative clause (as in (60)).

Only post-verbal dependents and ambiguously-attached phrases are included in the model, since they are the only ones that can modify the alternation. We are interested in possible weight effects emerging from these phrases, so we need a way of quantifying the weight of the post-verbal dependents and ambiguously-attached phrases. Length in words is used to quantify the weight effects for two reasons: it is simple to get an accurate count of the number of words in a phrase, and Szmrecsányi (2004) has shown that it is tightly correlated with other common measures of weight. The smoothed log length of the post-verbal dependents (in words) and the smoothed log length of the ambiguously-attached phrases (in words) are given separately to the model. If a sentence has more than one post-verbal dependents, their lengths are summed before the log-transform is applied. The same is done if there is more than one ambiguously attached phrase.

Table 1: Distribution of alternants in the BNC by modality.

Alternant	Spoken	Written	8.2% of alternants spoken, 92.8% written
<i>needs doing</i>	214	1125	Spoken: 44.1% <i>ing</i>
<i>needs to be done</i>	271	4319	Written: 20.7% <i>ing</i>

### 5.3.4 Modifying and Intervening Variables

Four additional variables are considered to investigate the effect of modifying the alternant and separating the alternant from its syntactic subject. Each is a presence/absence binary variable: presence/absence of modals, of negation, of intervening adverbs, and of intervening conjoined material. These are only triggered if the appropriate object occurs within a specified part of the sentence.

**Modals, Negations, and Adverbs** Modals and negations are noted only if they modify the instantiation of *need* in the alternant. An adverb counts as “intervening” only if it is between the syntactic subject of the alternant and the alternant itself. Thus (64a) would be considered to have an adverb, while (64b) would not.

- (64) a. The couch badly needs to be cleaned.  
 b. The couch needs to be cleaned badly.

**Conjoined Material** Intervening conjoined material refers to VPs that occur between the subject and the alternant and that share their subject with the alternant. Thus (65a) would be said to have intervening conjoined material, but (65b) would not.

- (65) a. The children [behaved unacceptably and need punishing].  
 b. [The children behaved unacceptably] and [their parents need punishing].

These intervening variables are included to determine if physically separating the subject from the alternant or modifying the action of the alternant affects speaker choice.

### 5.3.5 Sentence Modality

Lastly, we consider the modality of the sentence. The British National Corpus contains snippets of English from a variety of sources, varying in tone, purpose, and style. Part of this variety entails the inclusion of both spoken and written English. Spoken and written language are obviously quite different in their construction; whereas the written word may be revised over the course of time, and reworked to please the ear of the author, spoken language has no such revisability. Certain factors that we expect to have an effect in one modality may not be expected to have the same influence in the other modality. For instance, one might expect processing effects to be less noticeable in written language than spoken language because there is not the same time-pressure in writing as in speaking. Bresnan et al. (2007) showed that modality and genre did not affect the accuracy of their model for speaker choice in the dative alternation; a regression model trained on the Switchboard corpus performed approximately equally well when tested on the Penn Treebank’s Wall Street Journal corpus. Although this suggests that speaker choice in the dative alternation is driven by the same factors in the same amounts, regardless of modality, this is not necessarily true of other syntactic alternations.

A priori, there is some reason to suppose that modality may matter for the *needs doing* alternation. Table 1 shows the number of instances of each alternant in each modality. Although the alternation is used about as often (per word) in the two modalities, when the alternation is used in speech, it is more likely to take the *needs doing* form than when it is used in writing. It is unclear from this fact whether this preference is inherent to the different modalities or if it is a result of the different modalities having different distributions for the factors that influence speaker choice; this is the primary reason for including modality as a variable

in these experiments. However, as will be shown in the experiments, the *needs doing* alternation does not seem to be affected directly by modality.

### 5.3.6 Complete List of Variables

In summary, the full model uses the following variables:

Categorical variables		Continuous variables
animacy	modality	subject length
concreteness	telicity (aspect)	log verb length
definiteness	durativity (aspect)	log PVD length
pronominality	verb particle	log AAP length
relativization	negation	verb frequency
tense	modal	<b>Random effect</b>
adverb	conjoined material	

These variables define a 20-dimensional vector that the regression model uses to predict what construction is used in a sentence.

## 6 Results and Discussion

In this section, we discuss the regression’s model of speaker choice in the *needs doing* alternation and what it tells us about the four open questions from Section 2.4. We begin by determining which factors are useful in the model, which shows the factors that have significant effects on speaker choice, and assessing the model’s effectiveness given those factors. We then examine the nature of the random effect of the verb and show evidence for the Environment Prototypicality Hypothesis.

### 6.1 Feature Selection

Once a model has been generated and the values of  $\beta$  estimated, the question becomes which factors are relevant for classification. There are two points to consider: the first is the confidence that a factor’s effect is not spurious, and the second is whether the added predictive power of a more complicated model justifies the extra degrees of freedom from the additional factors.

Confidence in an effect’s validity is determined by comparing the estimated effect strength (i.e., the coefficient for that factor) against the null hypothesis that the coefficient is zero. To do this, we calculate the Wald statistic, the standardized deviation of the MLE estimate of the coefficient.

$$\frac{\beta_i - \hat{\beta}_i}{\text{StdErr}(\beta_i)} \sim \mathcal{N}(0, 1) \text{ [approx.]} \quad (7)$$

The Wald Statistic is approximately normally distributed with mean 0 and standard deviation 1, as long as the number of observations is large compared to the number of parameters (Agresti, 2002). Thus we can derive  $p$ -values for each  $\beta_i$  against the null hypothesis that the factor has no effect (in which case  $\beta_i$  would be zero.)

The other question a given model raises is whether the additional factors in a model make it better overall than a simpler model. This can be done with likelihood ratio tests. Likelihood is calculated for a logistic regression model by Eqn. 6. Having calculated the likelihoods for two models  $M_1$  and  $M_0$ , the log-likelihood ratio is the log of the quotient of likelihood for  $M_1$  divided by likelihood for  $M_0$ . Two times the log-likelihood ratio is called the deviance, and this value follows a chi-squared distribution  $k$  degrees of freedom, where  $k$  is the difference in the number of parameters in  $M_1$  and  $M_0$ .

$$2 \log \frac{\text{Lik}_{M_1}(X)}{\text{Lik}_{M_0}(X)} \sim \chi_k^2 \text{ [approx.]} \quad (8)$$

Table 2: Significance values for factors in the regression. Log-likelihood of the full model was -504.8; log-likelihood in the one-held-out models ranged from -504.8 to -538.7. Likelihoods further away from zero are worse, so a lower likelihood in this table indicates a more important factor. Significances were calculated using Equation 8; different factors have different degrees of freedom, so the same likelihood improvement could be significant for one factor but not for another.

Factor	Likelihood of Model w/o Factor	$\chi^2$ Significance
Animacy	-505.8	.14
Concreteness	-520.4	$2.1 \times 10^{-8***}$
Definiteness	-512.0	$7.2 \times 10^{-4***}$
Pronominality	-505.9	.12
Relative Clause	-510.8	$4.8 \times 10^{-4***}$
Tense	-509.8	.021*
Modality	-506.3	.079
Telicity	-510.7	.0027**
Durativity	-506.7	.14
Verb Particle	-507.8	.013*
Negation	-506.5	.06
Modal	-505.6	.18
Adverb	-504.8	.63
Conjoined Material	-505.2	.31
Verb Length	-513.1	$4.3 \times 10^{-5***}$
Subject Length	-504.9	.63
PVD Length	-538.7	$1.8 \times 10^{-16***}$
AAP Length	-519.0	$9.6 \times 10^{-8***}$
Verb Frequency	-513.3	$3.5 \times 10^{-5***}$

Because the deviance follows a chi-squared distribution, the deviance can be used to compare the simpler model against the more complicated model, to determine if the increased likelihood of the more complicated model is sufficient to justify the additional degrees of freedom in this model (Pinheiro & Bates, 2000). This test can be performed so long as both models share the same random effects and one model’s factors are a subset of the other’s. In presenting significance results in this section, we use significance values derived from the likelihood-ratio test instead of the values coming from the Wald statistic, because knowing that the increased predictive power of a factor is worth its extra degree(s) of freedom is what matters.

To determine the significant factors in the regression, we create a regression model  $M_0$  with all of the factors included, and calculate its likelihood,  $\text{Lik}(M_0)$ , over the training set. We then create models  $\{M_1, \dots, M_{19}\}$  each of which has one of the 19 factors listed in Section 5.3.6 held out. (So  $M_1$  would be trained without animacy,  $M_2$  without concreteness, and so on.) By comparing the likelihoods of each of these 18-factor models to the full 19-factor model using the likelihood-ratio test, we can determine significance values for each of the factors.

Table 2 shows significance values from the likelihood-ratio test in Equation 8 for each variable in the model. This shows that 10 factors are significant in the regression: concreteness, definiteness, relativization, tense, telicity, verb particle, verb length, verb frequency, post-verbal dependent length, and ambiguously attached phrase length. Two additional factors are close to significance: negation and modality. Post-verbal dependent length has the greatest effect on the model’s likelihood; in some sense this factor has the biggest effect on modelling speaker choice. We will investigate this as evidence for the Environment Prototypicality Hypothesis in Section 4.1.

## 6.2 Overall Results

Model accuracy can be measured in two ways. One forces the model to choose an alternant for each sentence and compares that to the attested alternant, and the other compares the model’s probability estimates to observed usage probabilities.

In both cases, the accuracy measurements are calculated using a 5-way cross-validation test. In this test, the dataset is divided into 5 subsets. Five separate regression models are trained, each using four of these subsets to train and testing on the fifth. Because the sentences in the test set are distinct from those in the training set, this test gives an honest estimate of how the final model proposed in this paper will perform on new sentences. If the test sentences were in the training sentences, the accuracy measures could be inflated due to overfitting. Different partitions of the data into the five subsets may result in different accuracy numbers, so all accuracy values in this section are given as the mean and standard deviation from 10 trial runs.

A basic measure of model accuracy can be obtained by forcing the model to choose an alternant for a sentence. This is essentially the task facing a speaker using this alternation; given the intended sentence  $S$ , one alternant must be chosen. We mimic this situation by taking the probability estimate  $P(\text{to be}|S)$  from the model and choosing the *to be* form if  $P(\text{to be}|S) \geq \frac{1}{2}$  and choosing the *ing* form if  $P(\text{to be}|S) < \frac{1}{2}$ . The model is said to succeed if it chooses the same alternant as the speaker, and to fail if it chooses the other alternant. It is important, therefore, to keep in mind that when the model is “wrong”, it is not necessarily choosing an alternant that creates an awkward sentence. In fact, the model can make the wrong choice even if it chooses the alternant that sounds better in a sentence, if the speaker chose the other form. Under this evaluation metric, the model averages  $74.8 \pm .6\%$  accuracy on the test sets. This is a significant improvement over the baseline accuracy, which is .5 since each alternant is equally likely in the retrospectively-sampled dataset.

This evaluation metric, though, collapses the probabilistic prediction from the model to a binary classification. The logistic regression model is not designed to directly guess the form the speaker has chosen; instead, it is estimating the probability of each alternant for each sentence. By forcing the model to choose an alternant, we are losing information and underestimating the model’s accuracy, assuming that speaker choice really is non-deterministic.

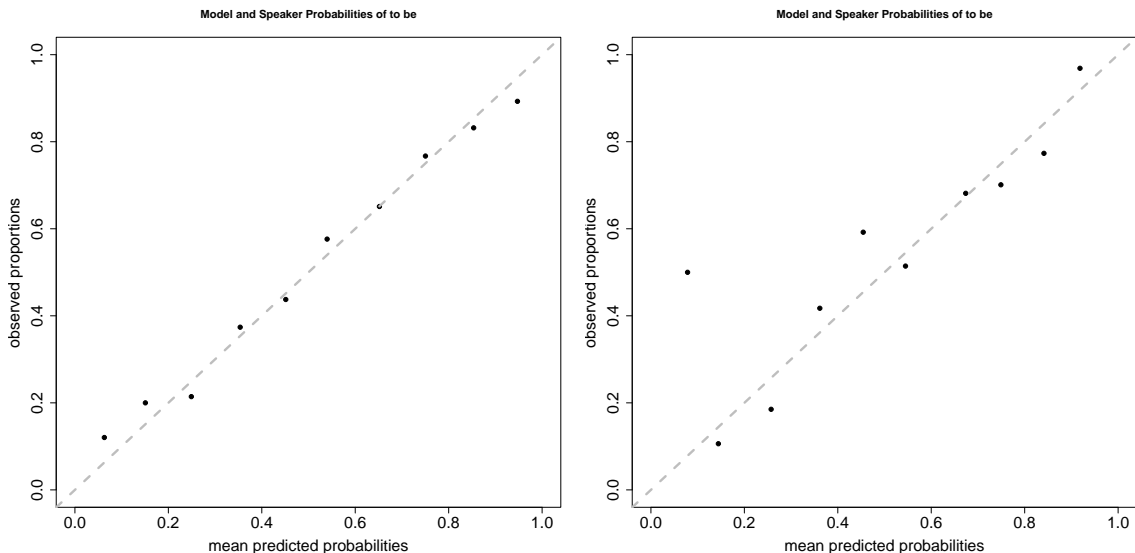
To illustrate this point: suppose that for a given sentence  $S$ , a speaker has a 65% probability of choosing the *to be* form. If the regression model is a perfect estimator for speaker choice in this sentence, it will estimate  $P(\text{to be}|S) = .65$ . Under the accuracy evaluation metric, the model will always choose the *to be* form, and will only be correct 65% of the time, even though it is a perfect estimator of speaker choice in sentence  $S$ . Forcing the binary choice can thus make a model look worse than it is.

A better measure of the model’s similarity to non-deterministic speaker choice is to examine the actual proportions of each alternant for sentences with similar model probabilities. For instance, if the model is a perfect estimator of speaker choice, and there is a set of one hundred sentences  $\mathbf{S} = \{S_1, \dots, S_{100}\}$  for which  $P(\text{to be}|S_i) \approx .65$  for all  $S_i \in \mathbf{S}$ , then we expect that approximately 65 of these sentences will use the *to be* alternant. If, on the other hand, only 20 of these sentences used the *to be* alternant, it would be unlikely that the model is a good estimator for speaker choice.

To formalize this measure, we follow the method used by Bresnan et al. (2007). Again, the dataset is split into five subsets, with predictions for the sentences in one subset coming from a model trained on the other four subsets. Each sentence  $S$  is then assigned a probability  $P(\text{to be}|S)$  from the appropriate model. Once every sentence in the dataset has a probability, the sentences are ranked by their values for  $P(\text{to be}|S)$ . The top 10% are put into one bin, the second 10% into another bin, and so on, resulting in 10 bins of decreasing *to be* probability. The model probability of *to be* for each bin is then taken to be the mean value of  $P(\text{to be}|S)$  over all sentences  $S$  in the bin. The speaker probability of *to be* for each bin is taken to be the number of sentences in the bin that use the *to be* alternant divided by the total number of sentences in the bin. The correlation between the speaker probability estimates and the model probability estimates can then be used to quantify the similarity between the two probabilities.

Also, this allows for a convenient visualization of the relationship between the two probabilities, as shown in Figure 1. The x-axis represents the mean probability of *to be* for sentences in a bin, as estimated by

Figure 1: Correlation between model and speaker probabilities in a sample trial. The x-axis gives the probabilities predicted by the model, and the y-axis gives the probabilities observed in the dataset. The left graph is for a model with all significant fixed factors, and the right graph is for a model with only a single fixed factor, concreteness. Both models include the random effect of verb as well.



the model’s values for  $P(\text{to be}|S)$ ; the y-axis represents the proportion of sentences in a bin that actually involved the *to be* form. Points above the dotted grey line represent bins where the model underestimates the proportion of the *to be* alternant, and points below the line represent bins where the model overestimates the *to be* alternant. The model run with all significant factors is shown in the left-hand graph, and a version of the model with a single factor (concreteness) is shown on the right. If the model were a perfect estimator for speaker choice, then all the points would sit on the dotted grey line and the correlation would be 1. This is not quite the case, but the correlation is surprisingly high, and the points sit close to the grey line. The mean correlation over ten trials is  $R^2 = .994 \pm .004$ , showing that the model probabilities very tightly fit the observed proportions, even on unseen data,

This can be compared to a model that uses only a single factor to determine speaker choice; we use concreteness in this comparison because its effect was both strong and highly significant. For the single-factor model, the points are a much worse fit for the line (Figure 1) and correlation drops to  $0.91 \pm .14$ . This limited model is not bad at estimating speaker choice, as concreteness was one of the most significant effects in the model. However, it is clear that the additional factors in the regression have substantial explanatory power in speaker choice, supporting the log-likelihood significance measures from Table 2.

The *needs doing* model’s accuracy is much lower than the accuracy of Bresnan et al’s model of the dative alternation, which performed with 95% accuracy. There are a few reasons for this. The first is due to the difference in the composition of the dataset. The *needs doing* dataset is retrospectively sampled, and thus the baseline for the current model is 50% accuracy. This means that a naive classifier that always chooses the same alternant for every sentence will only make the correct choice half of the time on this dataset. Bresnan et al’s dative dataset, though, is not retrospectively sampled. 79% of the sentences in their dataset are examples of the NP NP alternant, so their model’s baseline accuracy is 79%.

A second difference between the dative and *needs doing* datasets is that the former is contains 2360 sentences, more than double the number included in the present study. A larger dataset leads to better estimates for the coefficients for each factor, and thus should be expected to improve the overall accuracy. A

third difference is that the dative alternation has been extensively studied, with a wide variety of potentially relevant factors having been identified. By comparison, the *needs doing* alternation does not have been studied on its own before, so there are likely additional factors that speakers consider in the alternation but that this model does not take into account.

That said, the very high correlation between the model and speaker probabilities suggests that the model’s probabilities are nonetheless accurate. The model really is picking up on probabilistic trends; if the alternation were in free variation or dominated by categorical constraints, the correlation value would not be so high. The high correlation but only moderate accuracy suggests an intriguing result: perhaps the *needs doing* alternation is less predictable than the dative alternation. In the dative dataset, most contexts had strong preferences for one alternant over the other. Only a few sentences had 50-50 likelihoods in the model. Judging from Figure 1, a large proportion of the sentences are open to variation in the *needs doing* alternation, which suggests that the *needs doing* construction is less deterministic than the dative alternation. This is only a tentative conclusion, though. Our accuracy values are drawn exclusively from a correlational study. We would need to test the model’s predictions in a controlled study to confidently assert that this is the case.

### 6.3 Random Effect of Verb

Beyond the general accuracy of the model, we can learn more about the specific factors that drive speaker choice by delving into the effects that the model has identified. We start by analyzing the random effect of verb.

Recall that the random effect of verb is assigned on a verb-by-verb basis. Thus, every verb gets its own weight, but all instances of the same verb have the same weight. A positive verb weight indicates that the verb has an idiosyncratic preference for the *to be* alternant; a negative verb weight indicates a preference for the *ing* alternant. The random effects are constrained to fit a normal distribution with mean zero, so approximately as many verbs favor the *to be* alternant as favor the *ing* alternant.

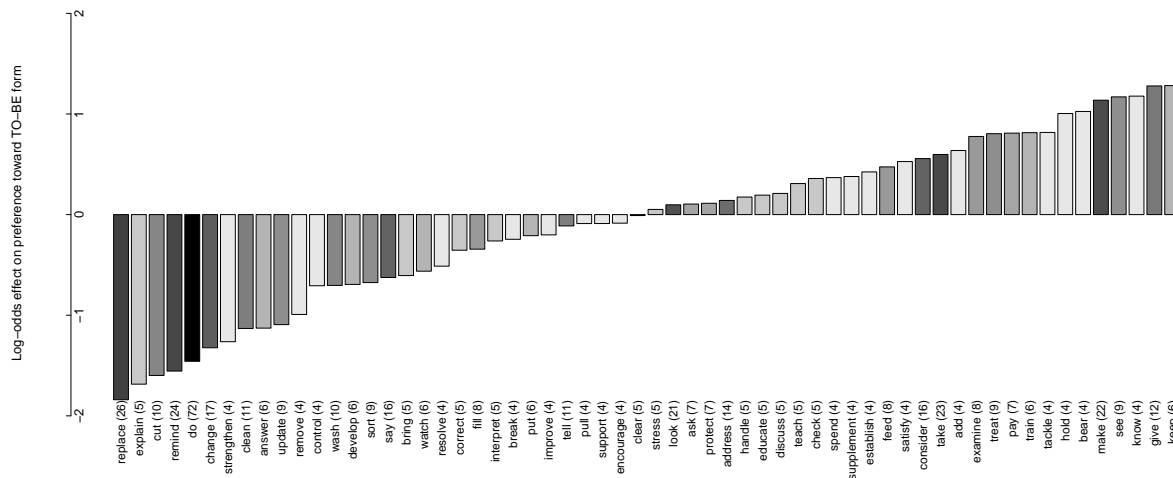
Allowing the model to fit verb-specific effects from the training dataset opens the door to concerns of overfitting. If the model is unconstrained in how it assigns its verb-specific effects, it seems that it could cherry-pick idealized baseline preferences and artificially inflate the accuracy of the model. However, the assignment of the verb-specific coefficients is protected from overfitting in three ways. The first is that the random effects of the verbs are constrained to follow a normal distribution with a mean of zero. Picking one verb’s random effect thus influences what can be chosen for the other verbs, so the model can’t choose the numbers that perfect the model. The second defense against overfitting, as discussed in the previous section, is that the model’s accuracy is cross-validated. If the random effect of verb were to cause overtraining to the training data, then the cross-validated accuracy would actually suffer. The last defense is that, due to the limited coverage of the training set, many of the verbs in the test set are not attested in the training set and thus have no inherent preference for an alternant when we model the test set. Thus we can be confident that these effects are not merely artifacts of overfitting.

Figure 2 graphs the values of the BLUPs (best linear unbiased predictors) for each verb that occurs at least four times in the dataset. Darker bars indicate verbs with more attestations in the dataset, and taller bars indicate a stronger preference for an alternant. Bars going up indicate a preference for *to be* and bars going down indicate a preference for *ing*.

The specific values of the verbs do not reveal any noticeable patterns. There are no clear semantic preferences that arise from these coefficients; semantically similar verbs often show preferences in opposite directions. One example of this is *cut* and *break*, which favor the *ing* and *to be* forms, respectively. It is important to keep in mind that the verb-specific preferences are calculated independent of verb length and frequency, because these are fixed effects in the model.

It is interesting to note that the verb-specific preferences are surprisingly strong. The normal distribution to which they are constrained has a standard deviation of 1.12, which is a bigger effect than more than half of the fixed effects have. More importantly, it is not only rare verbs that have large verb-specific effects. *make*, *remind*, *replace*, and *do* all have more than 20 attestations in the dataset — *do* has 72 — and each of

Figure 2: Graph of the random effect of verb for verbs with at least four attestations in the dataset.



these verbs has a coefficient with an absolute value of at least 1.4. By comparison, the strongest fixed effect in the model is concreteness, with a strength of 1.2.

Further evidence for the strength of the verb-specific preferences is given by a comparison between the mixed-effects model with an effect of verb and a logistic regression model with all the fixed effects but no verb-specific random effect. The log-likelihood of the model without the random effect of verb is -534 compared to a log-likelihood of -505 for the model with a random effect of verb, which is a highly significant improvement according to Equation 8.

We are seeing a very strong idiosyncratic effect of the verb on speaker choice. The reason for this is unclear; there may be some subtle semantic, pragmatic, or phonetic effect that are responsible for some of this effect. Because length and frequency are already in the model as fixed effects, we know that they are not influencing the random effects. It appears that independent of any clear patterns, verbs can have strong idiosyncratic preferences for an alternant.

This result has important consequences for syntactic judgments; the specific choice of a verb can have an unpredictable and idiosyncratic effect on speakers' judgments about the sentence. Given that various alternations often are affected by the same influences on speaker choice, this suggests that for a variety of alternations, a sentence that sounds perfectly acceptable with one verb could be wholly unacceptable with another verb. As an example, note that, from Figure 2, speakers simply prefer to use the gerundive alternant with *replace* and to use the infinitival alternant with *examine*. Then, given otherwise identical sentences, we could expect to see different preferences:

- (66) a. The computer's broken motherboard needs to be replaced.  
 b. The computer's broken motherboard needs replacing.

- (67) a. The computer's broken motherboard needs to be examined.  
 b. (??)The computer's broken motherboard needs examining.

It is difficult to see a strong difference in acceptability between (66a) and (66b), but (67a) does seem a bit better than (67b), suggesting that this lexical effect of verb may affect acceptability. This is only a tentative conclusion and merits further study in a magnitude estimation or related experiment. But even if the effect is not manifested in acceptability judgments, the model shows that speakers are affected by verb-specific preferences for the alternants, both for common and uncommon verbs, and that these effects are often of

Table 3: Coefficients for the significant factors, as fit by the regression. Positive coefficients indicate a preference for the *to be* alternant; negative coefficients indicate a preference for the *ing* alternant. Numeric measures are standardized to fit a normal distribution of mean zero and standard deviation one; this allows their effects to be compared to the categorical factors.

Factor Favoring Infinitive	Coefficient	Factor Favoring Gerund	Coefficient
Concrete Subject	-1.20	Post-Verbal Dependent Length	0.66
Relative Clause	-0.79	Verb Length	0.63
Verb Particle	-0.77	Indefinite Subject	0.61
Atelic Verb	-0.64	Ambig Phrase Length	0.51
Definite Subject	-0.60	Verb Frequency	0.22
Past Tense	-0.30	Telic Verb	0.15

greater strength than the fixed effects in the model.

## 6.4 Fixed Effects

The final version of the model uses the 10 fixed factors that were shown to be significant in Section 6.1. For each of these factors, the improvement in the log-likelihood from including the factor was sufficient to justify the extra degree(s) of freedom its inclusion entailed. Thus the set of factors included in the final model, and their effects in log-odds space is given in Table 3.

What are the theoretical implications of these factors?

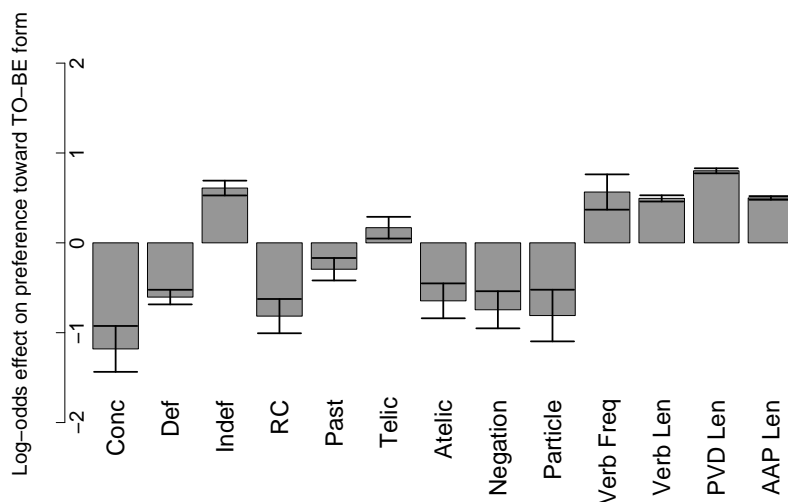
We start by looking at some factors that seemed to likely to affect speaker choice, but turned out to be insignificant. Durativity is perhaps the most surprising of these non-factors. Non-durative verbs (states and achievements) were singled out in Section 3.3.2 as potential influences on speaker choice because they generally lack a progressive *-ing* verb form. However, they are not significantly less likely to appear in the gerundive alternant than durative verbs are, even though this means that they must use an *-ing* verb form. Also, sentence modality (spoken versus written) had an insignificant effect on the final model. This is not unprecedented, as Bresnan et al. (2007) showed that a regression model for the dative alternation that was trained on spoken data performed well on written data as well, but given the sizable difference in alternant proportions in the two modalities, it is nevertheless surprising. The insignificant effect of animacy is also slightly odd, as concreteness was an important factor in the alternation.

As for the significant effects, we start with the accessibility factors: concreteness and definiteness. Concreteness of the subject is an especially strong effect, which is at first blush unremarkable; after all, concreteness has been shown to have an effect in other alternations. However, in other alternations, the effects of concreteness and definiteness were chalked up to a change in word order, as discussed in Section 2.3.5. This fits with the general, cross-linguistic preference for more accessible NPs to occur earlier in a sentence. In this alternation, though, the subject is in the same position in both alternants, so there is no apparent theoretical motivation for concreteness and definiteness to influence the choice of alternation. Because changes to the subject do not move anything within the sentence, these effects do not submit to a straightforward processing or information-theory account of speaker choice. Instead, they point to a hole in our understanding of speaker choice — and given the strength of the concreteness effect, the hole it points at is gaping. This is especially true in light of the clear effect of these factors on acceptability judgments:

- (68)
- a. **Concrete definite:** The cat will need feeding.
  - b. **Concrete indefinite:** ?A cat will need feeding.
  - c. **Abstract definite:** ?The idea will need justifying.
  - d. **Abstract indefinite:** \*An idea will need justifying.

Some of the other effects do not admit to any obvious explanation. The theoretical importance of the

Figure 3: Strengths of the control factors in the regression model. Positive values indicate preference for the *to be* alternant; negative values indicate a preference for the *ing* alternant.



effects of relativization, verb particles, and telicity, for example, are unclear. Perhaps the most important conclusion to be drawn from these effects at present is that they must be taken into account when building a theoretical model of speaker choice, with the underlying motivations for their effects to be filled in later.

The other effects suggest evidence of two deeper theoretical influences: production difficulty and environment prototypicality. Both of these hypotheses can explain the observed highly significant effects of post-verbal dependent and ambiguously-attached phrase lengths, but each has its own shortcomings.

#### 6.4.1 *To be* as a Disfluency

One possible explanation for the remaining effects is that the string *to be* in the *needs to be done* alternant functions like a syntactic disfluency. There are two hypotheses for the role of a syntactic disfluency. The first is that the disfluency is used by the speaker as a means of buying time to alleviate production difficulties. Ferreira and Dell (2000); Race and MacDonald (2003) argue that this drives *that*-omission; *that* is omitted when what follows is readily available, and *that* is included when what follows is more difficult to retrieve. On the other hand, it has also been argued that disfluencies are a case of audience design, a signal to indicate that the speaker anticipates production difficulty. Clark and Fox Tree (2002) argue that this is the case for the disfluencies *uh* and *um*. Jaeger (2005) extends this explanation to *that*-omission, finding that speakers include *that* when they expect to encounter production difficulty, but that including *that* does not itself alleviate the difficulty.

Both hypotheses on the purpose of disfluencies fit with the observed effects in the regression model. Under Ferreira and Dell; Race and MacDonald's alleviation hypothesis, the *needs to be done* alternant includes two short words, *to be*, that the *needs doing* alternant does not. Including *to be* adds a small amount of time to the production of the alternant, during which the speaker could begin preparing the upcoming words; thus, the alleviation hypothesis would predict the longer *needs to be done* alternant to be favored when the upcoming structure is difficult. Clark and Fox Tree; Jaeger's signal hypothesis predicts the same effect; the *to be* in the *needs to be done* alternant may be a signal to the audience of upcoming production difficulty. In the present study, it is impossible to distinguish the two hypotheses, as both would predict the same effects in the model.

The disfluency analysis has some promise in explaining the observed coefficients in the *needs doing*

model. Longer post-verbal dependents (PVDs) and ambiguously-attached phrases (AAPs) both favor the *to be* alternant. This supports the disfluency analysis, as increased phrase length is an indicator of increased production cost (as shorter phrases are generally less syntactically complex than longer ones).

Unfortunately, though, the disfluency analysis does not explain the preference of less common verbs for the *ing* alternant. One would expect that less common verbs would be more difficult to produce and therefore the speaker would prefer to use the *to be* alternant for these verbs to signal the production difficulty. One explanation for this is that because a less common verb indicates production difficulty within the alternant, the signal for approaching difficulty would not do any good within the alternant. Likewise, under the alleviation hypothesis, perhaps production difficulty can only be alleviated outside of the alternant that contains the disfluency. The fact that shorter verbs and less common verbs both favor the *ing* alternant is especially odd because verb length and verb frequency are negatively correlated ( $r = -0.34$ ). Shorter verbs tend to be more common, and longer verbs less common, so one would expect that shorter and more common verbs would favor the same alternant. If frequency or length is removed from the model, the coefficient of the remaining factor still points in the same direction, so this is not an result of an odd interaction between verb length and frequency.

Furthermore, it is unclear if *to be* could function as a disfluency in this alternation without the passive in general being a disfluency. After all, a passive clause has an extra word over its active counter-part; *The child **was** chased by the butterfly* has an extra word (two, really) before the second NP, as compared to *The butterfly chased the child*. Why, then, would the passive be a disfluency within the alternation but not elsewhere? One possibility is that the the passive does not function as a disfluency in general because there is no way to passivize a clause without changing the word order. Changing the word order will usually change the production difficulty by moving the harder NP to a different place in the sentence, and thus would render the disfluency of the passive superfluous. In the *needs doing* alternation, though, the word order is fixed and thus the disfluency effect of the passive could be useful to the speaker. That explanation, though, is clearly speculative. We would first need to find much clearer evidence that *to be* really functions as a disfluency before attempting to argue that the disfluency analysis is reasonable. One possibility, following Jaeger (2005), is to build a regression model fit to only spoken data encoded for disfluencies. In that model, it would be clear if use of the *to be* alternant predicts higher following disfluency rates to get a more direct measure of the effect of production difficulty on the alternation.

The disfluency hypothesis offers a possible explanation for two of the most significant effects in the model, but the observed verb frequency effect runs counter to the hypothesis, and the hypothesis itself is a bit discomfoting. This suggests that a more accurate measure of production difficulty is required to confidently assert that *to be* functions as a disfluency.

### 6.4.2 Structural Bias

Another explanation for the effect of post-verbal dependent and ambiguously-attached phrase length is the Environment Prototypicality Hypothesis of Section 4.1. Recall that the EPH proposed that the *needs doing* alternant would be favored in more nominal environments because it contains the partially-nominal gerund. We propose that the post-verbal dependent and ambiguously-attached phrase effects arise from a tendency of longer PVDs and AAPs to indicate less nominal environments. In that case, the observed preference for the *needs doing* alternant with shorter PVDs and AAPs is due to environment prototypicality effects.

First, let's look at the PVD effect in more depth. Post-verbal dependents are an important part of the environment of the alternation, as they directly modify the past participle or gerund. These dependents are something that would be expected to exert environment prototypicality effects, if these effects exist. Additionally, what is a common dependent for a VP and for an NP are quite different. Verbs tend to have adverbs, *by*-phrases, and CPs following them, whereas nouns tend to be followed by locative PPs and relative clauses. In our dataset, the post-verbal dependents tend to be ones that look more prototypical of VP dependents than of NP dependents. This suggests that increased post-verbal dependent length generally creates a more prototypically verbal environment. Importantly, this observation of more prototypically verbal post-verbal dependents exists despite the fact that the dataset is retrospectively sampled and therefore has equal numbers of *ing* and *to be* alternants.

This suggests that post-verbal dependent length is a crude measure of environment prototypicality. Consider sentences (69a,b).

- (69) a. I *need to be told* [<sub>CP</sub> that he eats candy]  
 b. (?)I *need telling* [<sub>CP</sub> that he eats candy]

In these sentences, the post-verbal dependent *that he eats candy* is a CP, which is a prototypical dependent for a verb but a non-prototypical dependent for a noun. Thus we expect to see the *to be* alternant used here, since it better fits the prototypical environment for a verb. It is important to note that using the post-verbal dependent as a measure of environment prototypicality does not require any assumption about the directionality of the effect. That is, we do not need to assume that the speaker first chooses the post-verbal dependent and then goes back and chooses the alternant, nor do we need to assume that the speaker chooses the alternant first and then chooses the dependent type. The relationship between dependent and environment is a correlation, not necessarily a causation. The relationship holds regardless of which is chosen first, and also holds if the two are chosen simultaneously.

To quantify the prototypicality effect, we introduce structural bias as an approximation of environment prototypicality. The structural bias for the alternation within a given sentence is the ratio of the probability of seeing the current environment if there is an NP in the alternant’s position to the probability of the environment with a VP in the alternant’s position:

$$\text{Structural bias} = \frac{P(\text{Environment}|NP)}{P(\text{Environment}|VP)} \quad (9)$$

High structural bias indicates that the probability of this environment is greater if an NP is observed than if a VP is observed. Thus the Environment Prototypicality Hypothesis predicts that high structural bias will favor the partially-nominal gerundive alternant, and low structural bias (indicative of a more verbal environment) will favor the strictly-verbal infinitival alternant. This probability is difficult to calculate, since it is difficult to define what the environment of the alternant encompasses. Instead, we approximate the bias by reducing the environment to only the post-verbal dependent. Specifically, we will use the first word ( $W_1$ ) and syntactic category of the post-verbal dependent ( $XP$ ) to estimate the bias.

$$\text{Structural bias} \approx \frac{P(W_1, XP|NP)}{P(W_1, XP|VP)} \quad (10)$$

We determine the component probabilities for the structural bias from the parsed Penn Treebank Wall Street Journal corpus. This was done with a series of `tgrep` searches and relative frequency estimations (detailed in the Appendix). The searches listed each word and each post-head modifier type that appeared in an NP or VP in the WSJ corpus. The results of these searches were two files, one listing the number of times each first word and XP pair occurred in an NP and the other listing the same for VPs. We can use Bayes’ Rule to simplify this approximation:

$$\frac{P(W_1, XP|NP)}{P(W_1, XP|VP)} = \frac{\frac{P(W_1, XP, NP)}{P(NP)}}{\frac{P(W_1, XP, VP)}{P(VP)}} \quad (11)$$

$$\propto \frac{\frac{\text{count}(W_1, XP, NP)}{\text{count}(NP)}}{\frac{\text{count}(W_1, XP, VP)}{\text{count}(VP)}} \quad (12)$$

$$\propto \frac{\text{count}(W_1, XP, NP)}{\text{count}(W_1, XP, VP)} \quad (13)$$

(The last step follows from the fact that  $\text{count}(VP)$  and  $\text{count}(NP)$  are constant with respect to the choices of  $W_1$  and  $XP$ .) Thus, the counts from the `tgrep` searches can be directly used to approximate

structural bias.<sup>28</sup> We augment the training data by estimating the structural bias for each sentence in the training set from the first word and syntactic category of its post-verbal dependent.

If PVD length is replaced in the model by structural bias, then structural bias is a significant factor ( $p < .0001$ ), with an effect in the predicted direction: higher structural bias favors the gerundive alternant and lower structural bias favors the infinitival. This supports the Environment Prototypicality Hypothesis, since we supposed that the post-verbal dependent length effect was due to environment prototypicality, and when post-verbal dependent length is replaced the more direct measure of environment prototypicality, this measure is significant.

When both PVD length and structural bias are added to the model described above, structural bias is not a significant factor; the improved likelihood of the model due to structural bias's inclusion is not sufficient to justify the additional degree of freedom. However, this is due to the tight correlation ( $\rho = -0.91$ ) between the post-verbal dependent length and structural bias. Therefore, while it appears that environment prototypicality does affect speaker choice, a better estimate of structural bias, one that is less collinear with PVD length, is needed before environment prototypicality can be definitively confirmed as an influence on the speaker choice. Using  $n$ -gram and/or Hidden Markov Models may lead to a better measure — one less collinear with PVD length — that can determine more definitively the effect of changing syntactic categories.

## 6.5 Model Conclusions

From the model, we see that speaker choice in the *needs to be done* ~ *needs doing* alternation is in many ways like other alternations. Speaker choice in this alternation is subject to accessibility and weight effects, as seen in other alternations. However, these factors affect the *needs doing* alternation despite the constant word order in the alternants — unlike the situation in the dative, genitive, and topicalization alternations. This suggests that word order variation is not the sole motivation behind accessibility effects in speaker choice.

In addition, we observe a surprisingly strong lexical effect of the verb on speaker choice. This effect suggests that when speakers make syntactic decisions, they are using lexical information as much, if not more than, syntactic information. The fact that these lexical effects appear despite the wide range of verbs used in the alternation is surprising, and suggests there is more going on in this alternation than meets the eye.

Furthermore, the alternation presents evidence suggesting an environment prototypicality effect influences speaker choice. This fits with the hypothesis that changes to the syntactic structure of an alternation directly influence speaker choice. This also suggests that the mixed-category membership of the gerund is salient to speakers during production. Finally, there is some evidence for *to be* functioning as a syntactic disfluency, although that result is tentative at best.

## 7 Future Work

There are two primary directions to go for future work. One is to improve the environment prototypicality measure so as to better disentangle it from post-verbal dependent length, allowing us to determine whether one or the other is responsible for the effects we have observed. One possible improvement is to use the head word of the post-verbal dependent, rather than its first word, to determine structural bias. This could improve the model because it is the head that primarily determines the purpose of the post-verbal dependent. A somewhat more possible improvement would be to extend the environment to take account of the whole post-verbal dependent. This could be done by training a Hidden Markov Model to estimate the likelihood of a given post-verbal dependent to modify a noun or a verb. The advantage of a Hidden Markov Model is that it can estimate the likelihood using both the part-of-speech tags and specific word choices for the entire post-verbal dependent. With a better measure of environment prototypicality, we will be able to determine more confidently the effect of changing the syntactic category.

---

<sup>28</sup>The estimate of structural bias from the counts is proportional to the intended estimate of structural bias and thus can be used in the regression model just as the intended estimate would be.

The second direction for future work is to expand beyond a corpus study. Bresnan et al's work on the dative alternation has shown that the production factors identified in corpus studies as affecting the dative alternation have processing correlates in magnitude estimation studies. However, magnitude estimation is an offline measure of processing; there has been no investigation into online effects of the production factors.

Two self-paced reading time studies could fill this gap. In one experiment, the subjects would be asked to read corpus sentences that use one of the alternants. The alternant used in the sentence will be varied across subjects; for instance, one subject would see *The dog needs to be fed* while another would see *The dog needs feeding*. The logistic regression model estimates the probability of choosing an alternant in a given sentence, so we can analyze the relationship between the alternant's probability in each sentence and the reading time in this alternant to see if readers use the probability information from production to predict the likely alternant in processing.

However, this experiment only accounts for information available to the reader before they reach the alternation (e.g., subject animacy), and the investigated environment prototypicality effect relies on post-alternation dependents. A similar self-paced reading time experiment could be run to investigate how salient environment prototypicality is during online processing. In this experiment, both the choice of alternant and the post-alternation dependent will vary. The dependent will alternate between a prototypical VP post-modifier (e.g., an adverbial phrase) and a prototypical NP post-modifier (e.g., a locative PP). If the environment prototypicality effect exists in comprehension, then an alternant/post-modifier mismatch should result in higher reading times on the dependent. Together, these two studies will establish whether production factors are salient during online processing and to what extent syntactic prediction is used in processing.

## 8 Conclusion

In summary, speaker choice in the *needs doing* alternation is determined by a variety of gradient factors, both expected and unexpected. Speaker choice, even in the relatively simple case of alternations, clearly involves a number of considerations that have just to be discovered. We find evidence in support both of the Environment Prototypicality Hypothesis and of *to be* as a syntactic disfluency, but neither of these hypotheses are completely conclusive. Future work on this and other constructions will hopefully shed further light on the mechanisms underlying speaker choice.

## References

- Aarts, B. (2004). Modelling linguistic gradience. *Studies in Language*, 28(1), 1–49.
- Agresti, A. (2002). *Categorical Data Analysis* (2nd ed.). Wiley.
- Baker, M., Johnson, K., & Roberts, I. (1989). Passive arguments raised. *Linguistic Inquiry*, 20(2), 219–251.
- Baron, J. (2000). *Thinking and Deciding*. Cambridge Univ. Press.
- Bock, J. K. (1986). Structural persistence in language production. *Cognitive Psychology*, 18, 355–387.
- Bock, J. K., & Irwin, D. (1980). Syntactic effects of information availability. *Journal of Verbal Learning and Verbal Behavior*, 19, 467–484.
- Bock, J. K., & Levelt, W. (1994). Language production: grammatical encoding. In M. Gernsbacher (Ed.), *Handbook of Psycholinguistics* (pp. 945–984). San Diego: Academic Press.
- Bolinger, D. (1968). Entailment and the meaning of structures. *Glossa*, 2(2), 119–127.
- Bresnan, J., Cueni, A., Nikitina, T., & Baayen, R. H. (2007). Predicting the dative alternation. In G. Bourne, I. Kraemer, & J. Zwarts (Eds.), *Cognitive Foundations of Interpretation* (pp. 69–94). Amsterdam: Royal Netherlands Academy of Science.
- Bresnan, J., & Nikitina, T. (To appear). On the gradience of the dative alternation. In L. Uyechi & L. H. Wee (Eds.), *Reality Exploration and Discovery: Pattern Interaction in Language and Life*.
- Clark, H., & Fox Tree, J. (2002). Using *uh* and *um* in spontaneous speaking. *Cognition*, 84, 73–111.
- Dahl, O., & Fraurud, K. (1996). Animacy in grammar and discourse. In T. Fretheim & J. Gundel (Eds.), *Reference and Referent Accessibility* (pp. 47–64). Amsterdam: Benjamins.

- Davies, M. (2008–). *The Corpus of Contemporary American English (COCA)*. Available from <http://www.americancorpus.org>
- Ferreira, V., & Dell, G. (2000). Effect of ambiguity and lexical accessibility on syntactic and lexical production. *Cognitive Psychology*, 40, 296–340.
- Goldberg, A. (2006). *Constructions at Work*. Oxford University Press.
- Green, G. (1971). Some implications of an interaction among constraints. In *Papers from the Seventh Regional Meeting, Chicago Linguistic Society* (pp. 85–100). Chicago, IL.
- Gropen, J., Pinker, S., Hollander, M., Goldberg, R., & Wilson, R. (1989). The learnability and acquisition of the dative alternation. *Language*, 65, 203–257.
- Haspelmath, M. (1999). Optimality and diachronic adaptation. *Zeitschrift für Sprachwissenschaft*, 18(2), 180–205.
- Jaeger, T. F. (2005). Optional *that* indicates production difficulty: Evidence from disfluencies. In *Proceedings of DiSS05, Disfluency in Spontaneous Speech Workshop* (pp. 103–109). Aix-en-Provence, France.
- Jaeger, T. F. (2006). *Redundancy and Syntactic Reduction in Spontaneous Speech*. Unpublished doctoral dissertation, Stanford U.
- Jaeggli, O. (1986). Passive. *Linguistic Inquiry*, 17, 587–622.
- Levy, R., & Andrew, G. (2006). Tregex and Tsurgeon: tools for querying and manipulating tree data structures. In *5th International Conference on Language Resources and Evaluation (LREC 2006)*.
- Levy, R., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In B. Schlökopf, J. Platt, & T. Hoffman (Eds.), *Advances in Neural Information Processing Systems* (Vol. 19, pp. 849–856). Cambridge, MA: MIT Press.
- Malouf, R. (2000). *Mixed Categories in the Hierarchical Lexicon*. Stanford: CSLI Publications.
- Mann, L., & MacWhinney, B. (1984). The Repeated Morph Constraint: Toward an explanation. *Language*, 60, 519–541.
- Murray, T., Frazer, T., & Simon, B. L. (1996). Need + past participle in American English. *American Speech*, 71(3), 255–271.
- Murray, T., & Simon, B. L. (1999). Want + past participle in American English. *American Speech*, 74(2), 140–164.
- Murray, T., & Simon, B. L. (2002). At the intersection of regional and social dialects: The case of like + past participle in American English. *American Speech*, 77(1), 32–69.
- Neuhaus, J., & Jewell, N. (1990). The effect of retrospective sampling on binary regression models for clustered data. *Biometrics*, 46(4), 977–990.
- Ortmann, A. (1998). The role of [ $\pm$  animate] in inflection. In R. Fabry, A. Ortmann, & T. Parodi (Eds.), *Models of Inflection* (pp. 60–84). Tübingen: Niemeyer.
- Partee, B. (2007). *Handout from MGU Seminar in Formal Semantics*. Available from [http://people.umass.edu/partee/MGU\\_2007/MGU0713\\_2up.pdf](http://people.umass.edu/partee/MGU_2007/MGU0713_2up.pdf)
- Pinheiro, J., & Bates, D. (2000). *Mixed-Effects Models in S and S-PLUS*. New York: Springer.
- Pinker, S. (1989). *Learnability and Cognition: The Acquisition of Argument Structure*. Cambridge, MA: MIT Press.
- Prat-Sala, M. (1997). *The production of different word orders: A psycholinguistic and developmental approach*. Unpublished doctoral dissertation, U. of Edinburgh.
- Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.
- Race, D., & MacDonald, M. (2003). The use of “that” in the production and comprehension of relative clauses. In *Proceedings of the 25th Annual Meeting of the Cognitive Science Society*.
- Rosenbach, A. (2003). Aspects of iconicity and economy in the choice between the s-genitive and the of-genitive in English. In G. Rohdenberg & B. Mondorf (Eds.), *Determinants of Grammatical Variation in English*. Berlin/New York: de Gruyter.
- Siewierska, A. (1988). *Word Order Rules*. Croom Helm.
- Silverstein, M. (1976). Hierarchy of features and ergativity. In R. M. W. Dixon (Ed.), *Grammatical Categories in Australian Languages* (pp. 112–171). Australian Institute of Aboriginal Studies.

- Smith, N., & Levy, R. (2008). Optimal processing times in reading: a formal model and empirical investigation. In *Proceeding of the 30th Annual Meeting of the Cognitive Science Society*.
- Snider, N. (2005). *A corpus study of left dislocation and topicalization*. Presentation at QPFest 2005. Available from [http://www.stanford.edu/snider/pubs/qpfest05\\_talk.pdf](http://www.stanford.edu/snider/pubs/qpfest05_talk.pdf)
- So, N. V. (1973). *The semantic interpretation of infinitives and gerunds as sentential complements*. Unpublished master's thesis, UCLA.
- Szmrecsányi, B. M. (2004). On operationalizing syntactic complexity. In *Proceedings of the 7th International Conference on Textual Data Analysis* (pp. 1032–1039).
- CMU Pronunciation Dictionary. (n.d.). Available from <http://www.speech.cs.cmu.edu/cgi-bin/cmudict>
- Reference Guide for the British National Corpus (XML Edition)*. (n.d.). Available from <http://www.natcorp.ox.ac.uk/XMLedition/URG/>
- Vendler, Z. (1957). Verbs and times. *The Philosophical Review*, 66(2), 143–160.
- Weiner, J., & Labov, W. (1983). Constraints on the agentless passive. *Journal of Linguistics*, 19, 29–58.
- Yamamoto, M. (1999). *Animacy and reference*. Amsterdam: Benjamins.
- Zaenen, A., Carletta, J., Garrettson, G., Bresnan, J., Koontz-Garboden, A., Nikitina, T., et al. (2004). Animacy encoding in English: why and how. In *ACL 2004 Workshop on Discourse Annotation*.
- Zaenen, A., O'Connor, M. C., & Wasow, T. (2007). *Slides from their class at the LSA 2007 Summer Institute: LSA 341, Paraphrase and Usage*.

## Appendix: Commands used in work on the dataset

### tgrep searches to create the dataset

The `tgrep` searches were run over a version of the BNC that had been automatically parsed by Doug Roland. Thus reference could be made to syntactic categories in the search strings. However, reliance on the syntactic structure was limited, because automatically parsed corpora are full of bad tree structures. Instead, linear ordering was the primary search criterion.

Three `tgrep` searches were used to compile the corpus:

```
(S << (/need(s|ed|ing)?\>/ . (/V(B|D|H|V)G/ < /[^ ]+ing/)))
(S << (/need(s|ed|ing)?\>/ . ((/[^ ]/ < /\<to\>/) . ((/[^ ]/ < /\<be\>/) . (/V(B|D|H|V)(D|N)/))))
(S << (/need(s|ed|ing)?\>/ . ((/[^ ]/ < /\<to\>/) . ((/[^ ]/ < /\<be\>/) . (/ [^ ]/ < /done/))))
```

The first of these collects all instances of a possibly inflected form of *need* followed by a verb ending in *ing*. That the following word had to be a verb was included to avoid false positives like *need something*. The second expression collects all instances of a possibly inflected form of *need*, followed by *to be*, followed by any verb in the past tense or the past participle. This needed to be augmented with the third search because *done* was never marked as a verb in the corpus, but usually as an auxiliary. The third search catches all instances of a form of *need* followed by *to be done*, and completes the dataset.

### False Positives

The `tgrep` searches returned 5926 sentences, but not all them were actual instances of the *needs to be done* or *needs doing* constructions. Some false positives are clear; the examples in Section 5.2 were clear examples of false positives, where the adjective *bored* was mislabelled as a past participle, or where *need* and a gerund were next to each other, but not members of the same constituent. However, in some sentences it is unclear if the gerund or past participle is sufficiently verbal to qualify as an example of the alternation:

(70) He would *need counselling*.

In (70), *counselling* could be either a mixed-category gerund, referring to the act of being counselled, or a strict noun, referring to a type of therapy. If it is the former, then (70) means the same as (71a); if it is the latter, then (70) means the same as (71b).

- (71) a. He would need to be counselled.  
b. He would need therapy.

Unfortunately, these two sentences have virtually identical meanings, so it is difficult to determine whether *counselling* in (70) is a true gerund or a noun. If it is a noun, then this is not an instance of the alternation, since it cannot alternate with *need to be counselled*. If it is a gerund, then this is an instance of the alternation and could be included in the dataset. A similar problem can arise with the *needs to be done* form:

- (72) [The seat] would *need to be curved* to avoid discomfort.

Here *curved* could be either a gerund or an adjective. If *curved* is a gerund, then (72) is the passive form of an agentless passive (73a); if *curved* is an adjective, then (72) has a meaning more like (73b).

- (73) a. They would need to curve the seat to avoid discomfort.  
b. The seat would need to be a curved seat to avoid discomfort.

Again, the meaning difference between (73a,b) is minor, so it is difficult to determine which is a better match for (72), and therefore whether (72) is an example of the alternation or not. Sentences like these are a double-edged sword. If they are included in the dataset and are not really instances of the alternation, they violate the requirement in Weiner and Labov (1983) that only alternating sentences can be included in the regression. If they are omitted but are really instances of the alternation, it could skew the verb-specific effects; *curve*, for instance, could appear to be biased strongly toward the *needs doing* alternant if every instance of *needs to be curved* is excluded as an adjectival usage.

In the end, sentences such as these were excluded unless the sentence's context strongly suggested that the speaker really could choose either alternant in the sentence. Fortunately, such sentences were rare, and there were no verbs that repeatedly occurred in such sentences, so it is unlikely that these exclusions will impact the verb-specific preferences significantly.

## Environment Prototypicality Calculations

In Section 6.4.2, structural bias was estimated from the frequency of different post-modifiers of VPs and NPs in the parsed Penn Treebank. This was estimated as follows.

First, the Penn Treebank was edited with *tsurgeon* (Levy & Andrew, 2006) to give cleaner counts of the post-modifiers, which consisted of three steps. The first was to mark coordinated NPs and VPs to exclude them from the calculations. This was done to avoid having to determine if a post-modifier modified both coordinates or only the last coordinate:

- (74) a. The cats and [dogs in the house]  
b. The [[cats and dogs] [in the house]]

The second step was to collapse the VPs and NPs. This was done to avoid overcounting the VPs and NPs; otherwise each argument or adjunct could be counted as a post-modifier in a different level of the NP or VP. This step converted a structure like (75a) to a structure like (75b).

- (75) a. The old man [<sub>VP</sub> [<sub>VP</sub> [<sub>V</sub> crept] [<sub>PP</sub> into the room]] [<sub>ADVP</sub> silently]].  
b. The old man [<sub>VP</sub> [<sub>V</sub> crept] [<sub>PP</sub> into the room] [<sub>ADVP</sub> silently]].

The final step was to excise NPs that were the immediate right-sister of the head of a VP. This was done to simulate passivization; recall that verbs participating in the alternation have their semantic objects occurring in the syntactic subject position. Thus, if there is a transitive verb in the Penn Treebank with its object as a post-verbal dependent, this object should not be counted as a post-modifier, since it would not be a

post-modifier if the verb were used in the alternation. (76) shows this; the object of *walk* moves outside the VP when *walk* is used in the alternation.

(76) I [walked *the dog*], because *the dog* needed walking.

These three steps were accomplished with the following `tsurgeon` commands:

```
VP=vp < CONJ|CC      relabel vp CVP
NP=np < CONJ|CC      relabel np CNP
@VP < (@VP=vp)       excise vp vp
@NP < (@NP=np)       excise np np
__ >># @VP $+ @NP=np relabel np DELNP
DELNP=np             delete np
```

These `tsurgeon` commands create a new version of the Penn Treebank that we can get post-head dependent counts from. From the annotations in the dataset, we know the first word ( $W_1$ ) and phrase type ( $XP$ ) for each post-verbal dependent observed in the training set. To get the counts for these ( $W_1, XP$ ) pairs, we use a series of `tregex` (Levy & Andrew, 2006) searches over the revised Penn Treebank:

```
tregex -h targ '(@NP <# (__ $+ (@ADVP <<, (__=targ !< __))) )'
```

This search will return the first word of each adverbial post-modifier that is the first post-modifier of an NP; these words can then be sorted to get the counts for each ( $W_1, ADVP$ ) pair. The same basic search can be run with NP or VP in the first position and with ADVP, SBAR, PP, or NP in the second position to amass the counts for each ( $W_1, XP$ ) pair that is required. The last count that is needed is a count of the number of VPs and NPs with no post-modifiers; these counts are used to estimate the structural bias of not having a post-verbal dependent. This count is calculated by finding the total number of VPs (or NPs) and subtracting the number of VPs (or NPs) with at least one post-modifier:

```
tregex '(@VP <# __)'
```

```
tregex '(@VP <# (__ $+ __))'
```