Environment prototypicality in syntactic alternation

GABRIEL DOYLE AND ROGER LEVY
UC San Diego Linguistics

1. Introduction

This paper investigates the effect of changing syntactic categories on speaker choice, using the *needs to be done* ∼ *needs doing* alternation as the testing ground. The two alternants in this alternation have different syntactic properties, and so we hypothesize that the syntactic preferences of the alternation's environment influence alternant choice. Our hypothesis is that, all else being equal, speakers prefer to use an alternant with a syntactic category that is more prototypical given the rest of the sentence. We show an effect of structural bias that argues for this hypothesis.

2. Speaker choice and syntactic alternations

For a given situation, there is a large, possibly infinite, set of sentences that could be said, all expressing the same core idea. Different sentences in the set may be more or less appropriate for the situation, and thus better or worse choices for the speaker. The relative appropriateness of a sentence is presumably a complicated calculation, yet speakers choose their sentences fluidly. So how do speakers choose among the various sentences they can use to express an idea?

This is a big question, too big to be answered completely with our current state of knowledge. However, instead of looking at speaker choice in sentence-level variation, we can simplify the problem by investigating how speakers choose between alternants in a syntactic alternation. A syntactic alternation is a situation in which there are multiple related phrases expressing the same semantic idea with different syntactic forms. Canonical examples of this are *that*-omission and the passive, dative, and genitive alternations. Looking at syntactic alternations rather than sentence-level variability simplifies the problem in three important ways:

- The alternants are nearly meaning-equivalent
- Limited set of alternants makes option comparison simpler
- Smaller region of variation allows for better experimental control

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.

These studies (esp. Weiner and Labov 1983, Bresnan and Nikitina 2007) have dispelled the notion that categorical semantic constraints are the primary determinants of speaker choice. Rather, speaker choice involves an interaction of categorical and gradient constraints. A categorical constraint is a hard effect on the probability of a speaker choosing an alternant; an alternant that violates a categorical constraint has no probability of being uttered except as a speech error. A gradient constraint is a soft effect on this probability; an alternant that violates a gradient constraint has a lower probability of being uttered than an alternant that does not violate the constraint, but the probability does not drop all the way to zero. Categorical constraints define the space of syntactic alternation: where each alternant can occur. When both alternants are available, gradient constraints determine which form will be used.

Previous work on alternations has identified two main types of gradient factors affecting speaker choice: accessibility and priming. In many alternations (Bresnan et al. 2007 for datives, Rosenbach 2003 for genitives, a.o.), speakers prefer the alternant that places an animate concrete NP earlier in the sentence. Rosenbach explains this as a result of a cross-linguistic preference for animate NPs to occur earlier in a sentence, which is argued to be a cognitive universal based on increased cognitive accessibility of animate and concrete NPs. Discourse status, pronominality/definiteness, and weight all exhibit similar accessibility effects on speaker choice (Rosenbach 2003; Bresnan et al. 2007).

Speaker choice is also influenced by the recent occurrence of an alternant. Bock (1986) and Bresnan et al. (2007) show that syntactic parallelism influences speaker choice in the passive and dative alternations, influencing the speaker to repeat the primed structure.

Although many factors have been identified that affect speaker choice, a variety of important factors remain uninvestigated. The present study looks at how the use of different syntactic categories in different alternants affects speaker choice. Many alternations have different structures for their alternants. For instance, the dative alternation switches between having two predicate NPs and having an NP and a PP in the predicate. Changing categories has not yet been studied in these alternations, perhaps because the effect of changing category structure is entwined with changes to the word order. Word order is fixed, though, in the *needs doing* alternation, so the effect of syntactic category can be disentangled from the effect of word order and studied for its influence on speaker choice. Note, though, that this is still not a perfectly clean contrast; *to be* intervenes between *needs* and the verb in one alternant but not the other.

3. The anatomy of the *needs doing* alternation

The *needs doing* alternation has not been extensively studied, receiving brief mentions in English grammars (e.g., Quirk et al. 1985) and little other notice. As such, only a quick overview of the alternation will be given. It should be noted that, although throughout the paper we refer to the alternation as the *needs doing* alternation, *do* is clearly not the only verb that can be used. It appears that any passivizable transitive verb can occur. Intransitive verbs cannot be used, because in both alternants "it is not the understood subject of the participle, but its understood object that is identified with the subject of the superordinate clause." (Quirk et al. 1985:1189) This is the only apparent restriction on verbs in the alternation.

As with other alternations, the *needs to be done* and *needs doing* alternants are usually, but not always, approximately equally acceptable. For instance, there is little clear or consistent difference in acceptability between (1a) and (1b), but the acceptabilities of (2a) and (2b) do differ (although (2b) is still attested; see Doyle 2008):

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

(2) a. You need to be shown the way.
b. *?You need showing the way.

This raises two important (and entwined) questions: what accounts for the differing acceptability judgments, and how does a speaker choose which alternant to use?

We assume, following previous work, that speaker choice is driven by a set of gradient constraints that influence which alternant will be chosen. We will stay intentionally agnostic about the specific cognitive mechanism of speaker choice, and simply assume that speaker choice is probabilistic. In this framework, we presume that, given a sentence environment S, the gradient constraints combine to yield an overall probability $P(\text{needs to be done}|S)$ of choosing one alternant in this sentence. The probabilistic framework implies that the same speaker can choose different alternants every time she sees a given environment, which fits intuitively with what we observe in actual usage. We estimate the probability $P(\text{needs to be done}|S)$ with a probabilistic model, mixed-effects logistic regression. 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; this is the standard model in most alternation studies. Using a mixed-effects regression allows different verbs to have idiosyncratic preferences for the alternants (Bresnan et al. 2007).

4. The gerund as a noun and a verb

The alternants in the *needs doing* alternation involve different syntactic categories. The past participle in *needs to be done* is verbal with no nominal properties, while the gerund in *needs doing* is has both verbal and nominal properties (Malouf 2000):

- Gerunds can govern NPs. [verbal]
- Gerunds are modified by adverbs, not adjectives. [verbal]
- A gerundive phrase has the same external distribution as an NP. [nominal]
- The gerund can take an optional genitive or accusative subject. [both]

There are mixed opinions on gerunds' category. Malouf argues that the gerund's properties are the result of mixed category membership; each token is simultaneously a verb and a noun. Others (e.g., Aarts 2004) contend that the gerund is merely category-ambiguous; each token is either a noun or a verb, albeit an atypical member of the category. We will sidestep this debate in the current paper, as the fact that gerunds can simultaneously exhibit both nominal and verbal properties is sufficient for our purposes.

Unlike the gerund, the past participle has no nominal characteristics. It may not be strictly verbal, as it can function as an adjective, but for this study it is sufficient that the past participle is not nominal.

5. EPH: Environment Prototypicality Hypothesis

This category difference leads to a simple hypothesis: speakers may prefer the more nominal *needs doing* alternant in more nominal environments. What do we mean by a "more nominal" environment? Consider sentences (3a,b).

(3) a. *The couch needs a to be cleaned
    b. The couch needs a cleaning.

Clean in each of these sentences is in a prototypical place for a noun. In both sentences, it follows a determiner. This is a highly prototypical place for a noun and a highly non-prototypical place for a verb — so highly non-prototypical, in fact, that it is grammatically unacceptable. Now consider sentences (4a,b).

(4) a. The paper needs to be completely rewritten.
    b. ?The paper needs completely rewriting.

Here *rewrite* in each sentence follows an adverb, which is a 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 awkward in this context.

These observations leads to the hypothesis that environment prototypicality is a factor affecting speaker choice, like accessibility or other previously observed effects. Specifically, 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. The only remaining issue is how to quantify the prototypicality of an environment. We answer this question in Section 7.3.

6. Modeling the alternation

To test the Environment Prototypicality Hypothesis, we first build a mixed-effects logistic regression model (Agresti 2002) to determine what previously-researched factors affect speaker choice in this alternation, then add a measure of environment prototypicality to the model to gauge its effect on speaker choice. We present an overview of the model construction here. Doyle 2008 provides more detail about the the dataset, the control factors, and lack of categorical constraints in the alternation.

6.1. Dataset

The regression model is trained on a dataset of 1004 sentences from the British National Corpus (BNC), each containing an instance of the alternation. This training set is a subset of 5926 sentences from the BNC, found by a tgrep2 search over a parsed version of the corpus (Doug Roland, p.c.). The 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, and thus includes some false positives. These false positives were removed when constructing the training set.

Before being included in the model dataset, each sentence was manually annotated for the control features (animacy, concreteness, etc.). 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 a majority 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 small 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.

Table 1: List of control factors included in the logistic regression model. Significant control factors, as determined by likelihood-ratio tests, are noted with asterisks. [* $-$ $p < .05$, ** $-$ $p < .01$, *** $-$ $p < .001$]

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***
relativiziation***	negation	verb frequency***
tense*	modal	
adverb	conjoined material	

6.2. Factors considered

We include 19 control factors in the regression model, based on the previously observed effects in Section 2. These factors are listed in Table 1. We highlight two factors here.

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 (PVDs), ambiguously-attached phrases (AAPs), and syntactically separate constituents.

Post-verbal dependents are constituents that unambiguously modify the alternation. Such dependents are arguments or adjuncts of the verb in the alternation. Ambiguously-attached phrases could modify either the verb in the alternation or the sentence as a whole. The last category, syntactically separate constituents, refers to constituents that are clearly unconnected to the alternation. Only PVDs and AAPs are included in the model, since syntactically separate constituents do not modify the alternation.

We are interested in possible weight effects emerging from these phrases, so we included the smoothed log of the length in words of these phrases in the model. If a sentence has more than one post-verbal dependent, their lengths are summed before the log-transform is applied. The same is done if there is more than one ambiguously attached phrase.

6.3. Potential categorical constraints on the alternation

Before we build a regression model, we must first account for any major categorical constraints that could restrict the alternation (Weiner and Labov 1983). Unlike better-studied alternations, there are few proposed constraints on the *needs doing* alternation; in fact, the only proposals in the literature are three speculative constraints from Lynne Murphy (cited in Murray, Frazer, and Simon 1996). We tested these constraints and two of our own devising, but counterexamples to each could be found in the British National Corpus or on

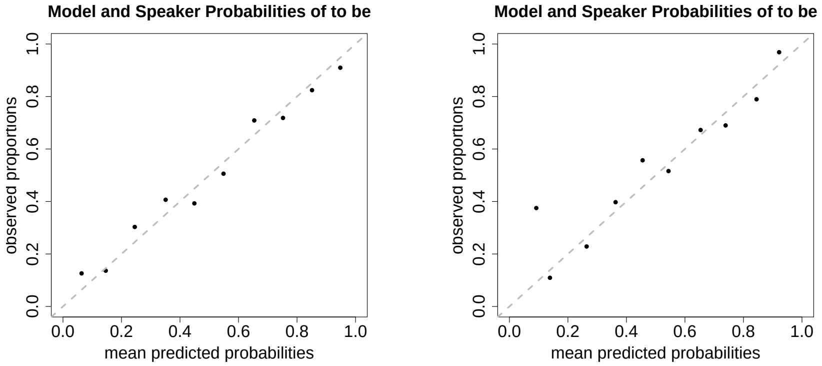


Figure 1: Correlation between model and speaker probabilities. The x-axis gives the probabilities predicted by the model, and the y-axis gives the probabilities observed in the dataset. The left graph is for a model with all significant control factors, and the right graph is for a model with only a single factor, concreteness.

the Web. These proposals and their counterexamples are discussed in Doyle 2008.

7. Results and Discussion

In this section, we discuss the regression model's ability to model speaker choice in the *needs doing* alternation. We begin by examining the model with just the control factors, then investigate the Environment Prototypicality Hypothesis as an explanation for the effects of one of the control factors.

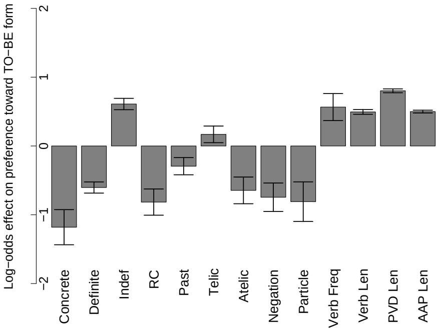
7.1. Model Accuracy

We estimate model accuracy in two ways: forced choice and probability matching. In both cases, the accuracy measurements are calculated using 5-way cross-validation, and all accuracy values are given as the mean and standard deviation from 10 trial runs.

The basic measure of model accuracy forces the model to predict the more likely alternant for each sentence in the dataset. The model succeeds if it predicts the alternant that the speaker used. Under this evaluation metric, the model averages $74.8 \pm .6\%$ accuracy on the test sets. The baseline accuracy is 50%, since each alternant is equally likely in the dataset. However, because we have assumed that speaker choice is probabilistic, forcing the model to choose an alternant loses probability information and potentially deflates its accuracy.

A better way to estimate the model's similarity to non-deterministic speaker choice is to compare the actual proportions of each alternant used in sentences with similar probabilities in the model. We follow the method used by Bresnan et al. (2007). Each sentence S is assigned a probability $P(\text{to be}|S)$ by the

Figure 2: 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.



model, and the sentences are divided into 10 bins based on these probabilities. The observed probability of *to be* for each bin is proportion of sentences in the bin that actually use the *to be* alternant, and this is compared to the mean value of $P(\text{to be}|S)$ in each 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.

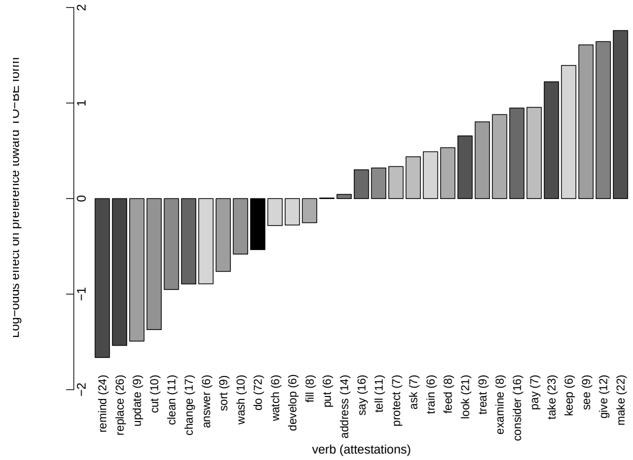
Figure 1 visualizes these results. If the model were a perfect estimator for speaker choice, then all the points would sit on the dotted grey line and the correlation would be 1. The correlation for the model using the control factors is $R^2 = .994 \pm .004$, showing that the model probabilities tightly fit the observed proportions. By comparison, a version of the model that only uses a single factor to predict speaker choice has a worse line fit and the correlation drops to $.91 \pm .14$. Thus we see that the model is successfully matching the actual probabilistic usage of the alternation.

It is worth noting that the *needs doing* model's accuracy is much lower than that of Bresnan et al's model of the dative alternation, which performed with 95% accuracy. The likely culprit here is retrospective sampling; baseline accuracy on our retrospectively sampled dataset is 50%, whereas their dataset has a baseline accuracy of 79%. The very high correlation between the model and speaker probabilities (comparable to Bresnan et al's) suggests that the model is nonetheless accurate at estimating the speaker probabilities.

7.2. Control Factors

The final version of the model uses the 10 significant control factors out of the original set of 19 (see Table 1). Significance of a factor was assessed by likelihood-ratio tests, which compare the likelihood of the dataset in a model with the factor to the likelihood of the dataset in a model without the factor.

Figure 3: The best linear unbiased predictors of the random effect for verbs with more than five attestations in the dataset. Positive values indicate preference for *to be*. Darker bars indicate verbs with more attestations in the dataset. There are no apparent patterns in the effect strengths for different verbs.



For each of the significant factors, the improvement in the likelihood from including the factor is sufficient to justify the extra degree(s) of freedom its inclusion entails. The strengths of these factors, as linear effects in the log-odds of the *needs to be done* alternant, are shown in Figure 2. A random effect of verb is also included in the model to account for verb-specific preferences between the alternants. The random effects of commonly attested verbs in the dataset are shown in Figure 3.

The control factor that most improves the model is post-verbal dependent (PVD) length. (The random effect of verb is a close second.) This is also one of the strongest control factors, with longer dependents favoring the *needs to be done* alternant. However, since the PVD occupies the same position in the sentence with either alternant, these effects are unlikely to be accessibility-related weight effects. Instead, these effects suggest an underlying environment prototypicality effect.

7.3. Structural Bias

We begin by looking at the PVD effect in more depth. Post-verbal dependents directly modify the past participle or gerund, so one would expect them 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 sentential complements following them, whereas nouns tend to be followed by locative PPs and relative clauses. In the training data, the post-verbal dependents for both constructions tend to look more prototypical of VP dependents than of NP dependents, despite equal numbers of *ing* and *to be* alternants in the dataset. This suggests that increased post-verbal dependent length generally creates a more prototypically verbal environment, and thus that the effect of PVD length is due in part to environment prototypicality. Consider sentences (5a,b).

(5) a. I *need to be told* [that he eats candy]
b. (?)I *need telling* [that he eats candy]

In these sentences, the post-verbal dependent *that he eats candy* is a sentential complement, which is a prototypical dependent for a verb but a non-prototypical dependent for most nouns. Thus we expect to see the *to be* alternant used here, since it better fits the prototypical environment for a verb.

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 [Eqn. 1]. These probabilities are difficult to calculate, since it is unclear what the environment of the alternant encompasses. Therefore, we approximate the bias by only considering the post-verbal dependent. Specifically, we use the first word (W_1) and syntactic category of the post-verbal dependent (XP):

$$\text{Structural bias} = \frac{P(Environment|NP)}{P(Environment|VP)} \approx \frac{P(W_1, XP|NP)}{P(W_1, XP|VP)} \tag{1}$$

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 *ing* alternant, and low structural bias (indicative of a more verbal environment) will favor the strictly-verbal *to be* alternant. We can simplify our approximation in Eqn. 1 using Bayes' Rule:

$$\begin{aligned}
\frac{P(W_1, XP|NP)}{P(W_1, XP|VP)} &= \frac{\frac{P(W_1, XP, NP)}{P(NP)}}{\frac{P(W_1, XP, VP)}{P(VP)}} && (2)\\
&\propto \frac{count(W_1, XP, NP)}{count(W_1, XP, VP)} \cdot \frac{count(VP)}{count(NP)} && (3)\\
&\propto \frac{count(W_1, XP, NP)}{count(W_1, XP, VP)} && (4)
\end{aligned}$$

Because the count ratio is proportional to the structural bias, it can be used in place of bias in the regression without changing the results (Agresti

2002). We estimate the counts in Eqn. 4 using tgrep searches over the Penn Treebank Wall Street Journal corpus (detailed in Doyle 2008).

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 *ing* alternant. 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 structural bias and PVD length are in the model, structural bias is not significant. This appears to be a result of the extremely tight correlation ($\rho = -0.91$) between the PVD length and our estimate of 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. In future work, we plan to explore n-gram and/or Hidden Markov Models as better measures of environment prototypicality to determine more definitively the effect of changing syntactic categories.

8. Conclusion

Speaker choice in the *needs doing* alternation is determined by a variety of gradient factors. We find evidence in support of the Environment Prototypicality Hypothesis by looking at the distribution of dependents of the alternation, although this is somewhat confounded with dependent length. Future work on this and other alternations will hopefully shed further light on syntactic categories' effects on speaker choice.

Acknowledgments

We wish to thank Joan Bresnan, Grant Goodall, T. Florian Jaeger, Andy Kehler, Tatiana Nikitina, Tom Wasow, Arnold Zwicky, and the members of the UCSD Computational Psycholinguistics Lab for comments and advice on this project.

References

Aarts, Bas. 2004. Modelling linguistic gradience. *Studies in Language 28*(1), 1–49.

Agresti, Alan. 2002. *Categorical Data Analysis* (2nd ed.). Wiley.

Bock, J. Kathryn. 1986. Structural persistence in language production. *Cognitive Psychology 18*, 355–387.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina, and R. Harald Baayen. 2007. Predicting the dative alternation. In G. Bourne, I. Kraemer, and J. Zwarts (Eds.), *Cognitive Foundations of Interpretation.* Royal Netherlands Academy of Science.

Bresnan, Joan and Tatiana Nikitina. 2007. On the gradience of the dative alternation. In L. Uyechi and L. H. Wee (Eds.), *Reality Exploration and Discovery: Pattern Interaction in Language and Life.*

Doyle, Gabriel. 2008. Determinants of variation in the needs doing construction. Unpublished manuscript, UC San Diego, May 2008 http://idiom.ucsd.edu/∼gdoyle/comps1.pdf.

Malouf, Robert. 2000. *Mixed Categories in the Hierarchical Lexicon.* Stanford: CSLI Publications.

Murray, Thomas, Timothy Frazer, and Beth Lee Simon. 1996. Need + past participle in American English. *American Speech 71*(3), 255–271.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language.* London: Longman.

Rosenbach, Annette. 2003. Aspects of iconicity and economy in the choice between the s-genitive and the of-genitive in English. In G. Rohdenberg and B. Mondorf (Eds.), *Determinants of Grammatical Variation in English.* de Gruyter.

Weiner, Judith and William Labov. 1983. Constraints on the agentless passive. *Journal of Linguistics 19*, 29–58.

Gabriel Doyle
University of California, San Diego
Department of Linguistics
9500 Gilman Dr., Mail Code 0108
La Jolla, CA 92093-0108

gdoyle@ucsd.edu