DOI: 10.26346/1120-2726-133

The role of individual variation in variationist corpus-based studies of priming

Michael Gradoville

School of International Letters and Cultures, Arizona State University, United States
michael.gradoville@asu.edu

This paper addresses the relationship between individual variation and priming effects in variationist corpus-based linguistic studies. Over the past decade, researchers have argued for the consideration of hierarchical relationships in sociolinguistic data (Johnson 2009) and corpus linguistic data (Gries 2015). While much of the emphasis on hierarchical relationships has pertained to macro-level predictors such as sociolinguistic social factors that are properties of the groupings in the data, speaker for example, than any one observation in particular, failure to account for this grouping relationship can affect the accuracy of estimates of micro-level predictors. Using a corpus of educated spoken Portuguese from Fortaleza, Brazil, this study investigates priming effects in the reduction of *para* 'to, for, in order to' to *p(r)a*. In particular, this study finds that many speakers in the sample show no positive evidence for priming and, furthermore, that the importance of the previous occurrence predictor is overstated due to the wide individual differences in the present data set[*].

KEYWORDS: hierarchical data, speaker-specific effects, generalized linear (mixed-effects) models, priming, Portuguese *para*

1. Introduction

Over the past decade, the issue of statistical methodology has received much attention in variationist sociolinguistic circles (Johnson 2009; Paolillo 2013; Roy 2013; Tagliamonte 2012: 129-130) as well as in corpus linguistic circles (Gries 2015). At issue in sociolinguistics is the continued use of versions of the variable rule program, the current versions of which are known as Goldvarb (Sankoff *et al.* 2005, 2015), in variationist sociolinguistic studies when other statistical programs of more general use and of (generally) more flexible capabilities have

[*] An earlier version of this project was presented at the 12th Conceptual Structure Discourse and Language Conference held at the University of California, Santa Barbara in November 2014. The author gratefully acknowledges the comments and suggestions of those in attendance as well as those of two anonymous reviewers. Any shortcomings are my own.

been developed since the introduction of the first variable rule programs more than four decades ago (Cedergren & Sankoff 1974). Among the drawbacks of the Goldvarb software identified by critics are its inability to handle continuous predictors and response variables, its reliance on an automated stepwise model selection procedure, its unique character-based input requirements, the relative difficulty of using Goldvarb to model interactions between predictors, and the program's inability to model grouping structures within the data.[1]

The present study is focused on this last issue of groupings in sociolinguistic data. Many social science data sets, including those used in variationist sociolinguistics, include hierarchical relationships, which violate the assumption made by single-level statistical models, such as Goldvarb, that "observations should be sampled independently from each other" (Snijders & Bosker 1999: 6). For example, in the field of education, one of the most recurrent examples of such a hierarchical relationship is pupils (micro-level) in a particular class (macro-level) on a standardized test. In its usual implementation, a single-level statistical model cannot account for the fact that, for example, multiple pupils have the same teacher and thus the performance of the group might be better attributed to the teacher instead of individual students. Snijders & Bosker (1999) assert that, in a statistical model that includes multiple students per class and multiple classes (see Figure 1 for a visual representation of this situation), the observations within each classroom are not independent of one another, a violation of the assumptions of single-level statistical tests.

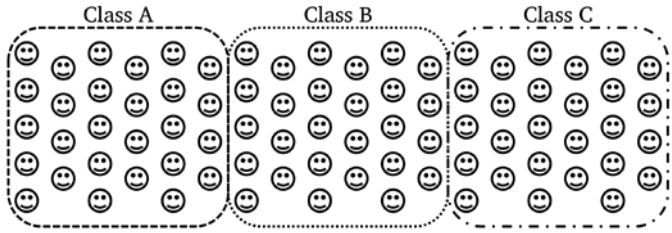

Figure 1. Visual representation of the hierarchical relationship between students and classes.

Research in linguistics, like the aforementioned education research, contains similar hierarchical relationships. Most linguistic studies involve multiple observations from each of several participants. These observations may thus be grouped according to the participants that

produced them. Snijders & Bosker (1999) refer to the model-building strategy in which multiple observations from each of multiple groups are included in the model as disaggregation. If such a model includes predictors relevant to both the micro and macro levels, which Gelman & Hill (2007: 27) refer to as the "[c]omplete pooling model", Snijders & Bosker (1999: 15) assert that "the miraculous multiplication of the number of units" results. If in a sociolinguistic study where a complete pooling model has been used a social factor such as age, gender, or socioeconomic class has been included, the significance of these predictors is based on the number of observations and not the number of speakers. Johnson (2009: 363) asserts that this results in an overestimation of the significance of such macro-level variables. This configuration has been the norm in variationist sociolinguistics studies and may still be found on occasion.

Most of the attention in the sociolinguistics literature has revolved around macro-level variables in disaggregation. With regard to micro-level variables, Snijders & Bosker (1999: 16) do not regard disaggregation in single-level models as incorrect, provided that it can be assumed that values within the macro-level grouping do not correlate. Thus, for example, in their analysis of the effects of word and speaker varying intercepts on a study of acoustic measurements of Colombian Spanish /s/, Gradoville *et al.* (2015) found that the inclusion of a speaker varying intercept only had an impact on the significance of internal linguistic predictors if the *p*-value of the predictor in the single-level model was already close to 0.05. If on the other hand observations within a macro-level grouping correlate with one another, the use of disaggregation in a single-level model may yield errors in the inferences made about the data (Snijders & Bosker 1999: 16). In linguistics studies, some examples where values of the response variable might correlate with the speaker/participant grouping include response time latencies in psycholinguistic studies (Baayen & Milin 2010), vocal tract size effects on acoustic measurements in sociophonetic studies (Drager & Hay 2012; File-Muriel *et al.* 2014), and studies of sociolinguistic variables where wide individual differences in variant use exist, particularly if both the response variable and a predictor are dependent on this relationship.

Gradoville *et al.* (2015: 109-110) raise the issue that certain corpus-based approaches to the study of priming of linguistic variants could be problematic if the researcher fails to account for the speaker grouping. In this paper, I assess the consequences of failing to address variation associated with the individual speaker in corpus-based approaches to priming where hypotheses related to priming are tested using a previous occurrence predictor. I show that models that fail to account for individ-

ual variation, at best, overstate the magnitude of effect and significance of the previous occurrence predictor and, at worst, could identify previous occurrence as significant when such a conclusion is unwarranted.

2. Priming/persistence phenomena in non-experimental linguistic production data

Corpus-based variationist research has frequently found that the previous occurrence of a linguistic variable may influence subsequent productions of the same variable. Let's suppose that the linguistic variable X has two possible variants, A and B. If a speaker produces (or is otherwise exposed to) variant A, the probability that the speaker will produce another variant A on the next occurrence of variable X increases. Likewise, if the same speaker produces (or is exposed to) variant B, the probability of a subsequent occurrence of B increases. This phenomenon has been variably referred to as priming, persistence, or formal parallelism. In this article, I adopt the term priming to refer to the phenomenon in question.

The remainder of this section provides an overview of major studies involving priming in non-experimental production data. Gries & Kootstra (2017) provide a more detailed discussion than is feasible to include in the present article. Among the first to observe that the sequential occurrence of the same linguistic variant in corpus data may not be by chance were Sankoff & Laberge (1978), who studied the phenomenon in three pronominal variables in Montreal French. They examined switch rates and the effect of syntagmatic proximity on the occurrences of the variables in question, finding that as two consecutive occurrences of the variable became more temporally or syntagmatically distant, a variant switch became more likely.

Another early study finding a dependency between sequential occurrences of a linguistic variable was Poplack's (1980) study of /s/ deletion in Puerto Rican Spanish. Many varieties of Spanish exhibit a weakening process where syllable- and word-final /s/ may be realized as [h] or phonetic zero. Since the nominal plural morpheme in Spanish contains /s/ (and usually nothing else) and Spanish noun phrases have plural agreement, the loss of /s/ can yield semantic ambiguity in an utterance. In her study, Poplack (1980: 63-64) found that the first element of a noun phrase most strongly favored retention, but subsequent occurrences of /s/ within the noun phrase were dependent on the immediately preceding occurrence of /s/: if the preceding occurrence was retained, the subsequent occurrence was likely to be retained; if it was

deleted, the following occurrence was likely to be deleted.

Weiner & Labov (1983) present another relatively early example of priming in their study of generalized active and agentless passive constructions in English. Their analysis demonstrated that a previous coreferential noun phrase in subject position favored the subsequent occurrence of the noun phrase in the same position. However, more important than this factor was whether a passive occurred in the previous five clauses, whether or not it was coreferential. If a passive occurred anywhere in the previous five clauses, the researchers found a strong likelihood of a subsequent occurrence of the passive.

Scherre & Naro (1991) examined priming-related phenomena in a variety of ways in both subject-verb and subject-predicate adjective agreement in the Portuguese of Rio de Janeiro. In the case of subject-verb agreement, Scherre & Naro (1991) looked at all semantically plural verbs. Those semantically plural verbs preceded by a same-subject verb lacking the plural morpheme were highly likely to also lack the plural morpheme. Those verbs preceded by a same-subject verb with the plural morpheme had a high degree of probability of occurrence with the plural morpheme. Likewise, within the same clause, if the last element of the subject NP lacked a plural morpheme, the verb was highly likely to lack a plural morpheme relative to the opposite condition. In their analysis of subject-predicate adjective agreement, Scherre & Naro (1991) found that plural morphemes in predicate adjectives were strongly favored when preceded by a predicate adjective with a plural morpheme and viceversa. Likewise, at the clausal level, the occurrence of the plural morpheme in a predicate adjective was strongly disfavored if either the last element of the subject NP or the semantically plural verb lacked the plural morpheme.

A linguistic variable frequently found to be subject to priming effects is Spanish variable subject expression (Cameron 1994; Flores-Ferrán 2002; Travis, Torres Cacoullos & Kidd 2017; Travis 2007). It is worthy of note that this finding has held regardless of the precise methodology employed by the researcher. In most cases, all grammatical subjects are included in the study, but some studies (Travis, Torres Cacoullos & Kidd 2017; Travis 2007) restrict study to a particular grammatical person. Additionally, what counts as a prime has varied from study to study. While in many cases only a previous coreferential subject is considered a prime, in some studies any previous occurrence of the referent, regardless of syntactic position, may count as a prime.

Szmrecsanyi (2006) studied priming in five different linguistic variables in English: comparison strategy choice, genitive choice, future marker choice, verb particle placement, and complementation strategy

choice. Although he found that the extent to which priming played a role in the phenomena examined varied (Szmrecsanyi 2006: 182), in all cases it factored into variant selection. Moreover, he found that the priming effect was strengthened when the prime and target shared more morphological material (i.e. the same verb lemma in particle placement and complementation strategy), a finding that corresponds to what Gries (2005) found for the same particle placement variable as well as for English ditransitive construction choice.

Although priming is often thought of as primarily a morphosyntactic phenomenon, Tamminga (2016) has found that it can also affect phonological variation. Tamminga (2016) studied variation in Philadelphia English *-ing* realization and final /t,d/ deletion. However, she found that priming only occurred when the prime and target were from the same morphological category, which she used as evidence from a modular theoretical perspective to suggest that the phenomenon could be used as a diagnosis of the nature of phenomena at the phonology-morphology interface.

Recent research (Rosemeyer 2015, Rosemeyer & Schwenter 2019) has suggested that priming may play a role in the preservation of moribund linguistic forms. Rosemeyer (2015) found that as the Spanish *be*-auxiliary became less frequent diachronically relative to the *have*-auxiliary, the strength of a *be*-auxiliary prime increased. Similarly, Rosemeyer & Schwenter (2019) found that in 20th century Spanish, the moribund *-se* imperfect subjunctive had a stronger priming effect than the more productive *-ra* imperfect subjunctive. Based on this evidence, the authors suggest priming may permit moribund linguistic forms to continue in a language.

3. The nature of the problem

Different speakers in a sociolinguistic corpus can be classified on the basis of the type of sequentiality of variant use that they exhibit. Figure 2 shows the logical possibilities with which hypothetical sociolinguistic variable X with variants A and B could occur for a given language user within an interview or other sample of speech. A speaker of Type 1 uses both variants A and B in a sequential manner, which can be used as evidence supporting the assertion that initial occurrences of one variant prime subsequent occurrences of that same variant. A speaker of Type 2 uses neither variant A nor variant B in sequence, which if such a speaker were to ever occur would serve as perfect counterevidence to the notion of priming as a factor in language variation. Probabilistically, a speaker of Type 2 is unlikely to be observed in data.

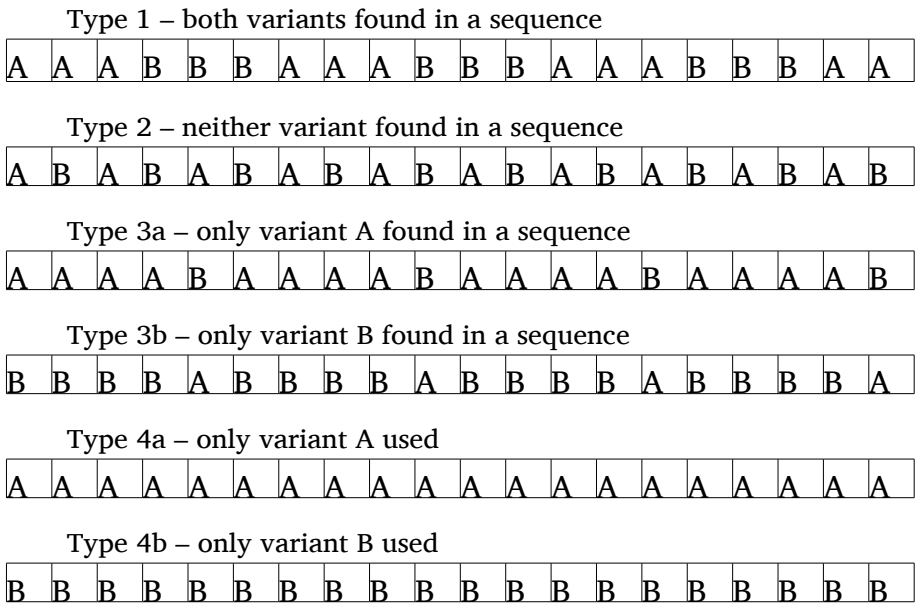

Figure 2. Sequential representations of logical possibilities of variant use.

Speakers of Types 3 and 4 are more troublesome for the analysis of priming as a factor in language variation. Although Type 3 speakers use both variants, they are only observed to produce one variant occurring in a sequence. Thus, although these speakers show sequential use of a variant, we cannot exclude the possibility that such sequential use is merely a consequence of the high rate at which the speaker uses one variant or the other (80% in Figure 2). Type 4 speakers exhibit no variation, using one variant or the other exclusively, at least over the course of the speech sample. Of course, sociolinguistic studies use speech samples, which are intended to represent the population of utterances that a language user is capable of employing in a given context. Thus, although a given speech sample may show use of only one variant (Type 4) or only one variant sequentially (Type 3), this does not mean that the sample exhaustively represents the speaker's repertoire, especially given that the use of linguistic variants is normally subject to constraints other than priming. Nevertheless, speakers of Types 3 and 4 provide no evidence that priming is occurring.

The problem occurs when a common methodology for testing hypotheses related to priming, namely the realization of the preceding

occurrence of the linguistic variable or certain related methods, is used. Although testing the hypothesis usually requires analysis of a similar variable, the probabilities of both the prime and the target occurrences of the variable are influenced by the speaker's overall predisposition to use a particular variant. In other words, different speakers will use different rates of a given linguistic variable. As such the probability of two sequential occurrences of a given variant is also dependent on the speaker that uttered them. When such data are analyzed using a single-level model, failing to account for the influence of the individual speaker, the importance of the previous occurrence variable, and thus priming, may be overestimated. At the extremes are speakers of Types 3 and 4 who, while using one variant sequentially, show no positive evidence of priming.

The present study has been guided by the following research questions:

1. To what extent can the concern over individual speaker variation in studies of priming be mitigated by studying individual speakers separately?
2. To what extent can the influence of priming be overstated when failing to account for individual speaker variation?
3. Could a research study obtain a significant result for previous occurrence when there is no evidence in individual speech samples for a priming effect?
4. If individual variation is minimal, is the concern about overestimation of priming mitigated?

4. Method

4.1. Linguistic variable and data

The linguistic variable to be used to evaluate the research questions is the form variation surrounding the Portuguese preposition *para* 'to, for, in order to'. In informal spoken Portuguese, this word, represented in standard orthography as *para*, may be reduced to *pra* or *pa*, the former of which is regarded as the default variant (Perini 2002, Thomas 1969). Reduced variants are also subject to contraction with frequently co-occurring following words (Ilari *et al.* 2008; Kewitz 2006). This variable has been subject to study regarding both social variation (Felgueiras 1993; Ferreira 2014; Gradoville 2015; Lucena 2001; Maya 2004; Silva 2010; Vellasco 1998) and internal linguistic constraints (Felgueiras 1993; Gradoville 2017; Huback 2012). Two main findings from previous studies of Portuguese *para* form variation make this variable a logical

test vehicle for the research questions. First, Felgueiras (1993: 87) found a fairly strong effect for her previous occurrence predictor (Varbrul range = 45), suggesting a very strong priming effect. Second, Gradoville (2015) observed a notably wide range of individual speaker behavior with individual reduction rates varying between 5% and 100%.

The corpus of study comes from the material produced for the Corpus Português Oral Culto de Fortaleza (Educated Oral Portuguese of Fortaleza Corpus, Monteiro 1993), which is a 500,000-word spoken corpus from the 1990s that includes speech from 75 different individual speakers in one of three speech styles: dialogues between two people that know one another (*diálogos entre dois informantes*, 26 speakers), dialogues between an informant and a researcher (*diálogos entre um informante e um documentador*, 30 speakers), and classes and formal lectures (*elocuções formais*, 19 speakers). All tokens of unreduced *para*, reduced rhotic *pra*, and reduced non-rhotic *pa* as well as any contractions involving these forms were exhaustively extracted from the corpus transcriptions using the techniques described by Gries (2009). Tokens from individuals other than the research subjects (interviewers, students, others) were excluded from study, yielding a total of 4749 observations.

4.2. Variables

4.2.1. Response variable

The three variants, namely *para*, *pra*, and *pa* were reduced to a binary response variable. Specifically, the two reduced forms *pra* and *pa* were together placed in opposition to unreduced *para*. This division is motivated by two different issues. First, both *pra* and *pa* are reduced temporally relative to their unreduced counterpart *para*. Second, the opposition between rhotic *pra* and non-rhotic *pa* may be a result of a complex onset simplification process for which unreduced *para* would be ineligible to participate. When it is necessary to refer to the reduced variants collectively, they will henceforth be referred to as *p(r)a*. For the purposes of this study, contractions of *p(r)a* with following words are treated as the same as *p(r)a*. Thus, for example, the contraction *pro*, a result of the fusion of the reduced form *pra* with the definite article *o*, is treated as another token of *pra*, the contraction process being considered a separate issue.

4.2.2. Predictor variables

The predictor of principal interest in the present study is that of the variant produced in the previous occurrence of the *para* variable (Previous Occurrence). While some studies of priming have limited the

temporal distance of the previous occurrence predictor in one way or another, this is unnecessary for the purpose of testing the hypotheses in question. Moreover, any temporal distance division would be arbitrary. This predictor thus had three possible values: unreduced *para* (see (1a)), reduced rhotic *pra* (see (1b)) or non-rhotic *pa* (see (1c)), and the first occurrence in the sample. The reference level of regression models was set at unreduced *para*.

(1) a. *outr-o* *foi* para *Minas Gerais…* *outr-o* *foi* pr-*o* *Rio…*
other-M go.PST.3SG *para* Minas Gerais other-M go.PST.3SG *para*-ART.M Rio
'Another went to Minas Gerais… Another went to Rio.' (Interview 23)

b. *então* *a* *idéi-a* *se-ria* *junt-ar* *a-s* *duas* *forç-a-s* *polític-a* pra
so ART.F idea-F be-COND.3SG join-INF ART.F-PL two.F force-F-PL political-F *para*
v-er *se* *arrast-a* pra */colá* (Dialogue 45)
see-INF if pull-PRS.3SG *para* there
'So, the idea would be to the two political forces to see if it pulls (it) there.'

c. *vai* p*/* *o-s* *cant-o* *marc-a* *entrevist-a* *vai…* pra
go.PRS.3SG *para* ART.M-PL place-M set-PRS.3SG interview-F go.PRS.3SG *para*
rá::dio
radio
'(He/she) goes places… sets an interview… goes to the radio…' (Dialogue 28)

The remaining predictors were included on the basis of findings of their importance in previous studies. Of interest in the present study are the internal linguistic constraints that have been found to affect *para* form variation. First, previous studies have found the frequency of co-occurrence of *para* with a flanking word to affect *para* reduction (Gradoville 2017, Huback 2012): *para* reduces to *p(r)a* more often in high frequency sequences. While this has been found to be true of both the *para* + WORD and WORD + *para* strings (Gradoville 2017), given the greater importance of the *para* + WORD frequency predictor as well as problematic nature of including both predictors in the same regression,[2] the present study will only include the *para* + WORD frequency predictor (*para* + WORD Frequency). The frequency of *para* + WORD bigrams was determined within the corpus of study since it is the largest spoken corpus available for the variety of Portuguese in question. A logarithmic transformation was applied to frequency values to account for the exponential nature of the frequency predictor (File-Muriel 2010). Moreover, the logarithmically-transformed frequency values were *z*-scored due especially to the relative intolerance of mixed-effects models of unstandardized interval values. In

the multifactorial models, certain groups of tokens involve missing data (see Gradoville 2017: 98-99 for more detailed discussion of this issue): (i) tokens in utterances truncated immediately following *para* or in the word thereafter (see (2a) and (2b)) and (ii) tokens of *para* followed by a feminine singular NP that begins with or potentially could begin with the definite article *a* (see (3)), since due to the aforementioned contraction process it is not always possible to empirically verify whether the article is present, thereby making it impossible to accurately classify these bigrams. Since these groups of data have to be treated as having missing values, the average value of *para* + WORD bigram frequency (0, after the previously mentioned transformations) was imputed in these groups for this variable, following Gelman & Hill (2007: Ch. 25).

(2) a. *vai* *d-ar* pra… *pra* *que* *eu* *possa* *fal-ar*
go.PRS.3SG give-INF *para* *para* that I be.able.to.PRS.SBJV.1SG talk-INF
'It will work out so that I can talk.' (Dialogue 39)

b. *e* *vai* *agora hav-er* *u::m-a…* pra *je/* *nova jerusa/ jerusalém* *né?*
and go.PRS.3SG now there.be-INF ART-F *para* Nova Jerusalém right?
'And now there's going to be one for Nova Jerusalém, right?' (Interview 13)

(3) *para* *o* *bem* *d-a* *famíli-a* *brasileir-a…* pr(-)a *formação…* *de*
para ART.M sake of-ART.F family-F Brazilian-F *para*(-ART.F?) education of
um-a *nov-a* *geração…* (Interview 44)
ART-F new-F generation
'For the sake of the Brazilian family… for the education of a new generation.'

Another predictor included in the present study accounts for certain exceptional groupings known to occur in Portuguese *para* reduction (Gradoville 2017). Specifically, although the effect of *para* + WORD bigram frequency is robust, following definite and indefinite articles as well as the adverbial conjunction *para que* 'so that' have exceptionally low reduction rates that fall well outside the normal range for other bigrams in the data set. This predictor (Grouping) thus accounts for the variance associated with these groups. The possible values for this predictor are: definite article (see (1a) and preceding occurrence in (1c), indefinite article (see (4)), *para que* (see (2a)), and other. The reference level for this predictor is other.

(4) *imagin-e* pr/ *um-a* *criança…* */tá?* (Interview 21)
imagine-IMP *para* ART-F child okay?
'Imagine for a child, okay?'

Preceding context has also been found to play a role in the reduction of Portuguese *para*. Previous studies (Felgueiras 1993, Gradoville 2017) have found that the reduction of *para* is disfavored when following a pause. In order to fully account for this variable (Preceding Context), the predictor was coded for the following values: unstressed syllable (see (1b) and (4)), stressed syllable (see (1a) and (2a)), pause (see (3)), truncation (see second occurrence in (2a)), sentence initial (see (5)), and unclear. The reference level in this case is unstressed syllable.

(5) pa/ *aprend-er* *a* *lín::gu-a* *né?…* (Dialogue 47)
para learn-INF ART.F language-F right?
'To learn the language, right?'

While other internal constraints have been tested for their effect on the reduction of Portuguese *para*, on the basis of previous studies, the greatest degree of confidence can be placed in the role of the aforementioned predictors. The focus of the present study, furthermore, is on properly accounting for priming effects and not on testing additional constraints on *para* reduction.

4.3. Analysis

Analysis of the data included a variety of techniques, all of which used the R programming language (R Core Team 2017). The first set of techniques focused on understanding the behavior of Previous Occurrence. A general Fisher's exact test including all data testing the overall relationship between previous occurrence and reduction was performed. Thereafter, individual Fisher's exact tests were carried out for each individual speaker to see whether significant priming effects could be identified for each speaker. Speakers were also classified according to the type of sequential pattern of variant use that they exhibited (see discussion of Figure 2). Finally, appropriate plots were generated in order to visualize the data.

The second set of techniques involved fitting a series of generalized linear (mixed-effects) models in order to test the effect of the previous occurrence predictor under various circumstances. Three types of models were fit. First, a single-level generalized linear model without a fixed effect for speaker was fit in order to establish a base line that is here considered to be equivalent to the expected behavior of the Varbrul family of programs under similar circumstances. Second, a single-level generalized linear model with a fixed effect for speaker was fit in order

to account for both the previous occurrence predictor and the varying reduction rates of individual speakers. Third, a multi-level generalized linear mixed-effects model with a varying intercept for speaker was fit using the lme4 (Bates *et al.* 2015) and the optimx (Nash & Varadhan 2011) packages in order to achieve the same end as the second model.[3] While mixed-effects models have generally been prescribed to account for speaker effects in sociolinguistic data (Johnson 2009), Paolillo (2013) has argued that it may be more appropriate to model speaker variation using fixed effects when speakers have not been selected randomly. Since the purpose of the present study is to show the necessity of accounting for speaker effects in studies of priming and not to advocate for the precise manner in which this is achieved, both types of models have been fit.[4]

Four different sets of these models were fit. The first set of models includes data from all speakers in order to see the overall effect of individual variation on estimates of priming in *para* reduction. The second and third sets of models include only those speakers with reduction rates in the ranges 75%-95% and 75%-85%, respectively, to determine whether the overestimation of priming effects occurs when individual variation is controlled.[5] The fourth set of models includes only those speakers of Type 3 in order to assess whether due to wide speaker variation it would be possible to obtain a significant result for Previous Occurrence where one is not warranted.

5. Results

Results of the analysis of the data on *para* reduction in the spoken Portuguese of Fortaleza show that 82.9% (3935/4749) of the occurrences in the data were reduced to *p(r)a*, indicating that reduced variants represent a sizable majority of use. Table 1 shows the overall distribution of these tokens according to Previous Occurrence. As we can see, while the first occurrence of *para* by a speaker has a very similar reduction rate to the overall mean, a preceding unreduced *para* yields a much lower reduction rate (49.3%) than the other conditions. Conversely, *para* reduces to *p(r)a* much more frequently when it is preceded by *pra* or *pa* (89.9%). This distribution between preceding *para* and preceding *p(r)a* is statistically significant according to the Fisher's exact test (Odds Ratio $= 9.152515$; $p < 2.2 \times 10^{-16}$). However, these massive observed rate differences will be shown to be partially a consequence of large individual differences in variant usage.

	Reduction Rate	N
First Occurrence	76.0%	75
para	49.3%	799
p(r)a	89.9%	3875
Overall	82.9%	4749

Table 1. Reduction rates of *para* according to previous occurrence

The following sections address first individual variation and its relationship to Previous Occurrence and second the impact of this on multifactorial regression models incorporating Previous Occurrence.

5.1. Individual variation

Individual variation plays a major role in the wide distributional differences observed in Table 1. Each speaker was classified according to the pattern of sequential variant use observed in the data. Similar Fisher's exact tests were also carried out for each individual speaker to determine whether individually Previous Occurrence played a significant role for that speaker. The results of these analyses have been combined into one pie chart (see Figure 3 below). Fisher's exact tests are inappropriate for speakers of Type 4 (exclusive use of one variant) and they nearly always yield a *p*-value of 1 for speakers of Type 3 (only one variant observed to occur sequentially). As a consequence, only Type 1 speakers (both variants occur sequentially) have been subdivided according to the results of the Fisher's exact test.

As we can see in Figure 3, slightly more than half of the 75 speakers (39 speakers; 52.0%) belong to Type 1, meaning that while more than half of the speakers show potential evidence for priming effects, nearly half of the 75 do not. Some 22 speakers (29.3%), although using both variants, belong to Type 3, meaning that they use only one variant consecutively. A further 14 speakers (18.7%) use the reduced *p(r)a* variant exclusively. No speaker in the sample used unreduced *para* exclusively.

Although more than half of the speakers in the sample use both variants in sequence, thereby showing potential positive evidence for priming effects, the actual number of speakers to show a statistically significant difference on the basis of Previous Occurrence is actually quite small. Of the 39 speakers of Type 1, we can see in Figure 3 that only 10 speakers (13.3% of all speakers in the sample) had a Fisher's exact test with a *p*-value below 0.05. As a consequence, on an individual level, we can only affirm that priming matters for two of every 15 speakers in the sample.

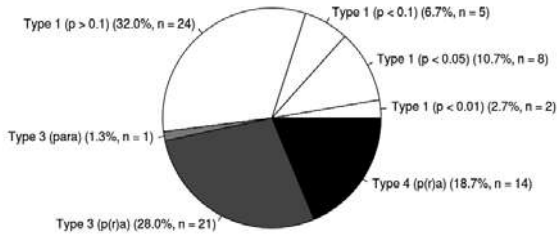

Figure 3. Pie chart of speakers according to sequential use of variants and Fisher's exact tests for relationship between reduction and Previous Occurrence

As previously discussed, the widely varying rates in Table 1 are largely a consequence of individual variation. Figure 4 is a kernal density curve of the reduction rates of individual speakers. As we can see, the overwhelming majority of the speakers are concentrated at the high end of the reduction scale. More than half of the speakers in the sample (54.7%; 41/75) have reduction rates in excess of 90%. As previously discussed, 14 of these speakers reduce *para* to *p(r)a* categorically. Although speaker behavior is highly concentrated at one end of the reduction spectrum, it is also apparent that individual speakers are capable of having widely varying reduction rates. As we can see in Figure 4, speakers in the sample regularly have reduction rates as low as 31.9%. One exceptional speaker even has a reduction rate of 7.7%. It is from these speakers with very low reduction rates that in the overall sample the reduction rate of tokens preceded by unreduced *para* can be 49.3% since speakers with low reduction rates are likely to have large numbers of unreduced *para* produced sequentially. There is a strong relationship between speaker type and reduction rate. As we can see in Figure 5, Type 1 speakers exhibit reduction rates throughout the range. There is only one speaker of Type 3a and so naturally no variation. Speakers of Type 3b exhibit variation, but tend to have reduction rates above 90%. Speakers of Type 4b, by definition, all have reduction rates of 100%.

A secondary issue affecting the results in Figure 3 is the number of observations per speaker, which varies dramatically. While speakers in the sample produce an average of 63.32 observations with the median being 56, production ranges between 6 and 189 observations with an interquartile range between 34.5 and 82. Speakers with a smaller number of observations are less likely to be of Type 1 or have a significant Fisher's exact test, although the more dramatic effect seems to be the reduction rate issue. Table 2 is a generalized linear model of predictors of a speaker

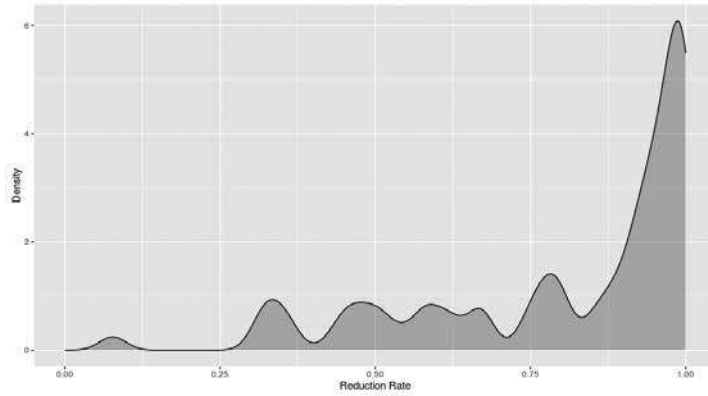

Figure 4. Kernal density curve of speakers' reduction rates

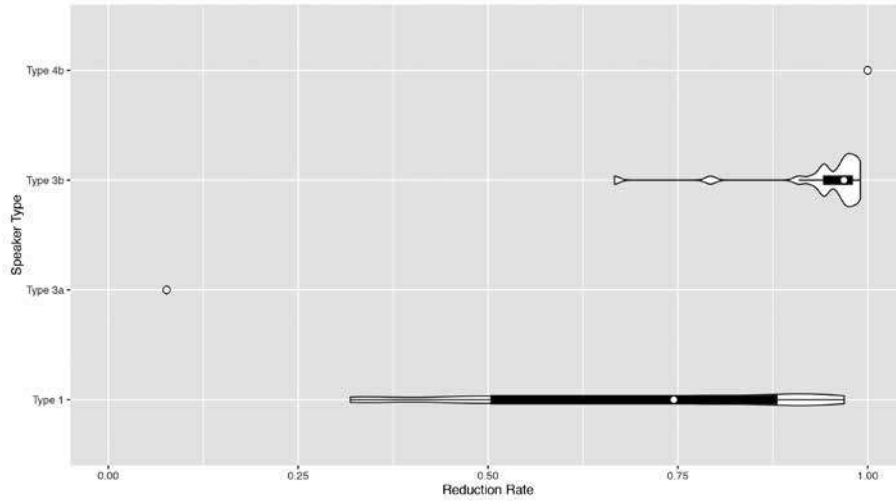

Figure 5. Violin and box plot of speakers' reduction rates by speaker type

being Type 1 or not. Tested predictors were number of tokens and distance from categoricity, which is defined as the absolute value of the difference between the speaker's rate of variant use and the nearest categorical value (0% or 100%). For example, speakers with reduction rates of 40% and 60% would both be 40% from categorical. Both values were logarithmically transformed and *z*-scored so that they would be on the same scale. We can see in Table 2 that the effect of distance from categorical is far more robust (estimate = 4.6355) than number of tokens (estimate =

1.4686). In fact, number of tokens is not significant unless the logarithmic transformation is applied. In sum, although the probability of a speaker being Type 1 increases with both greater distance from categoricity and greater number of tokens, the former effect is much stronger than the latter.

	ESTIMATE	*p*-VALUE	
INTERCEPT	-0.1339	0.753047	
DISTANCE FROM CATEGORICAL	4.6355	< 0.001	***
TOKENS	1.4686	0.012339	*

Table 2. Generalized linear model of predictors of a speaker being Type 1

5.2. Multifactorial regression models

This section presents the results of the multifactorial regression models. The first set of regression models includes the data from all of the speakers. The second set of regression models includes data from speakers observed to reduce *para* 75% to 95% of the time. The third set of regression models includes data from speakers observed to reduce *para* 75% to 85% of the time. The fourth set of regression models includes data from Type 3 speakers.

5.2.1. Complete data set

Table 3 presents the models of the complete data set side by side in the interest both of space and usability. This table follows a common format that will later be used in Tables 4, 5, and 6. The first column presents the predictors and, in the case of factors, each individual value of the factor. The results of the regular generalized linear model (GLM) are presented first, those for the GLM with the fixed effect for speaker are presented second, and those for the generalized linear mixed-effects model with the varying intercept for speaker last. For each model, the estimates, *p*-values, and a standard shorthand for significance level (***: $p \leq 0.001$; **: $p \leq 0.01$; *: $p \leq 0.05$; .: $p \leq 0.10$) are presented. In the interest of space, standard errors and *z*-values are not presented, but will be discussed where relevant. Although the GLM with the fixed effect for speaker has estimates for speaker, these are not presented since speaker is acting merely as a moderator variable and the fitted estimates are not of interest. Additionally, the variance and standard deviation for the speaker varying intercept are not presented, but will be discussed where relevant.

	GLM			GLM (speaker fixed effect)			GLMM (speaker intercept)		
	Estimate	p-value		Estimate	p-value		Estimate	p-value	
Intercept	0.75834	< 0.001	***	18.89620	0.9836		2.63940	< 0.001	***
Previous Occurrence									
Unreduced *para*	** REFERENCE LEVEL **			** REFERENCE LEVEL **			** REFERENCE LEVEL **		
Reduced *p(r)a*	2.28369	< 0.001	***	0.78519	< 0.001	***	0.90243	< 0.001	***
First Occurrence	1.14134	< 0.001	***	-0.10983	0.7709		-0.01805	0.96067	
para + WORD Freq.	0.46809	< 0.001	***	0.43972	< 0.001	***	0.44674	< 0.001	***
Grouping									
Other	** REFERENCE LEVEL **			** REFERENCE LEVEL **			** REFERENCE LEVEL **		
Definite Article	-2.79220	< 0.001	***	-3.37024	< 0.001	***	-3.31849	< 0.001	***
Indefinite Article	-1.62994	< 0.001	***	-2.02037	< 0.001	***	-1.99856	< 0.001	***
Feminine Singular NP	-1.01290	< 0.001	***	-1.03393	< 0.001	***	-1.04073	< 0.001	***
para que	-1.49415	< 0.001	***	-1.35702	< 0.001	***	-1.38401	< 0.001	***
Preceding Context									
Unstressed Syllable	** REFERENCE LEVEL **			** REFERENCE LEVEL **			** REFERENCE LEVEL **		
Stressed Syllable	0.33252	0.003205	**	0.32297	0.0135	*	0.33827	0.00867	**
Pause	-0.85036	< 0.001	***	-0.83363	< 0.001	***	-0.83658	< 0.001	***
Turn Initial	0.06661	0.799296		-0.07809	0.7893		-0.06225	0.82990	
Truncation	0.13704	0.545314		0.08344	0.7370		0.08179	0.73917	
Unclear	-0.03339	0.962124		-0.48574	0.5728		-0.38370	0.64491	

Table 3. Multifactorial regression models of predictors of *para* reduction (complete data set)

With respect to Previous Occurrence in Table 3, in the single-level GLM that does not account for the speakers in the sample, relative to the reference level where unreduced *para* precedes the token in question, a previous reduced *p(r)a* variant appears to yield significantly more reduction to *p(r)a* (estimate = 2.28369). In the GLM with the speaker fixed effect what is most notable is the reduction of the magnitude of the estimate to 0.78519, which is 34.4% of the magnitude without the speaker fixed effect. In the mixed-effects model, the reduction in the magnitude of the estimate for previous reduced *p(r)a* is not as drastic (0.90243; 39.5%), but it is much closer to the estimate in the GLM with the speaker fixed effect.[6] In other words, whether or not the speaker grouping in the data is accounted for strongly influences the strength of Previous Occurrence.

With respect to the behavior of the first occurrence, in the single-level GLM initial occurrences are significantly more likely to yield reduction (estimate = 1.14134) than when the token is preceded by unreduced *para*. However, this estimate loses significance the moment that individual speakers are taken into consideration, be it with a fixed effect or varying intercept for speaker.

The fitted intercepts for the models in Table 3 are worthy of note. Although they are not expected to be the same, the intercept in the GLM with a fixed effect for speaker has an inflated estimate (18.89620) that is not significant because of its much larger standard error (917.8).

Regarding the behavior of the other predictors in Table 3, although the inclusion of a fixed effect or varying intercept for speaker has no practical consequences in terms of changes in significance, some estimates change notably. The estimate for *para que* in Grouping, for example, notably reduces in magnitude. The magnitudes of the estimates for definite and indefinite articles, on the other hand, increase. It is worthy of note that none of the models in Table 3 account for WORD + *para* and *para* + WORD bigrams, which results in much lower *p*-values for Grouping and Preceding Context relative than those reported in Gradoville (2017).

5.2.2. Speakers with reduction rates between 75% and 95%

The second set of regression models includes only speakers with reduction rates between 75% and 95%. These models include 1579 observations (33.2% of tokens) from 23 speakers (30.7% of sample). Table 4 presents these models in the same format as Table 3 presented those of the complete data set.

	GLM			GLM (speaker fixed effect)			GLMM (speaker intercept)		
	Estimate	*p*-value		Estimate	*p*-value		Estimate	*p*-value	
Intercept	1.8965	< 0.001	***	2.75598	< 0.001	***	2.1799	< 0.001	***
Previous Occurrence									
Unreduced *para*	** REFERENCE LEVEL **			** REFERENCE LEVEL **			** REFERENCE LEVEL **		
Reduced *p(r)a*	1.5942	< 0.001	***	1.27937	< 0.001	***	1.4071	< 0.001	***
First Occurrence	0.6831	0.306541		0.39418	0.548183		0.5092	0.440707	
para + WORD Freq.	0.3829	0.004926	**	0.45168	0.001306	**	0.4235	0.002299	**
Grouping									
Other	** REFERENCE LEVEL **			** REFERENCE LEVEL **			** REFERENCE LEVEL **		
Definite Article	-3.7025	< 0.001	***	-4.00945	< 0.001	***	-3.8672	< 0.001	***
Indefinite Article	-1.7400	< 0.001	***	-1.90752	< 0.001	***	-1.8271	< 0.001	***
Feminine Singular NP	-1.3806	< 0.001	***	-1.28473	< 0.001	***	-1.3304	< 0.001	***
para que	-1.7169	< 0.001	***	-1.86412	< 0.001	***	-1.7906	< 0.001	***
Preceding Context									
Unstressed Syllable	** REFERENCE LEVEL **			** REFERENCE LEVEL **			** REFERENCE LEVEL **		
Stressed Syllable	0.1108	0.608703		0.09428	0.679268		0.1172	0.599409	
Pause	-0.4725	0.060048	.	-0.53931	0.041738	*	-0.5109	0.048434	*
Turn Initial	-0.2171	0.634016		-0.48003	0.291206		-0.3637	0.424075	
Truncation	0.4326	0.335461		0.39060	0.390099		0.3909	0.386204	
Unclear	-1.3491	0.220005		-1.35028	0.255725		-1.4121	0.215588	

Table 4. Multifactorial regression models of predictors of *para* reduction (only speakers with reduction rate between 75% and 95%)

With respect to Previous Occurrence, in the single-level GLM in Table 4 the estimate of 1.5942 for a preceding reduced *p(r)a* is much smaller in magnitude than that of the equivalent model in Table 3 (2.28369; 69.8% of the estimate for the complete data set). The estimate in the GLM with a speaker fixed effect has a smaller magnitude estimate (1.27937) than the single-level GLM (80.3% of the GLM estimate). The mixed-effects estimate (1.4071) is between the single-level GLM and GLM speaker estimates, although closer to the latter. Thus, although the influence of the speaker effect is more moderate when interspeaker variation has been reduced, individual variation still functions to inflate the effect of Previous Occurrence when speakers are not accounted for. Interestingly, the GLM with speaker and mixed-effects estimates in Table 4 are much higher in magnitude than the equivalent estimates in Table 3 (162.9% and 155.9% of the estimates, respectively), possibly a consequence of the inclusion of Type 4 speakers in the models in Table 3. There is no significant difference between a preceding *para* and a first occurrence in any of the models in Table 4.

Regarding the other effects in Table 4, while the intercepts are not the same in the three models, there are no inflated estimates like there were for the GLM speaker model in Table 3. The effects of Grouping and *para* + WORD Frequency in Table 4 are largely the same as in Table 3. On the other hand, the significant effect for preceding stressed syllables (compared to unstressed syllables) in Table 3 is gone in Table 4. Finally, the significant effect for a preceding pause in Table 3 only attains significance in the GLM speaker model and the mixed-effects model in Table 4.

5.2.3. Speakers with reduction rates between 75% and 85%

The third set of regression models includes only speakers with reduction rates between 75% and 85%, thereby showing the behavior of Previous Occurrence when individual variation is minimal. These models include 540 observations (11.4% of tokens) from 8 speakers (10.7% of sample). Table 5 presents these models in the same format as Tables 3 and 4.

In this case, the effect of a preceding reduced *p(r)a* (compared to a preceding unreduced *para*) is basically the same in the three models (estimates = 1.215, 1.2056, 1.20262). Although these estimates in Table 5 are lower in magnitude than those in Table 4, they are still much higher than the equivalent estimates in Table 3 in the models where speaker is accounted for. As was the case in the models in Table 4, there is no significant difference in any of the models in Table 5

	GLM			GLM (speaker fixed effect)			GLMM (speaker intercept)		
	Estimate	*p*-value		Estimate	*p*-value		Estimate	*p*-value	
Intercept	1.702	< 0.001	***	3.2826	< 0.001	***	1.75580	< 0.001	***
Previous Occurrence									
Unreduced *para*	** REFERENCE LEVEL **			** REFERENCE LEVEL **			** REFERENCE LEVEL **		
Reduced *p(r)a*	1.215	< 0.001	***	1.2056	< 0.001	***	1.20262	< 0.001	***
First Occurrence	0.046	0.96127		0.1409	0.88457		0.05141	0.95637	
para + WORD Freq.	0.499	0.01376	*	0.5268	0.01013	*	0.50384	0.01335	*
Grouping									
Other	** REFERENCE LEVEL **			** REFERENCE LEVEL **			** REFERENCE LEVEL **		
Definite Article	-4.511	< 0.001	***	-4.9212	< 0.001	***	-4.60933	< 0.001	***
Indefinite Article	-1.156	0.17233		-0.9891	0.25963		-1.10222	0.19824	
Feminine Singular NP	-1.665	< 0.001	***	-1.6413	< 0.001	***	-1.66153	< 0.001	***
para que	-1.657	0.00912	**	-1.9113	0.00425	**	-1.70883	0.00783	**
Preceding Context									
Unstressed Syllable	** REFERENCE LEVEL **			** REFERENCE LEVEL **			** REFERENCE LEVEL **		
Stressed Syllable	-0.401	0.21904		0.3458	0.30412		0.39181	0.23448	
Pause	-0.205	0.57456		-0.4118	0.28356		-0.27693	0.45417	
Turn Initial	0.000	0.99990		-0.1519	0.84408		-0.02382	0.97588	
Truncation	2.498	0.04428	*	2.2425	0.06615	.	2.38688	0.05284	.
Unclear	-1.753	0.19302		-2.1307	0.16358		-1.82672	0.18162	

Table 5. Multifactorial regression models of predictors of *para* reduction (only speakers with reduction rate between 75% and 85%)

between a first occurrence and a previous unreduced *para*.

Regarding other estimates in the models in Table 5, there again is some variability in the estimates of the intercepts, but no inflated estimates of the sort found in Table 3. The estimates for *para* + WORD Frequency in Table 5 are higher in magnitude than in either Table 3 or 6 and, likely due to the small sample size, *p*-values are much higher than in the other models. The small sample size has likely had a number of other effects. In Grouping, although the effect of the definite article continues to be quite robust, having the highest magnitude estimates in Table 5 compared to any other models (-4.511, -4.9212, -4.60933), there is no significant difference between *para* + INDEFINITE ARTICLE and other sequences in these models. The estimates of feminine singular NPs and *para que* in Table 5 follow the pattern of Table 4, having higher magnitude effects than the models in Table 3.

5.2.4. Type 3 speakers

The fourth set of regression models includes only Type 3 speakers, speakers that only produce one variant sequentially. These models include 1536 observations (32.3% of tokens) from 22 speakers (29.3% of the sample). Table 6 presents these models in the same format as Tables 3, 4, and 5.

In Table 6, which only includes speakers shown to only produce one of the variants sequentially, the single-level GLM assigns the largest magnitude estimate to a previous reduced *p(r)a* (relative to a preceding unreduced *para*) of any model we have seen so far (estimate = 2.5540). In the GLM that includes the speaker fixed effect, the estimate for previous reduced *p(r)a* loses significance, but its magnitude is inflated (-16.2043; standard error = 739.71072). While the mixed-effects model in Table 6 does not have the inflated estimate for previous reduced *p(r)a*, the estimate has rightly lost significance. Regarding the first occurrence, there is once again no significant difference from a previous unreduced *para*, although it is worth mentioning that the estimate in the GLM speaker model is also highly inflated (-17.1019; standard error = 739.71123).

With respect to the other estimates in Table 6, although some patterns remain from previous models, some patterns differ. As occurred in Table 3, the magnitude of the estimate of the intercept in the GLM speaker model is inflated. In the single-level GLM, *para* + WORD Frequency has its highest magnitude estimate of any model in the study (0.5613). However, once individual speakers are accounted for, the *p*-value increases and, in the case of the GLM with a speaker fixed effect,

	GLM			GLM (speaker fixed effect)			GLMM (speaker intercept)		
	Estimate	*p*-value		Estimate	*p*-value		Estimate	*p*-value	
Intercept	1.7882	< 0.001	***	21.8116	0.97648		5.0889	< 0.001	***
Previous Occurrence									
Unreduced *para*	** REFERENCE LEVEL **			** REFERENCE LEVEL **			** REFERENCE LEVEL **		
Reduced *p(r)a*	2.5540	< 0.001	***	-16.2043	0.98252		-0.7996	0.2974	
First Occurrence	0.6890	0.33933		-17.1019	0.98155		-1.8147	0.1086	
para + WORD Freq.	0.5613	0.00254	**	0.4811	0.05446	.	0.4922	0.0333	*
Grouping									
Other	** REFERENCE LEVEL **			** REFERENCE LEVEL **			** REFERENCE LEVEL **		
Definite Article	-3.4101	< 0.001	***	-3.7939	< 0.001	***	-3.6711	< 0.001	***
Indefinite Article	-1.7869	0.01598	*	-2.3592	0.03514	*	-2.0020	0.0533	.
Feminine Singular NP	-0.9605	0.03932	*	-1.4538	0.00856	**	-1.2999	0.0149	*
para que	-1.9758	0.00879	**	-3.3098	< 0.001	***	-3.0782	< 0.001	***
Preceding Context									
Unstressed Syllable	** REFERENCE LEVEL **			** REFERENCE LEVEL **			** REFERENCE LEVEL **		
Stressed Syllable	0.3719	0.24333		0.4352	0.26100		0.4563	0.2179	
Pause	-0.7077	0.03039	*	-0.6449	0.13704		-0.5814	0.1490	
Turn Initial	0.4454	0.56999		-0.0514	0.95113		0.1411	0.8644	
Truncation	0.6327	0.35528		0.0651	0.92519		0.1682	0.8043	
Unclear	-1.0323	0.42577		-1.3702	0.33682		-1.3856	0.3040	

Table 6. Multifactorial regression models of predictors of *para* reduction (only Type 3 speakers)

it loses significance. Regarding Grouping, aside from the estimates for *para* + INDEFINITE ARTICLE, the other estimates for the variable in Table 6 are characterized by high *p*-values, although still significant with the exception of *para* + DEFINITE ARTICLE in the mixed-effects model. Finally, the only significant effect in Table 6 for Preceding Context is the contrast between a preceding pause (estimate = -0.7077) and the unstressed syllable reference level in the single-level GLM, an effect lost in the models that account for speaker.

6. Discussion

The results of this study inform our understanding of the relationship between individual variation and priming effects in corpus-based studies. Samples from different speakers vary in the extent to which they support the existence of a priming effect. While some speakers in a sample of data may show no variation at all, other speakers may produce only one variant sequentially, which is generally a consequence of a high rate of use of one variant, and such sequential use should not be considered evidence in favor of priming effects. Evidence in favor of priming occurs when a speaker uses each variant sequentially. Switch rates, however, are to a great extent a function of rates of variant use. Figure 6 is a scatterplot of speakers in the sample according to switch rates and distance from categoricity, as previously defined in section 5.1. Different plotting symbols have been used for each speaker type. Plotting symbols have been rendered transparent so that, when they overlap, it is apparent that multiple speakers are represented at the coordinates in question. As we can see in Figure 6, there is a strong relationship between switch rate and distance from categoricity. Speakers with nearly categorical variant use naturally have a very low switch rate, which means that they use the same variant in long sequences. Speakers with moderate reduction rates, being far from categoricity, have higher switch rates, although the pattern is much less uniform. While the trend line in Figure 6 is relatively consistent, when speakers are between 40% and 50% from categorical, the trend line is basically flat, which is an indication that, even though speakers use both variants productively, on average switch rates in this range are constant.

The strong relationship between variant use rate and switch rate has implications for how corpus-based studies of priming must be approached. The results from the analyses of the entire data set in Table 3 show that, if a linguistic variable is subject to wide variation in individual rates of variant use and if those speakers are not accounted for

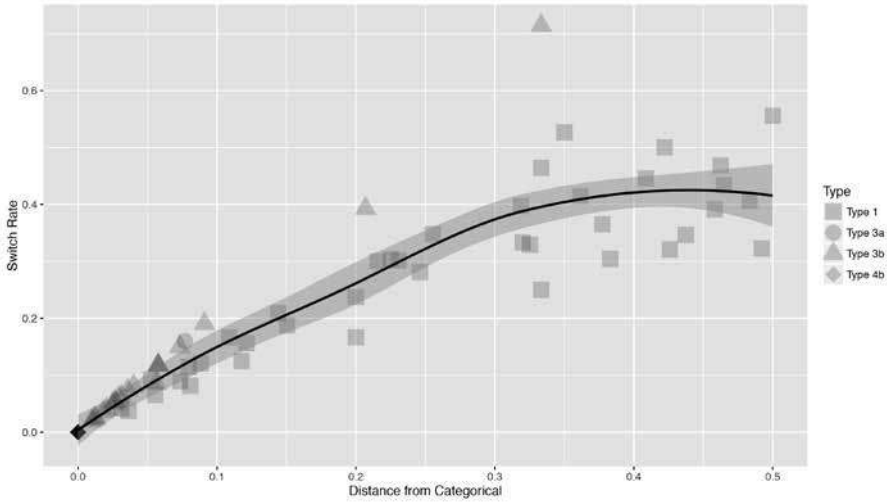

Figure 6. Scatter plot of speakers according to distance from categoricity (*x*-axis) and switch rate (*y*-axis)

in the statistical model, the magnitude of effect of Previous Occurrence may be massively overstated. The analyses in Table 4, where the range of individual variant use was restricted to 20%, show that even under these circumstances the effect of Previous Occurrence can be overstated, although the differences between these models were more muted than those found in Table 3. It was only when the range of individual variant use was restricted to 10% in Table 5 that the behavior of Previous Occurrence was basically uniform between the three models. As a consequence, it is of crucial importance to track individual variation when conducting corpus-based studies of priming effects. In cases where the range of individual variation is very small as it was in the sample in Table 5 (10% range), accounting for individual speakers may not be necessary to get an accurate estimate of the importance of the priming effect. However, in cases where individual speakers' rates vary more (as little as the 20% range in Table 4), the effect of Previous Occurrence will be overstated since the statistical model assigns all of the probability of the sequential occurrence of the variants to the tokens and not to the speaker that uttered them. The results of the analysis of Type 3 speakers (Table 6) show that it is possible in an extreme case to obtain a significant effect for Previous Occurrence despite the fact that none of the speakers show any positive evidence for priming.[7] In these cases, individual speakers must be accounted for in some way as a moderator

variable to avoid overstating the importance of Previous Occurrence. In the case of corpora in which the individual producer of a token cannot be identified, the particular text from which the token was extracted could be used instead; however, if the corpus interface does not provide at least this level of metadata, such a corpus is not suitable for the analysis of priming effects. In terms of evaluating previous research on priming effects that has not accounted for individual speakers, it may be necessary to revisit these studies for confirmation. If through other means we know that the linguistic variable in question is not subject to wide individual variation, the results may stand as is. However, if we do not know the extent to which individual variation plays a role or if we know that the variable is subject to wide individual variation, it would be appropriate to conduct a follow-up study accounting for the individual speakers. In any case where individual speakers were not included in the model and Previous Occurrence was found to be one of the most important predictors of the variation, such a result should be treated with caution.

With respect to the reduction of Portuguese *para*, Felgueiras (1993) previously found relatively strong priming effects in her data from Rio de Janeiro (Varbrul range = 45). While the apparent priming effect in these data from Fortaleza is not as strong once individual speakers are accounted for, Previous Occurrence continues to be significant, indicating that the result obtained by Felgueiras (1993) that there is a priming effect is supported by these data. Future research should investigate the extent to which temporal distance between the prime and target moderates the priming effect. Future research, furthermore, should investigate the extent to which similarities between the prime and target enhance the priming effect.

Returning to the research questions that guided this study, the first question addressed whether the concern over individual speaker variation could be mitigated by studying priming in individual speakers separately. The results of this study suggest that, although it is important to pay attention to individual speakers' use patterns, restricting the study of priming to individual speakers is unlikely to be very fruitful because it is difficult to obtain many significant results without a large number of tokens per speaker and the small number of tokens per speaker makes it difficult to carry out multifactorial studies. The second question pertained to whether the influence of priming can be overstated in studies that fail to account for individual speaker variation. The results of this study indicate that in many circumstances the influence of priming is overstated when individual variation is not accounted for. Moreover, regarding the third question, in extreme circumstances it is possible to

obtain a significant priming effect where none should exist. Finally, with respect to the fourth question about whether this concern is mitigated when individual variation is minimal, this question can be answered in the affirmative; however, this does not absolve the researcher of the responsibility to track individual variation. Moreover, in this study, the overstatement of the priming effect only disappeared when individual variation ranged 10%, a small range that may not be reasonable to expect from very many linguistic variables.

7. Conclusion

This study aimed to determine the extent to which priming effects may be overstated when individual variation is unaccounted for using a data set of the reduction of *para* to *p(r)a* in the spoken Portuguese of Fortaleza, Brazil. The results of this study indicate that it is of crucial importance to track individual variation in corpus-based studies of priming effects and, in cases where individual variation is anything but minimal, to include individual speakers in the statistical model as a moderator variable in order to obtain an accurate estimate of the importance of priming in the study in question.

In the case of the present study, although priming is suggested to play a role in the variation surrounding the reduction of *para*, its effect is nowhere near as strong as is suggested when individual variation is not accounted for in the statistical model. Therefore, when examining past research on priming effects that have not considered individual variation, it is important to consider what else is known about the variation in question in order to ascertain whether the conclusions drawn are appropriate. This study contributes to our knowledge of best practices in the study of priming effects in corpus-based studies of language variation.

Notes

[1] Paolillo (2013) has argued that Goldvarb may be used to estimate fixed effects for individual speakers and, furthermore, that it may be more appropriate to use fixed effects to model individual variation when participants were not selected randomly.

[2] Both frequency predictors are numeric; however, there are instances where one or the other has no logical value. For example, while we can easily determine how often *vai para* 'goes to' and *é para* 'is for' occur relative to one another, it is not appropriate to quantify *para* occurring at the beginning of a phrase in the same way since it is qualitatively different from a two-word sequence. While statistical programs in the Varbrul family allow for such instances to be treated as missing data (Roy 2013),

generalized linear models (GLM) and their mixed-effects counterparts do not natively allow for such missing data.

[3] Given the fact that the individual speakers are nested into three different speech styles and the results of Gries' (2015) examination of how to treat such multi-level hierarchical relationships in corpus data, a reviewer has argued for the appropriateness of also accounting for this relationship in these data. In order to test the impact of speech style, additional multi-level generalized linear mixed-effects models were fit with both speaker and speech style varying intercepts. In all cases, speech style accounted for a fraction of the variance accounted for by speaker and in no case did the inclusion of the speech style varying intercept have any more than a superficial impact on the estimates and significance of the fixed effects, Previous Occurrence included. While there is a statistically significant difference according to the *anova()* function between the generalized linear mixed-effects model in Table 3 and its counterpart with speech style varying intercepts, it is not robust ($p = 0.02471$). In the interest of space, the models with a speech style intercept have thus been omitted, but Gries' (2015) point is well taken on the importance of accounting for all hierarchical relationships in a statistical model.

[4] This should not be interpreted as an assertion that these are the only methods to model speaker-specific effects. It is, however, necessary to limit the range of models considered.

[5] These models have been included in this paper in order to treat the topic fairly, since the issue at hand is considerably less serious for linguistic variables where individual variation is minimal. Given the very skewed distribution (see Figure 4) and the fact that partitioning the data set into smaller pieces can result in models with very few observations, the specific reduction rate ranges were selected in order to maximize the number of speakers/observations in each model, while also excluding speakers with reduction rates above 95%.

[6] A reviewer has rightly pointed out that the inclusion of speakers that use one variant categorically skews, to some extent, the effect being discussed here. A comparable model to the GLM presented in Table 3 in which Type 4 speakers have been excluded yields an estimate of 2.08660, which is somewhat lower than the model in Table 3 (2.28369), although still much higher than the models that include the speaker effect. Comparable differences in the models where the speaker effect is accounted for are, as can be expected, superficial. While under some circumstances it may be appropriate to simply exclude Type 4 speakers, if a researcher has not accounted for speaker effects in any way, there is no way to exclude Type 4 speakers. The inclusion of Type 4 speakers in these models is intended to represent the scenario where individual speakers have been completely ignored, although the mere exclusion of Type 4 speakers does not resolve the issue at hand.

[7] It is important to note that this significant effect obtained as a result of the inclusion of the one Type 3 speaker that only produced *para* sequentially. An identical GLM was fit with all Type 3 speakers except that one and in that case Previous Occurrence was not found to be significant.

Bibliographical References

Baayen, R. Harald & Milin, Petar 2010. Analyzing reaction times. *International Journal of Psychological Research* 3,2. 12-28.

Bates, Douglas; Maechler, Martin; Bolker, Ben & Walker, Steve 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67,1. 1-48.

Cameron, Richard 2004. Switch reference, verb class and priming in a variable syntax. In Beals, Katharine (ed.), *Papers from the Regional Meeting of the Chicago Linguistic Society. Parasession on variation in linguistic theory*. Chicago: Chicago Linguistic Society. 27-45.

Cedergren, Henrietta & Sankoff, David 1974. Variable rules. Performance as a reflection of competence. *Language* 50. 333-355.

Drager, Katie & Hay, Jennifer 2012. Exploiting random intercepts. Two case studies in sociophonetics. *Language Variation and Change* 24,1. 59-78.

Felgueiras, Carmen Maria 1993. *Analise da variação no uso da preposição* para. M.A. thesis. Universidade Federal do Rio de Janeiro, Rio de Janeiro.

Ferreira, Melissa Osterlund 2014. *A variação da preposição* para *na fala de Londrina pelos dados do Varsul*. Senior undergraduate thesis. Universidade Federal do Rio Grande do Sul, Porto Alegre.

File-Muriel, Richard J. 2010. Lexical frequency as a scalar variable in explaining variation. *The Canadian Journal of Linguistics / La revue canadienne de linguistique* 55,1. 1-25.

File-Muriel, Richard J.; Brown, Earl K. & Gradoville, Michael S. 2014. Methodological considerations for the study of the sibilant *s*. The role of speaker physiology. Presentation at the 12th Conference on Conceptual Structure Discourse and Language.

Flores-Ferrán, Nydia 2002. *Subject personal pronouns in Spanish narratives of Puerto Ricans in New York City. A sociolinguistic perspective*. Munich: Lincom Europa.

Gelman, Andrew & Hill, Jennifer 2007. *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.

Gradoville, Michael 2015. Social and stylistic variation in the use of phonetic variants of Fortalezense Portuguese *para*. *Sociolinguistic Studies* 9,4. 373-398.

Gradoville, Michael 2017. The cognitive representation of multi-word sequences. A usage-based approach to the reduction of Fortalezense Portuguese *para*. *Lingua* 199. 94-116.

Gradoville, Michael; Brown, Earl K. & File-Muriel, Richard 2015. The effect of varying intercepts on findings in sociophonetic data. Some observations from Caleño Spanish. *Journal of Research Design and Statistics in Linguistics and Communication Science* 2,2. 105-130.

Gries, Stefan Th. 2005. Syntactic priming. A corpus-based approach. *Journal of Psycholinguistic Research* 34,4. 365-399.

Gries, Stefan Th. 2009. *Quantitative corpus linguistics with R. A practical introduction*. London: Routledge, Taylor and Francis Group.

Gries, Stefan Th. 2015. The most under-used statistical method in corpus linguistics. Multi-level (and mixed-effects) models. *Corpora* 10,1. 95-125.

Gries, Stefan Th. & Kootstra, Gerrit Jan 2017. Structural priming within and across languages. A corpus-based perspective. *Bilingualism: Language and Cognition* 20,2. 235-250.

Huback, Ana Paula 2012. Chunking and the reduction of the preposition *para* 'to, for' in Brazilian Portuguese. *Studies in Hispanic and Lusophone Linguistics* 5,2. 277-295.

Ilari, Rodolfo; Castilho, Ataliba de; Almeida, Maria Lúcia Leitão de; Kleppa,

Lou-Ann & Basso, Renato 2008. A preposição. In Ilari, Rodolfo & Neves, Maria Helena de Moura (eds.), *Gramática do português culto falado no Brasil*. Campinas: Editora Unicamp. 623-808.

Johnson, Daniel Ezra 2008. Getting off the GoldVarb standard. Introducing Rbrul for mixed-effects variable rule analysis. *Language and Linguistics Compass* 3,1. 359-383.

Kewitz, Verena 2006. *Gramaticalização e semantização das preposições A e PARA no português brasileiro (séculos XIX e XX)*. Ph.D. dissertation. Universidade de São Paulo, São Paulo.

Lucena, Rubens Marques de 2001. *Comportamento sociolinguístico da preposição PARA na fala da Paraíba*. M.A. thesis. Universidade Federal da Paraíba, João Pessoa.

Maya, Leonardo Zechlinski 2004. *A variação da preposição* para *na fala de Porto Alegre*. M.A. thesis. Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre.

Monteiro, José Lemos 1993. O português oral culto de Fortaleza – PORCUFORT. <http://www.geocities.com/jolemos.geo/> (accessed 2006/02/27)

Nash, John C. & Varadhan, Ravi 2011. Unifying optimization algorithms to aid software system users. optimx for R. *Journal of Statistical Software* 43,9. 1-14.

Paolillo, John C. 2013. Individual effects in variation analysis. Model, software, and research design. *Language Variation and Change* 25. 89-118.

Perini, Mário 2002. *Modern Portuguese. A reference grammar*. New Haven: Yale University Press.

Poplack, Shana 1980. The notion of the plural in Puerto Rican Spanish. Competing constraints on (s) deletion. In Labov, William (ed.), *Locating language in time and space*. New York: Academic Press. 55-67.

R Core Team 2013. *R. A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. <http://www.R-project.org>

Rosemeyer, Malte 2015. How usage rescues the system. Persistence as conservation. In Adli, Aria; García García, Marco & Kaufmann, Göz (eds.), *Variation in language. System- and usage-based approaches*. Berlin: De Gruyter. 289-311.

Rosemeyer, Malte & Schwenter, Scott A. 2019. Entrenchment and persistence in language change. The Spanish past subjunctive. *Corpus Linguistics and Linguistic Theory*. 15,1. 167-204.

Roy, Joseph 2013. Sociolinguistic statistics. The intersection between statistical models, empirical data and sociolinguistic theory. In Barysevich, Alena; D'Arcy, Alexandra & Heap, David (eds.), *Proceedings of Methods XIV. Papers from the Fourteenth International Conference on Methods in Dialectology, 2011*. Bern: Peter Lang. 261-275.

Sankoff, David & Laberge, Suzanne 1978. Statistical dependence among successive occurrences of a variable in discourse. In Sankoff, David (ed.), *Linguistic variation. Models and Methods*. New York: Academic Press. 119-126.

Sankoff, David; Tagliamonte, Sali A. & Smith, Eric 2005. *Goldvarb X. A variable rule application for Macintosh and Windows*. Toronto: University of Toronto Department of Linguistics.

Sankoff, David; Tagliamonte, Sali A. & Smith, Eric 2015. *Goldvarb Yosemite. A variable rule application for Macintosh*. Toronto: University of Toronto Department of Linguistics.

Scherre, Maria Marta Pereira & Naro, Anthony J. 1991. Marking in discourse. 'Birds of a feather.' *Language Variation and Change* 3. 23-32.

Silva, Nahete de Alcântara 2010. *A preposição para e suas variantes no falar araguatinense*. M.A. thesis. Universidade Federal da Paraíba, João Pessoa.

Snijders, Tom & Bosker, Roel 1999. *Multilevel analysis. An introduction to basic and advanced multilevel modeling*. London: Sage Publications.

Szmrecsanyi, Benedikt 2006. *Morphosyntactic persistence in spoken English. A corpus study at the intersection of variationist sociolinguistics, psycholinguistics, and discourse analysis*. Berlin: De Gruyter.

Tagliamonte, Sali A. 2012. *Variationist sociolinguistics. Change, observation, interpretation*. Chichester, West Sussux: Wiley-Blackwell.

Tamminga, Meredith 2016. Persistence in phonological and morphological variation. *Language Variation and Change* 28. 335-356.

Thomas, Earl W. 1969. *The syntax of spoken Portuguese*. Nashville: Vanderbilt University Press.

Travis, Catherine E. 2007. Genre effects on subject expression in Spanish. Priming in narrative and conversation. *Language Variation and Change* 19. 101-135.

Travis, Catherine E.; Torres Cacoullos, Rena & Kidd, Evan 2017. Cross-language priming. A view from bilingual speech. *Bilingualism: Language and Cognition* 20,2. 283-298.

Vellasco, Ana Maria de Moraes Sarmento 1998. Um estudo da variação da preposição *para* no português do Brasil. In Grosse, Sybille & Zimmermann, Klaus (eds.), *'Substandard' e mudança no português brasileiro*. Frankfurt am Main: Teo Ferrer de Mesquita. 291-314.

Weiner, E. Judith & Labov, William 1983. Constraints on the agentless passive. *Journal of Linguistics* 19. 29-58.