languages

MDPI

Article

Reanalyzing Variable Agreement with *tu* Using an Online Megacorpus of Brazilian Portuguese

Scott A. Schwenter *, Lauren Miranda, Ileana Pérez and Victoria Cataloni

Department of Spanish and Portuguese, The Ohio State University, Columbus, OH 43210, USA; miranda.137@osu.edu (L.M.); perez.641@osu.edu (I.P.); cataloni.1@osu.edu (V.C.)
* Correspondence: schwenter.1@osu.edu

Abstract: We reanalyze the phenomenon of verbal (non)agreement with the 2SG *tu* in a megacorpus of Brazilian Portuguese compiled from the web. Unlike previous research, which has analyzed sociolinguistic interview data and regional differences, we examine these data with a focus on the internal linguistic factors that constrain the variability. Our analysis of 4860 tokens of *tu* + verb reveals that non-agreement with the 3SG verb form is by far the most common pattern, 2SG agreement being relatively infrequent. Individual verb lexemes show highly distinct rates of (non)agreement. In addition, the specific tense/aspect/mood forms and main/auxiliary status are likewise significant factors affecting the variation. We conclude that future studies of this phenomenon should not ignore these internal linguistic factors. We situate our study within a group of other recent studies in Romance linguistics, which have found that individual verbal and constructional patterns can have diverse effects on morphosyntactic variation.

Keywords: Brazilian Portuguese; pronouns; variation; online corpus data; quantitative analysis

check for updates

Citation: Schwenter, Scott A., Lauren Miranda, Ileana Pérez, and Victoria Cataloni. 2024. Reanalyzing Variable Agreement with *tu* Using an Online Megacorpus of Brazilian Portuguese. *Languages* 9: 197. https://doi.org/10.3390/languages9060197

Academic Editor: Patrícia Amaral

Received: 4 November 2023
Revised: 15 March 2024
Accepted: 19 May 2024
Published: 28 May 2024

1. Introduction: The Phenomenon

While Brazilian Portuguese (BP) is widely known for its widespread use of *você* to denote second-person singular (2SG) reference (Faraco 1996), there are regions of the country that still make use of the pronoun *tu*, either exclusively or in conjunction with *você*, in order to effect 2SG reference. Probably most famous is the use of *tu* in the South of Brazil, especially in the state of Rio Grande do Sul, where *você* is rarely used, neither as a subject nor a direct object (cf. Schwenter et al. 2018). In other regions, for example, in the North (e.g., Pará) and some of the Northeast states (e.g., Maranhão, Ceará) of Brazil, the two pronouns are used to differing extents. The variation is even found in the Southeast dialect of Rio de Janeiro, which Paredes Silva (2003) has characterized as a "return" of *tu* to the Carioca dialect.

Beyond the presence or absence of the pronoun *tu* in Brazil, there is also variability in BP regarding verbal agreement with *tu* between the normative and historical 2SG paradigm found in European Portuguese (where *tu* occurs invariably with 2SG agreement) and the "mixed" system that pairs *tu* with 3SG agreement. The latter is the pattern found in BP with the subject *você*, which derives from an originally third-person possessive NP used for 2SG reference (*vossa mercê* 'your mercy'), which evolved to be the most widespread 2SG pronoun in BP. Indeed, most grammatical descriptions of BP (e.g., Perini 2002; Kato et al. 2022) overlook this variation completely and simply state that 2SG *tu* is found invariably with 3SG agreement in Brazil. A recent overview article on variable agreement in BP, both nominal and verbal (Mendes and Oushiro 2015), likewise makes no mention of the variation in agreement found with *tu*. More specialized studies (e.g., Scherre et al. 2015) have actually revealed considerable nuance in the situation, however, and identified regional variation in the patterns of (non)agreement. This variable phenomenon is evidenced in the following examples of present indicative (1), simple past (2), imperfect (3), and future subjunctive (4). The (a) versions in each case illustrate the normative 2SG conjugation (as found in European

Portuguese), while the (b) versions show the common BP pattern of non-agreement between *tu* and the 3SG verb conjugation.

(1a) Tu és (2SG)
You be.PRES:2SG
'You are'

(1b) Tu é (3SG)
you be.PRES:3SG
'You are'

(2a) Tu comeste (2SG)
you eat.PRET:2SG
'You ate'

(2b) Tu comeu
you eat.PRET:3SG
'You ate'

(3a) Tu escrevias (2SG)
you write.IMPF:2SG
'You wrote/were writing'

(3b) Tu escrevia (3SG)
you write.IMPF:3SG
'You wrote/were writing'

(4a) (Se) tu achares (2SG)
if you believe.FUTSUBJ.2SG
'(If) you believe'

(4b) (Se) tu achar (3SG)
if you believe.FUTSUBJ.3SG
'(If) you believe'

This variation between the 2SG and 3SG verb forms has traditionally been treated as a dialectal phenomenon, though all regions where *tu* is found seem to show at least a minimal amount of 2SG agreement. One issue with this research is that the data sets that have been analyzed in the many regions where *tu* is found are not uniform in size or in the methods of collection used, making it difficult to draw valid comparisons. One goal of this paper is to examine a much larger data set that allows for greater lexical and constructional variety, and also for the random sampling of the data. We hope that this will provide at least a partial model for future studies of the phenomenon, no matter what the source of the data or its manner of collection.

As stated above, previous studies that have investigated (non)agreement with *tu* have mainly concentrated on the regional distribution and frequencies of the variation between 2SG and 3SG agreement patterns. Rather large discrepancies in (non)agreement rates have been found between these studies, even in studies of the same region or city. Thus, to take one example from a city often cited as having a high rate of 2SG agreement, Loregian (1996) found 39% agreement in sociolinguistic interview data from Florianópolis and a similar rate (43%) in her own follow-up study (Loregian-Penkal 2004). However, a more recent study by Davet and Campos-Antoniassi (2014), using a different corpus of interviews, found only 14% 2SG agreement in Florianópolis. While the latter authors consider the possibility of change in progress (which we deem unlikely in corpora separated by only 20 years), another possibility could be that the varying results are due to interspeaker differences in the populations surveyed or are due to internal linguistic factors, such as the particular verbs that occur in the corpus, the tense/aspect/mood forms in which they appear, their functions as main verbs or auxiliaries, etc.

Beyond the geographical variability of (non)agreement, some have considered education levels (*nível de escolarização*) as an independent variable, with some researchers finding increasing amounts of 2SG agreement as education increases (e.g., Davet and Campos-Antoniassi 2014 in Florianópolis) but others finding higher levels of this agreement among speakers with lower education levels in other localities (e.g., Loregian-Penkal 2004 in Porto Alegre). The spoken versus written mode has also been analyzed in several studies, with written language showing more 2SG agreement overall than spoken, but once again with widely varying rates depending on the study (Guimarães 1979; Loregian-Penkal 2004). However, in none of these prior studies has the effect of individual speaker variation been taken into account, and it is well known that individual speech patterns can heavily skew data sets. Indeed, when individual speakers are identifiable in the data, they should be treated as random effects in any variationist study (Baayen 2008; Johnson 2009); this has not been the case in prior studies of 2SG agreement in BP.

In their summary of the prior research on 2SG pronoun variation and (non)agreement in BP, Scherre et al. (2015; also Scherre and Duarte 2016) argue for a classification of six different pronominal systems for 2SG reference. They base the classification of these six systems on the following four factors, which can have positive or negative values. First, there is the possible presence of *tu* in the system, since there are major dialects or dialect regions (e.g., São Paulo, Belo Horizonte) in which *tu* does not play a role. Second, there is the frequency of *tu* in the system, which can range from rather sporadic in the dialects of Rio de Janeiro to the nearly exclusive usage of *tu* in the dialect of Porto Alegre. The third factor is the presence or absence of the canonical verb inflection *-s*, and the fourth factor is the average rate of overt 2SG agreement. The third factor is very limited in its application, since not all of the verb conjugations with *tu* are marked by *-s*. For instance, the simple past form is marked by *-ste*, as in *falaste* or *comeste*.

Scherre et al. (2015) go on to state that a "remarkable feature of the system" is non-overt agreement: "the lack of the overt 2nd person singular agreement mark *-s*", which is not associated with any particular social stigma, a position corroborated by Souza and Chaves' (2015) study of speaker evaluations of (non)agreement in Florianópolis. However, two of the factors that Scherre et al. (2015) use to distinguish 2SG pronominal systems are precisely related to 2SG verbal inflection, i.e., the presence or absence of 2SG agreement and the average rate of overt agreement with the pronoun *tu*. Thus, it would be more accurate, in our view, to simply state that 3SG agreement is always the most frequent variant, no matter what region is analyzed, and the rates of 2SG agreement with *tu* can vary by region.

The updated survey of 2SG pronoun research presented in Scherre et al. (2020) reduces the variation in (non)agreement considerably to the North region of Brazil (Pará, Maranhão), with low rates of agreement (<25%) in the region. The South region, specifically in Santa Catarina, shows the highest rates of agreement (just over 25%), while agreement in Rio Grande do Sul is sporadic, and lower than 5%.[1]

Scherre et al. (2020, p. 274) conclude that more work is needed on the geographic spread of *tu* and of the patterns of (non)agreement in Brazil: "[P]ara que tenhamos um mapa ainda mais próximo da realidade, são necessárias e urgentes mais pesquisas no vasto território brasileiro, com o controle de, pelo menos, cinco possibilidades disponíveis no português brasileiro: *você*, *ocê*, *cê*, *tu* com concordância e *tu* sem concordância, com o controle rigoroso dos contextos sintácticos e das nuances interacionais" ['For us to have a map that is closer to reality, more research is necessary and urgent in the vast Brazilian territory, with control of, at least, five possibilities available in Brazilian Portuguese: *você*, *ocê*, *cê*, *tu* with agreement and *tu* without agreement, with rigorous control of syntactic contexts and interactional nuances'—our translation]. While we of course recognize that more work along these lines is always useful to clarify the empirical reality of Brazilian Portuguese, it must be pointed out that the map created by Scherre et al. (2020) is the fruit of nearly 60 studies (Scherre et al. 2020, p. 270). In this paper, we present a distinct approach with a new data source that may help further clarify this variation, and especially the variation in verbal (non)agreement with the subject *tu*. We are pessimistic about the possibility of interview methods allowing for the "rigorous control of syntactic contexts", as in the quote above (Scherre et al. 2020, p. 274). It is for this reason that we decided to shift the empirical focus in this study away from sociolinguistic interviews to a megacorpus of Brazilian Portuguese web data, which will permit us to examine the variation in question with greater control of several internal linguistic factors, which we enumerate below.

In the remainder of this paper, we present the results of the corpus analysis of verbal (non)agreement with *tu* in a random sample taken from the web-based megacorpus used for this study. As alluded to above, we are interested in whether there are differing patterns according to the verb lexeme, verb form (tense/aspect/mood), lexical vs. auxiliary verb, phonic salience (determined by the morphophonological differences between verb forms), frequency, etc. The next section presents our methods in more detail, including the independent variables that we coded for. Section 3 presents both the descriptive statistics of our data (Section 3.1) and the inferential statistical analysis (Section 3.2). We discuss the broader contributions of our study in Section 4.

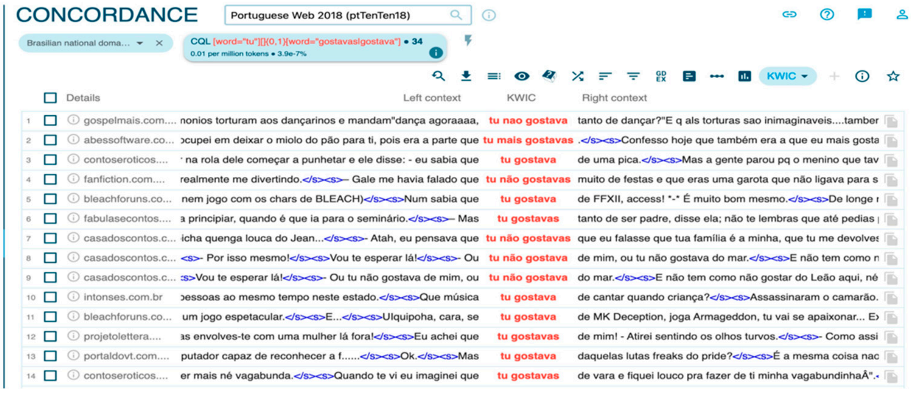
2. Methods

To collect data for the study, we utilized the Brazilian National Domain (.br) of the Portuguese Web 2018 (ptTenTen18) corpus from the Sketch Engine family of corpora (http://www.sketchengine.eu; accessed on 10 April 2023; see Kilgarriff et al. 2014). Sketch Engine was developed by Lexical Computing CZ and functions as a corpus manager, where users can analyze authentic texts of billions of words, known as their text corpora. The Brazilian National Domain corpus alone contains 4.7 billion words from diverse online sources. Due to that diversity, we do not claim that the data analyzed are necessarily representative of either spoken or written BP. However, the data **are** representative of Brazilian Portuguese as it is found on the web, where both spoken and written texts are available. Moreover, we believe that the constraints we uncover in the analysis to come likely resemble those found in BP more generally, and these constraints can be uncovered more easily by using a diverse online megacorpus. For this study, we opted to use Sketch Engine because it allowed us to look at a broader variety of data from distinct registers, topics, and styles, and also examine the lexical and constructional differences between verbs and different verb forms, facets which previous studies using sociolinguistic interviews have not included in their analysis.

Using the data from the ptTenTen18 corpus, we aim to determine what linguistic factors license 2SG verbal agreement or non-agreement with the pronoun *tu* in these corpus data. Following research on the effect of verbal lexemes and their frequency on variable phenomena in other Romance varieties (e.g., Poplack et al. 2018), as well as the intuitions of several native BP speakers who speak a *tu*-dominant dialect, we hypothesize that individual verbs will show distinct rates of second singular (non)agreement. We also hypothesize, following Naro (1981) among others, that phonic salience will play a critical role in the rates of 2SG agreement and that there will be more agreement when there is higher phonic salience between the 2SG and 3SG variants, i.e., when there are greater morphophonological differences between the two forms.

To select the verbs for our analysis, we used a random number generator to select 10 of the 20 most frequent verbs in the corpus. These verbs were *ser*, 'to be', *estar*, 'be', *ir*, 'go', *ter*, 'have', *ficar* 'stay', *poder* 'to be able to', *achar*, 'think', *falar*, 'speak', *gostar* 'like', and *escrever*, 'write.' The other 10 verbs that we included in our analysis were randomly selected from the top 1000 most frequent verbs, excluding the top 20 verbs since these had already been included in the first round of verb selection. The less-frequent verbs included *abrir*, 'open', *aparecer*, 'appear', *chamar*, 'call', *crer*, 'believe', *escolher*, 'choose', *lembrar*, 'remember', *mudar*, 'change', *preferir*, 'prefer', and *sair*, 'to go out'. The reasoning behind this method was to attempt to include verbs of varying frequency among those used in the analysis, but at the same time knowing that less frequent verbs were likely to have lower numbers of usable tokens. As we detail below in Section 3.1, we ended up separating two distinct forms of one of these verbs (*estar*) in our analysis due to broadly diverging patterns of (non)agreement between these forms.

For each of the 20 verbs, we used Sketch Engine's built-in random sample generator to select 100 tokens of the verbs for four different finite forms (see below for explanation). We therefore attempted to collect at least 400 tokens (100 per finite form) for each of the 20 verbs. Verbs that were less frequent in the corpus did not always meet the 100-token goal; therefore, for those verbs, we included all the tokens available in the corpus for each verb tense. In Sketch Engine's search interface, we used the corpus query language (CQL) to search for examples with the explicit co-occurrence of the subject pronoun *tu*; see the sample search results in Figure 1. Using the CQL, we controlled for the (non)agreement of the pronoun but not necessarily the use of the variant forms in general. This was due to the fact that the corpus does not always provide sufficient accessible context to determine the reference of third-person singular forms that, without a co-occurring subject pronoun, are ambiguous between 2SG (*você/tu/o senhor/a senhora*) and 3SG (*ele/ela*) reference. Note also that we found no cases of the inverse of what we were investigating, i.e., there are no examples in the corpus of a 2SG verb form co-occurring with a subject pronoun that normatively takes 3SG agreement. Thus, while both *tu falas* 'you speak' and *tu fala* occur, as well as *você fala*, the non-agreeing form **você falas* does not occur in the data (nor are we aware of it being used in BP).



Portuguese Web 2018 (ptTenTen18)

Brasilian national doma...

CQL [word="tu"][]{0,1}[word="gostavas|gostava"] • 34
0.01 per million tokens • 3.9e-7%

	Details	Left context	KWIC	Right context
1	gospelmais.com....	nonios torturam aos dançarinos e mandam"dança agoraaaa,	tu nao gostava	tanto de dançar?"E q als torturas sao inimaginaveis....tamber
2	abessoftware.co...	ɔcupei em deixar o miolo do pão para ti, pois era a parte que	tu mais gostavas	.</s><s>Confesso hoje que também era a que eu mais gosta
3	contoseroticos....	' na rola dele começar a punhetar e ele disse: - eu sabia que	tu gostava	de uma pica.</s><s>Mas a gente parou pq o menino que tav
4	fanfiction.com....	realmente me divertindo.</s><s>— Gale me havia falado que	tu não gostavas	muito de festas e que eras uma garota que não ligava para s
5	bleachforuns.co...	nem jogo com os chars de BLEACH)</s><s>Num sabia que	tu gostava	de FFXII, access! *-* É muito bom mesmo.</s><s>De longe r
6	fabulasecontos....	a principiar, quando é que ia para o seminário.</s><s>— Mas	tu gostavas	tanto de ser padre, disse ela; não te lembras que até pedias
7	casadoscontos.c...	icha quenga louca do Jean...</s><s>- Atah, eu pensava que	tu não gostavas	que eu falasse que tua família é a minha, que tu me devolves
8	casadoscontos.c...	<s>- Por isso mesmo!</s><s>Vou te esperar lá!</s><s>- Ou	tu não gostava	de mim, ou tu não gostava do mar.</s><s>E não tem como n
9	casadoscontos.c...	:s>Vou te esperar lá!</s><s>- Ou tu não gostava de mim, ou	tu não gostava	do mar.</s><s>E não tem como não gostar do Leão aqui, né
10	intonses.com.br	ɔessoas ao mesmo tempo neste estado.</s><s>Que música	tu gostava	de cantar quando criança?</s><s>Assassinaram o camarão.
11	bleachforuns.co...	um jogo espetacular.</s><s>E...</s><s>Ulquipoha, cara, se	tu gostava	de MK Deception, joga Armageddon, tu vai se apaixonar... E›
12	projetolettera....	as envolves-te com uma mulher lá fora!</s><s>Eu achei que	tu gostavas	de mim! - Atirei sentindo os olhos turvos.</s><s>- Como assi
13	portaldovt.com....	putador capaz de reconhecer a f......</s><s>Ok.</s><s>Mas	tu gostava	daquelas lutas freaks do pride?</s><s>É a mesma coisa nac
14	contoseroticos....	er mais né vagabunda.</s><s>Quando te vi eu imaginei que	tu gostavas	de vara e fiquei louco pra fazer de ti minha vagabundinhaÂ".‹

Figure 1. Example search result from Sketch Engine.

For each of the tokens, we coded our dependent variable (agreement or non-agreement with the overt second-person singular pronoun *tu*) and seven independent variables, chosen on the basis of findings from prior studies on agreement in BP (e.g., Scherre et al. 2007) and also based on our hypotheses about what factors might constrain the patterns of (non)agreement. The dependent variable contrasted the normative form, which has agreement corresponding to the 2SG *tu*, with the non-agreeing 3SG verb form that Brazilian speakers often use instead of the normative 2SG variant. For our independent variables (see Table 1), we coded the verb lexeme in the form of the infinitive for each of the 20 verbs included in the study. We also included low, mid, and high phonic salience, which is the degree of phonological difference between the two possible forms (to be elucidated further in Section 3), an important variable that has been used in other studies on (non)agreement in BP since the seminal work of Naro (1981). In addition, we coded for frequency, specifically the frequency per million words of the verb lexeme in the corpus, hypothesizing that frequency could affect rates of (non)agreement. In the coding scheme, we also distinguished between main and auxiliary verb function, which we hypothesized could show differing rates of agreement—specifically, lower rates for auxiliaries and higher rates for main verbs, since the former tend to be more phonologically reduced and semantically bleached than the latter. We coded as well for the particular finite form of the verb, specifically the present, preterit, imperfect and future subjunctive forms for each verb, since these allowed us to determine different degrees of phonic salience. Finally, we included coding for the polarity of the sentence, i.e., affirmative or negative, as well as intervening words, that is, whether and how many words occurred between the pronoun *tu* and the finite verb in each token, which is another factor often included in studies of (non)agreement. Given the paucity of tokens with more than one intervening element between the subject pronoun and verb, in the end we only compared the binary contrast between zero and one intervening word. Collinearity between polarity and intervening words can arise if a negative word occurring in the preverbal position between the subject pronoun *tu* and the verb is also counted as an intervening word. As a result, we chose not to include such cases of negation as an intervening word, thus guaranteeing the independence of the two categories. Intervening words in the data were therefore mainly limited to adverbs such as *já* 'already', *sempre* 'always', or clitic pronouns such as the (direct or indirect) object *me* 'you' or reflexive *te* 'you' (as in *tu te chama[s]* 'you call yourself').

Table 1. Independent variables and their values.

Independent Variables	Values
Verb Lexeme	verb (labeled by infinitive)
Phonic Salience	low, mid, high
Verb Lemma Frequency	tokens per million
Verb Type	main or auxiliary
Intervening Words	0 or 1
Polarity	affirmative, negative
Verb Form	present, preterit, imperfect, future subjunctive

3. Results

3.1. Descriptive Statistics

This section presents a description, using rates of (non)agreement, of the independent variables found to be significant in our inferential statistical analysis (see Section 3.2). In total, we extracted and coded 4860 tokens of *tu* with a conjugated verb from the Portuguese Web 2018 corpus. We performed a descriptive and inferential statistical analysis of these tokens in R (R Core Team 2023). The overall results of (non)agreement in this data set were 1514 cases of 2SG agreement (31.2%) versus 3346 cases (68.8%) of 3SG agreement (i.e., non-agreement with 2SG *tu*). This distribution is shown graphically in Figure 2.

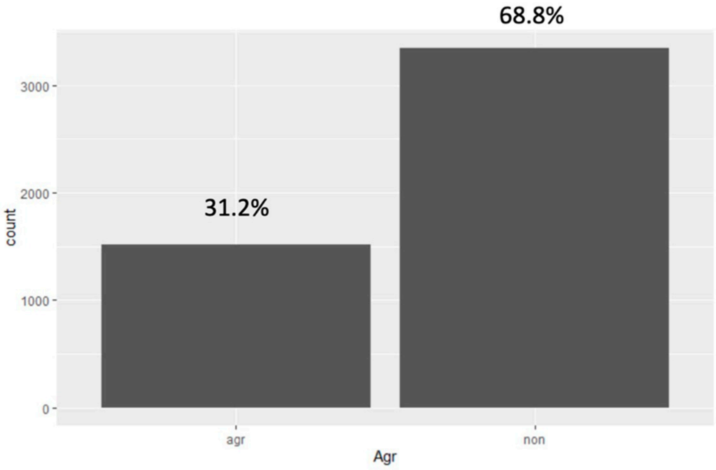


Figure 2. Overall distribution of (non)agreement in data set.

There were notable differences between the verbs under consideration. These are seen in Table 2, where each verb lexeme is presented, ordered by its frequency per million in the corpus, followed by its rates of agreement and non-agreement with the subject *tu* in our data. The relative frequency here stands for the rate per million words of occurrences of the verbal lexeme in the .br domain of the ptTenTen18 corpus. As can be seen, there is a wide range of variability by verb: the reduced form *tar* (<*estar*; see below for more discussion of this form) is the verb that shows the lowest rate of agreement at only 1.9%, in stark contrast to its unreduced counterpart *estar*, whose agreement rate of 63.2% is surpassed only by the verb *crer*, a lexeme which, in our data, is heavily restricted to religious contexts (BP uses the verbs *acreditar* and *achar* more commonly for 'to believe'), at 65.3%.

Table 2. Overall rates of (non)agreement, by verb, ordered by frequency per million.

	ser	*estar*	*tar*	*poder*	*ter*	*fazer*	*ir*	*ficar*	*falar*	*achar*	*chamar*
Freq/million	17,426.63	4007.25	N/A	3217.78	2503.25	2258.97	1720.42	1212.86	606.29	412.89	377.94
Agr	150 (44.0%)	199 (63.2%)	4 (1.9%)	95 (36.3%)	109 (29.9%)	103 (28.7%)	95 (23.1%)	48 (19.1%)	64 (24.7%)	88 (18.6%)	88 (55.3%)
Non-Agr	191 (56.0%)	116 (36.8%)	204 (98.1%)	167 (63.7%)	255 (70.1%)	256 (71.3%)	317 (76.9%)	203 (80.9%)	195 (75.3%)	385 (81.4%)	71 (44.7%)
	gostar	*sair*	*lembrar*	*escrever*	*abrir*	*mudar*	*escolher*	*aparecer*	*preferir*	*crer*	
Freq/million	330.87	323.6	265.65	236.05	220.9	205.93	198.46	189.27	93.25	92.05	
Agr	51 (18.8%)	48 (19.1%)	51 (34%)	108 (48.9%)	15 (19.7%)	69 (39.7%)	49 (40.5%)	7 (10.3%)	19 (26.8%)	64 (65.3%)	
Non-Agr	221 (81.2%)	203 (80.9%)	99 (66%)	113 (51.1%)	61 (80.3%)	105 (60.3%)	72 (50.5%)	61 (89.7%)	52 (73.2%)	34 (34.7%)	

A visualization of these results is given in Figure 3, ordered from the greatest rates of 2SG non-agreement on the left to the greatest rates of 2SG agreement on the right. We see that only three verbs (*chamar*, *crer*, *estar* [excluding *tar*]) have rates of 2SG agreement over 50%. Most of the other verbs are well under that percentage and nearly half of them are at or under a rate of 25% agreement. Again, however, there is considerable variability among the verbs analyzed, and their behavior is far from uniform with respect to (non)agreement. This result suggests, therefore, that sociolinguistic studies on 2SG agreement that utilize interview data to determine rates of (non)agreement with *tu* could be heavily skewed by individual verb frequency, as well as the potential (lack of) diversity of verbs in the data analyzed. To take an obvious hypothetical example, a data set of informal BP conversation that includes copious amounts of a highly frequent verb lexeme in the 2SG like *ser*, but few others, would most likely have a very low overall rate of 2SG agreement, but this overall rate for the region or dialect in question could be artificially suppressed precisely due to the high frequency of *ser* in the data.

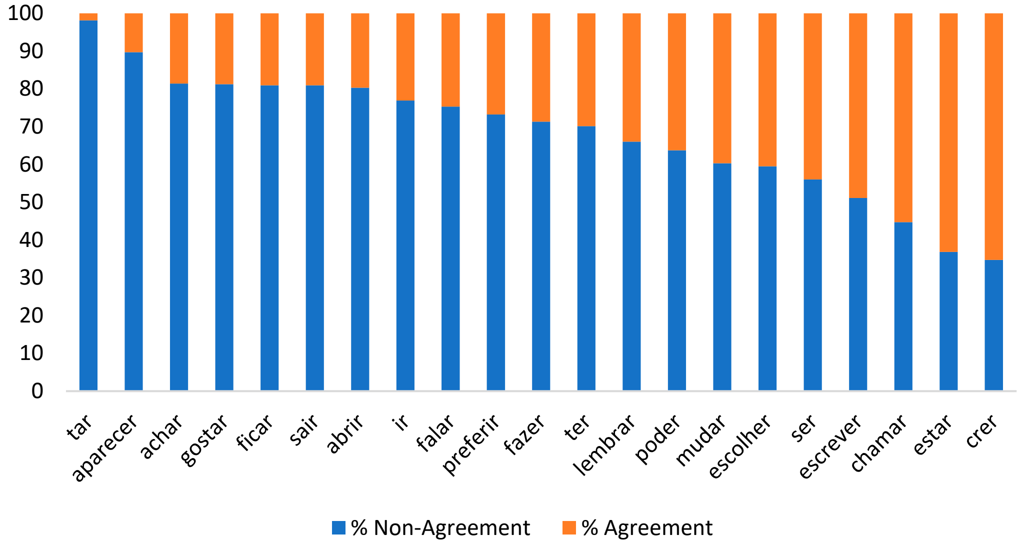


Figure 3. Rates of (non)agreement by verb.

Indeed, one of the most curious cases in our data comes from the stark contrast between the copular verb *estar* ‘to be’, originally included in the random sample of 20 verbs to be analyzed from the corpus, and its reduced variant *tar*, which turned out to be extremely frequent in our searches for the distinct forms of *estar*. In BP, *estar* is often reduced in speech and informal writing (e.g., on social media or informal chats between family or friends such as on WhatsApp) via the deletion of the initial syllable, as shown in the examples that follow.

(5) Eu estou > Eu tô
'I am'
(6) Tu estava(s) > Tu tava(s)
'You were'
(7) Ela esteve > Ela teve
'She was'

When we analyzed these two verbs separately with respect to (non)agreement, we encountered a large discrepancy: 2SG agreement was the majority variant in the case of unreduced *estar* (63.2%) but there was virtually no such agreement (4/208 tokens or 1.9%) for reduced *tar*. See the full results in Table 3.

Table 3. Rates of (non)agreement by *estar* vs. *tar*.

	Estar	***Tar***
Agreement	199 (63.2%)	4 (1.9%)
Non-agreement	116 (36.8%)	204 (98.1%)

This contrast is, we believe, reflective of the normative view of agreement, which is mirrored by the use of the full, normative, bisyllabic (or more in some tense/aspect/mood conjugations) verb form. We should note, however, that the reduced form created several issues for our corpus searches, since some of the reduced forms deriving from *tar* become homophonous with (and are also homographs of) forms of completely different verbs. Thus, for example, the preterit *tu teve* 'you were' with a nonagreeing 3SG form (instead of 2SG *tu tiveste*) could have been a reduced form of *estar* (*teve* < *esteve*), or the 3SG preterit form of the verb *ter* 'to have'. A preliminary search of these preterit forms revealed that their tagging in the corpus was inconsistent and therefore we chose to exclude them from the analysis. For this reason, the number of tokens for *estar* (*n* = 315) and those for *tar* (*n* = 208) are not equivalent, and not directly comparable for each of their tense/aspect forms.

The results for the four different tense/aspect forms that were extracted from the corpus and analyzed for each of the 21 verbs are provided in Table 4. The rates of agreement follow the hierarchy Imperfect > Present > Past > Future Subjunctive. As we will discuss immediately below, this particular independent variable is highly correlated with that of phonic salience. In our inferential analysis (see Section 3.2 below), it was not advisable to include both of these factors due to their collinearity. Once again, the discrepancy in the totals for the different forms is due to the paucity of some forms (see e.g., future subjunctive) in collocation with *tu* in the corpus, a problem that does not arise for the more abundant present tense forms.

Table 4. Rates of (non)agreement by verb form.

	Present	**Past**	**Imperfect**	**Future Subjunctive**
Agreement	644 (32.2%)	384 (28.4%)	303 (37.7%)	183 (26.2%)
Non-agreement	1359 (67.8%)	970 (71.6%)	501 (62.3%)	516 (73.8%)

The results for the factor phonic salience (PS) are presented in Table 5. As mentioned in Section 2, the different levels of PS corresponded to different verb forms, here exemplified with the competing forms of the verb *falar* 'to talk, say'.

Table 5. Rates of (non)agreement by phonic salience.

	Low	**Mid**	**High**
Agreement	928 (34.3%)	202 (25.2%)	384 (28.4%)
Non-agreement	1776 (65.7%)	600 (74.8%)	970 (71.6%)

- Low Phonic Salience: Present (*tu fala/falas*) and Imperfect (*tu falava/falavas*)

- Mid Phonic Salience: Future Subjunctive (*tu falar/falares*)
- High Phonic Salience: Preterit (*tu falou/falaste*)

These distinctions were made based on the degree of difference between the 2SG and 3SG forms for each verb form. In the case of low PS, the only difference between the two forms is the addition of the morpheme *-s* in the 2SG form, while the 3SG form lacks that morpheme. For mid PS, the difference is found not only in the addition of the morpheme *-es* to make the 2SG form, but also in the additional syllable created by that morpheme when compared to the 3SG form. Finally, for high PS, there is a much greater change when comparing the past tense morphemes, *-ou* (or *-eu/-iu*) for 3SG versus *-aste* (or *-este*/*-iste)* for 2SG, in addition to the extra syllable that the 2SG morpheme adds to the word.

The overall rates for (non)agreement depending on Phonic Salience show that low PS forms show higher rates of 2SG agreement than mid or high PS, a result that is contrary to prior results in the literature for 3PL verb forms (Naro 1981, among many others), where high PS forms showed higher rates of agreement than lower PS forms. We interpret this result as owing to the relative ease of converting the low PS form from 3SG to 2SG, and speakers' recognition of *-s* as the normative marker of 2SG agreement for *tu*. While it may be true that Brazilians learn (but possibly forget) the full paradigm of 2SG in the formal educational system, the overall frequency of the low PS forms most likely gives speakers more familiarity with them than the preterit or future subjunctive forms (which, in the case of the latter, are considered to be more pedantic, according to several native speakers of BP *tu* dialects that we asked).

The next set of descriptive results we present here concerns the distinction between the main and auxiliary uses of verbs. As noted in grammaticalization studies by researchers such as Bybee et al. (1994) and Heine (1993), the reduction in semantic content in the diachronic shift from main (lexical) to auxiliary verb is paralleled by a reduction in the formal properties and other features, such as agreement, that are characteristic of lexical verbs. In the case of 2SG agreement, we hypothesized that auxiliary verbs would show lower rates of agreement than their main verb counterparts, since the principal lexical content is not conveyed by the auxiliary but by the main verb. As Tables 6 and 7 show, this hypothesis is true not only for the full data set (Table 6), but is also more clearly true for those verbs that have both main and auxiliary verb uses (Table 7), such as *estar*, *ter*, and *ir*.

Table 6. Rates of (non)agreement for main vs. auxiliary verb uses, full data set.

	Main	**Auxiliary**
Agreement	1271 (32.0%)	243 (27.3%)
Non-agreement	2699 (68.0%)	647 (72.7%)

Table 7. Rates of (non)agreement for main vs. auxiliary verb uses, only verbs with both uses.

	Main	**Auxiliary**
Agreement	390 (31.5%)	152 (23.8%)
Non-agreement	848 (68.5%)	487 (76.2%)

The factor Intervention distinguished between tokens where no words intervened between the subject pronoun *tu* and the following verb, and tokens where one word intervened between the subject and verb. As noted above, there were extremely few tokens containing more than one element intervening between the subject and the verb, and we made the resulting decision to limit this factor to a binary comparison. As can be seen in Table 8, there was considerably more agreement in the case of one intervening element than in the case of no intervening elements, which leads us to believe that there may be a constructional effect of subject pronoun *tu* + 3SG verb collocations (e.g., *tu come* 'you eat'), which is not as robust when intervening elements such as temporal adverbs or clitic pronouns intervene between the subject and the verb (e.g., *tu sempre comes* 'you always eat'). Due to space limitations, we do not discuss this possibility further here, but it is undoubtedly an important issue for future research.

Table 8. Rates of (non)agreement by intervening elements (0 vs. 1).

	0 Intervening Elements	**1 Intervening Element**
Agreement	1299 (29.7%)	215 (45.5%)
Non-agreement	3088 (70.3%)	258 (54.5%)

Regarding the factor Polarity, there was slightly more non-agreement found in negative sentences (71.2%) than in affirmatives (68%), as seen in Table 9. The differences in the (non)agreement rates for Intervention in Table 8 versus the rates for Polarity in Table 9 provide further justification for our decision to separate these two factors in the analysis.

Table 9. Rates of (non)agreement by Polarity.

	Affirmative	**Negative**
Agreement	1399 (32.0%)	115 (28.8%)
Non-agreement	3062 (68.0%)	284 (71.2%)

The last set of descriptive statistics that we present in this section relates to the corpus frequency of the verbal lexemes included in our study in combination with the 2SG pronoun *tu*. Again, there were vast discrepancies in the frequencies per million words of the different verb forms analyzed, as seen above in Table 2, ranging from 17,426.63/million in the case of *ser* to only 92.05/million for *crer*, with the median frequency being 330.87/million, indicating the considerable rightward skew of verb frequency. Therefore, we opted to normalize the verb frequencies by log-transforming them in order to meet the necessary assumptions needed for valid statistical analysis. After the log transformation, a conditional inference tree was used to examine the effects of frequency on the data set overall, and it was determined that the verbs could be split into two groups, namely high and low frequency, where the high group included essentially those verbs at or above the median frequency and the low group those below. The overall rates of (non)agreement by high/low log frequency are given in Table 10; as can be seen, the rates were nevertheless very similar in both the high and low-frequency groups.

Table 10. Rates of (non)agreement by binary log frequency.

	High	**Low**
Agreement	803 (32.0%)	711 (30.3%)
Non-agreement	1709 (68.0%)	1637 (69.7%)

Summarizing this section, the descriptive results we have presented show the effects of different linguistic constraints on (non)agreement with *tu* in our data. There are clear patterns to this variation, but perhaps most importantly, we have corroborated our principal hypothesis, which is that individual verbs would display distinct rates of (non)agreement with *tu*. Some verbs have relatively high rates of 2SG agreement, but most do not, leading to an overall rate of agreement of around 30% for the random sample we analyzed. We turn now to the presentation and discussion of the results of our inferential statistical analysis.

3.2. Inferential Statistical Analysis

A mixed-effects logistic regression analysis in R was performed based on the output of a random forest and also by carrying out the step function in order to determine the variables with the greatest potential effect on the dependent variable. The random forest showed that Verb accounts for a substantial amount of the variation, and therefore should be treated as a random effect (cf. Tagliamonte and Baayen 2012). Random effects are typically independent variables whose individual values cannot be exhausted in the data set, such as speaker in studies where a subset of speakers in a given community is included (Johnson 2009). In our case, it was clearly not possible to include all Portuguese verbs in the analysis, and as shown above, the verbs that we did include in our random sample vary greatly in terms of both their frequency per million and rates of (non)agreement. For these reasons, we included Verb as a random effect. In conjunction with the use of the step function to determine significant factors, we used the random forest to build a set of explanatory models using the step-up method, which included both the factors described above as fixed effects and Verb as a random effect. We then compared those models using the ANOVA function in R to determine the best-fit model for the data set, and checked the interactions using a conditional inference tree and by running additional regression models with interaction terms. Although the variable Form initially appeared to be a potentially significant factor based on the random forest calculation, it was ultimately excluded from our models due to its high degree of collinearity with PS (low PS = present and imperfect; mid PS = future subjunctive; high PS = preterit) and also due to the fact that PS appeared to have stronger effects on the variation, as indicated by the random forest and the results of the step function. The stronger effects of PS were also corroborated by a lower AIC value for the regression model that included PS compared to the model that included Form, thus indicating that the model better fit the data.

The best-fit regression model appears in Table 11, which summarizes the significance of each factor included in this model. To orient readers to our analysis, the values in the Estimate column refer to the likelihood of non-agreement with *tu*. A positive estimate value indicates higher rates of non-agreement (ergo, lower rates of agreement) between *tu* and its corresponding verb. A negative estimate indicates higher rates of agreement (ergo, lower rates of non-agreement).

Table 11. Best-fit logistic regression model output for verbal (non)agreement with 2SG *tu* (* = $p < 0.05$).

	Estimate	**Std. Error**	**Z-Value**	***p*-Value**
(Intercept)	0.17242	0.26360	4.944	<0.001
PS (High)	0.33152	0.08051	4.118	<0.001 *
PS (Mid)	0. 70859	0. 10377	6.829	<0.001 *
Intervening (1 element)	−0.54970	0.11013	−4.991	<0.001 *
Main.aux (Main)	−0.44254	0.13753	−3.218	0.00129 *

For additional orientation, because we used contrast coding for the logistic regression, it should be noted that one variant of each fixed effect is omitted from Table 11 (e.g., PS (low) in the case of phonic salience). The likelihood of the fixed effects variants listed in the table affecting verbal (non)agreement is calculated in comparison to the variant *not* listed in the table. Thus, verbs with a high PS and mid PS both show significantly more non-agreement than verbs with a low PS. Tokens with one intervening element are compared to those with no intervening elements, and there is a statistically significant difference between these two values, such that the former show more agreement. Lastly, main verbs had significantly higher rates of agreement than auxiliary verbs.

To explore potential interactions in our data, we created conditional inference trees (cf. Tagliamonte and Baayen 2012), as shown in Figure 4, which includes only the significant predictors from the logistic regression illustrated in Table 11. As can be seen, there are no interactions between other factors and tokens with one intervening element (n = 473), which show significantly higher rates of agreement than the rest of the data set with no intervening material. A low degree of phonic salience interacts significantly with the Main vs. Aux status, such that auxiliary verbs show lower rates of 2SG agreement than main verbs. There is no interaction, however, between high/mid phonic salience and the distinction between Main vs. Aux status.[2]

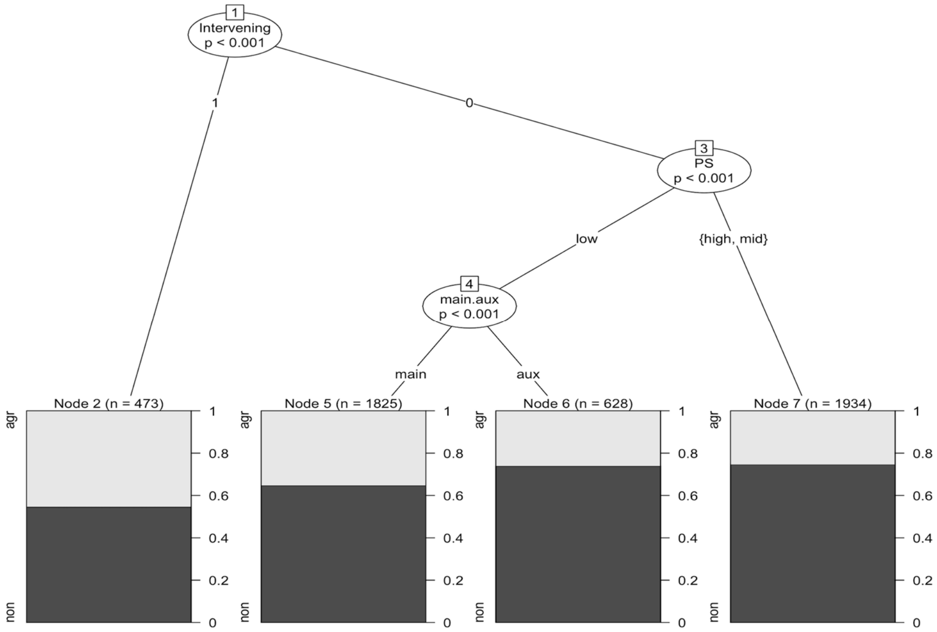


Figure 4. Conditional inference tree showing interactions between significant factors.

In this section, we have shown, using inferential statistical analysis, that (non)agreement with *tu* in our data is significantly constrained by several linguistic factors, the strongest of which, by far, is the individual verb lexeme in question. However, when we consider verb lexeme as a random effect in our models, other factors emerge as significant predictors of the variation: phonic salience, main vs. auxiliary verb status, and the presence/absence of intervening elements.

4. Discussion and Conclusions

In this article, we have shown, using a large random sample of naturally occurring data extracted from online sources in the Portuguese Web 2018 corpus, that 2SG agreement with the pronoun *tu* in this megacorpus is largely a lexically regulated phenomenon in BP, even though the overall rate of 2SG agreement is low (around 30%). Some verbs in our data are highly resistant to 2SG agreement, as shown for the particular case of *tar* (1.9%), the reduced version of the verb *estar* 'to be', which is a form typical of spoken BP or of written language that attempts to represent or mimic spoken styles. The unreduced version of this verb has a completely different behavior with regard to 2SG agreement (63.2%), insofar as it is typical of writing and other kinds of more careful linguistic production. This distinction, we believe, will be relevant to future studies using sociolinguistic interviews as their data source. Other verbs, such as *crer* 'to believe', which is highly restricted to religious contexts in our data and thus can also be assumed to be more careful in production, likewise show higher levels of agreement (65.3%). The rest of the verbs in our sample show rates ranging from a low of 10.3% in the case of *aparecer* 'to appear' to a high of 55.3% in that of *chamar* 'to call'.

In addition to the individual verb rates, there are also significant constructional effects seen in different tense/aspect forms in our data set. While previous research (e.g., Naro 1981) found important differences in third-person plural forms with respect to greater phonic salience, in the case of the second-person forms analyzed here, agreement is more common in the forms that only require *-s* to make the 2SG, i.e., present indicative and imperfect indicative forms. This pattern contrasts with the findings (Naro 1981) for third-person plural forms where the more distinct forms in the paradigm were those that tended to conserve agreement more than forms that only, e.g., required the nasalization of a final vowel to mark plurality.

The main vs. auxiliary verb status is also a significant predictor of agreement in our data for verbs that show both types of uses, such that auxiliaries display significantly less agreement with *tu* than main verbs. This reflects the greater degree of grammaticalization of the auxiliary uses of the verbs in question, which leads to a further reduction in their phonological content. An important conclusion we draw from these findings is that future studies of variable 2SG agreement in BP, no matter what the source of the data, must take both lexical and constructional factors into account for a more complete explanation of the variation.

Lastly, we found significant effects for Intervention, the variation between a total lack of intervening elements between *tu* and the following verb and one element intervening between *tu* and the verb. When an element intervenes, there is significantly more 2SG agreement than when nothing intervenes, thus suggesting that there may be additional constructional effects of *tu* + verb-3SG that are reduced when an intervening element is present. This contrasting pattern is worthy of further research.

From a theoretical perspective, these results buttress the status of BP as a language "in which the primary function of personal pronouns is carried out by independent personal pronouns that occur as arguments" instead of by affixes that appear on the verb (Bhat 2004, p. 15). As is well known, BP has reduced the paradigm of verbal morphology greatly when compared to European Portuguese, such that the main distinction is now between 1SG forms and all the other person/number combinations (Azevedo 2005; Kato et al. 2022; Perini 2002). The overwhelming tendency for 2SG pronoun *tu* to co-occur with 3SG morphology (the same morphology that co-occurs with the 2SG pronoun *você* throughout Brazil) is another indication of this reduction in the complexity of the verbal conjugation paradigm and the severing of the link between pronouns and their erstwhile verbal morphology. This reduction in the verbal paradigm has also been tied to the growing obligatoriness of subject pronouns in BP (Tarallo 1996), which in that sense seems to be following the same path as French and English, two languages with near-obligatory subjects and a highly reduced verbal morphology. These languages also have a highly fixed SVX order, which is yet another characteristic that increases their resemblance to BP (Silva 2001).

This paper also contributes to recent work across languages showing that many variable phenomena show broad variation across different lexical types (such as verbs) or constructions (such as distinct tense/aspect/mood forms or main vs. auxiliary verb uses). For Romance languages, this has been shown most prominently for the choice of the indicative vs. subjunctive mood (Poplack et al. 2018), which is best considered a case of the lexical routinization of certain main clause verbal governors. In turn, individual languages or dialects (Schwenter and Hoff 2020 for Spanish) differ in their degree of conventionalizing these patterns of mood choice. Likewise, in Spanish, the choice between the past subjunctive forms in *-ra* and *-se* is also heavily restricted to a handful of the most frequent verbs in the language (Rosemeyer and Schwenter 2019). For Portuguese, the variability between alternate forms of past participles (*particípios duplos*) for the same verbs (e.g., *pagado* vs. *pago* 'paid') has also been shown to be an overwhelmingly lexically regulated phenomenon. Schwenter et al. (2019) found that of 584 irregular participles in a corpus of over 1000 participles from both Brazilian and European Portuguese, three verbs (*pagar* 'to pay', *ganhar* 'to win', and *gastar* 'to spend') accounted for 64% (377/584) of all the irregular participles in their data set. Similar lexically and construction-specific findings for a much larger set of BP data were reported more recently by Dickinson (2022, 2024), who also found parallel patterns for Spanish. Our analysis in this paper advances evidence that 2SG (non)agreement should be included in this growing body of variable phenomena conditioned by verbal lexemes.

In our own future research, we plan to analyze, in more detail, the effects of topic and register in the updated Portuguese Web 2020 corpus, since we have already seen that there are clear differences to be found in religious contexts (reflected in our data by *crer* 'to believe'), and the updated corpus now provides more options for selecting data by register, topic, or style. In addition, we hypothesize that persistence (aka priming) likely also has strong effects in the data, which we did not code for in this study. These effects would be expected, given the low overall frequency of 2SG agreement, and in view of the general patterns of persistence in other low-frequency variants that are seemingly on their way to obsolescence (Rosemeyer and Schwenter 2019).[3]

Author Contributions: Conceptualization, S.A.S.; methodology, S.A.S. and L.M.; software, L.M.; validation, I.P. and V.C.; formal analysis, L.M.; investigation, V.C.; resources, I.P.; data curation, S.A.S. and L.M.; writing—original draft preparation, All; writing—review and editing, All; visualization, L.M. and I.P.; supervision, S.A.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Interested parties can contact the first-named author via e-mail to receive a spreadsheet with the coded data set.

Acknowledgments: We thank Kendra Dickinson, Martín Fuchs, Mark Hoff, and two anonymous reviewers for their excellent comments, which improved this paper considerably, as well as the audiences at OSUCHiLL 2023 and HLS 2023 at BYU for their useful feedback. We are also grateful to the other participants in Schwenter's Portuguese 5611 class in Spring 2023 for helping with the initial stages of the project.

Conflicts of Interest: The authors declare no conflicts of interest.

Notes

[1] This summary of the (non)agreement rates across Brazil is consonant with folk ideas about where agreement with *tu* is found, especially with regard to the phenomenon in Maranhão, which is often considered by laypersons to be the Brazilian state where the "best" Portuguese is spoken (Bagno 2009).

[2] We created an additional conditional inference tree testing interaction between the significant factors seen in Figure 4 as well as Log Frequency (binary division) and Polarity. It appeared there was a boosting effect of Frequency on Polarity, such that only those verbs with high frequency showed a sensitivity to Polarity (cf. Erker and Guy 2012). There appeared to be a further interaction effect of Frequency and main vs. auxiliary verb, but given the low number of low frequency auxiliary tokens with high/mid PS (n = 11), this potential interaction appeared dubious. Further testing of these potential interactions in a regression model, however, revealed that they were not statistically significant.

[3] It is not possible to access a large amount of the prior (or following) context in the Portuguese Web corpus (2018 or 2020 versions). However, there is a finite set of characters available for analysis before each target token, and this would provide an inherent limit on the distance between the prime (e.g., a prior token with 2SG agreement) and the target (a following token with 2SG agreement). While this limit on context is not ideal, it would at least offer a basis for consistent analysis across the full set of data (cf. Rosemeyer and Schwenter 2019).

References

Azevedo, Milton M. 2005. *Portuguese: A Linguistic Introduction*. Cambridge: Cambridge University Press.

Baayen, R. Harald. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.

Bagno, Marcos. 2009. *Não é errado falar assim!* São Paulo: Parábola.

Bhat, D. N. Shankara. 2004. *Pronouns*. Oxford: Oxford University Press.

Bybee, Joan L., Revere Perkins, and William Pagliuca. 1994. *The Evolution of Grammar*. Chicago: University of Chicago Press.

Davet, Julie, and Paula Isaias Campos-Antoniassi. 2014. Variação na concordância verbal de segunda pessoa do singular—Um estudo de fala florianopolitana. *UFSC Working Papers em Linguística* 15: 95–111. [CrossRef]

Dickinson, Kendra V. 2022. Past Participles in Spanish and Brazilian Portuguese: A Usage-Based Approach to Grammatical and Social Variation. Ph.D. dissertation, The Ohio State University, Columbus, OH, USA.

Dickinson, Kendra V. 2024. Regularization and innovation: A usage-based approach to past participle variation in Brazilian Portuguese. *Languages* 9: 52. [CrossRef]

Erker, Danny, and Gregory R. Guy. 2012. The role of lexical frequency in syntactic variability: Variable subject personal pronoun expression in Spanish. *Language* 88: 526–57. [CrossRef]

Faraco, Carlos. 1996. O tratamento "você" em português: Uma abordagem histórica. *Fragmenta* 13: 51–82. [CrossRef]

Guimarães, Ana Maria Mattos. 1979. A Ocorrência de 2" Pessoa: Estudo Comparativo Sobre o Uso de tu e Você na Linguagem Escrita. Ph.D. dissertation, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil.

Heine, Bernd. 1993. *Auxiliaries: Cognitive Forces and Grammaticalization*. Oxford: Oxford University Press.

Johnson, Daniel Ezra. 2009. Getting off the GoldVarb standard: Introducing RBrul for mixed-effects variable rule analysis. *Language and Linguistics Compass* 3: 359–83. [CrossRef]

Kato, Mary, Ana Maria Martins, and Jairo Nunes. 2022. *Português Brasileiro e Português Europeu: Sintaxe Comparada*. São Paulo: Editora Contexto.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The Sketch Engine: Ten years on. *Lexicography* 1: 7–36. [CrossRef]

Loregian, Loremi. 1996. Concordância Verbal Com o Pronome *tu* na Fala do Sul do Brasil. Master's thesis, Universidade Federal de Santa Catarina, Florianópolis, Brazil.

Loregian-Penkal, Loremi. 2004. (Re)análise da Referência de Segunda Pessoa na Fala da Região Sul. Ph.D. dissertation, Universidade Federal do Paraná, Curitiba, Brazil.

Mendes, Ronald Beline, and Lívia Oushiro. 2015. Variable number agreement in Brazilian Portuguese: An overview. *Language and Linguistics Compass* 9: 358–68. [CrossRef]

Naro, Anthony Julius. 1981. The social and structural dimensions of a syntactic change. *Language* 57: 63–98. [CrossRef]

Paredes Silva, Vera Lúcia. 2003. O retorno do *tu* ao falar carioca. In *Português Brasileiro: Contato Linguístico, Heterogeneidade e História*. Edited by Cláudia Roncarati and Jussara Abraçado. Rio de Janeiro: 7Letras, pp. 160–69.

Perini, Mário. 2002. *Modern Portuguese Grammar*. New Haven: Yale University Press.

Poplack, Shana, Rena Torres Cacoullos, Nathalie Dion, Rosane Berlinck, Salvatore Digesto, Dora LaCasse, and Jonathan Steuck. 2018. Variation and grammaticalization in Romance: A cross-linguistic study of the subjunctive. In *Manuals in Linguistics: Romance Sociolinguistics*. Edited by Wendy Ayres-Bennett and Janet Carruthers. Berlin and Boston: de Gruyter, pp. 217–52.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online: https://www.R-project.org (accessed on 14 March 2024).

Rosemeyer, Malte, and Scott A. Schwenter. 2019. Entrenchment and persistence in language change: The Spanish past subjunctive. *Corpus Linguistics and Linguistic Theory* 15: 167–204. [CrossRef]

Scherre, Maria Marta Peireira, and Maria Eugênia Lammoglia Duarte. 2016. Main current processes of morphosyntactic variation. In *The Handbook of Portuguese Linguistics*. Edited by W. Leo Wetzels, João Costa and Sergio Menuzzi. New York: Wiley and Sons, pp. 526–43.

Scherre, Maria Marta Peireira, Anthony Julius Naro, and Carolines Rodrigues Cardoso. 2007. O papel do tipo de verbo na concordância verbal no português brasileiro. *DELTA* 23: 283–317. [CrossRef]

Scherre, Maria Marta Peireira, Carolina Queiroz Andrade, and Rafael de Castro Catão. 2020. Redesenhando o mapa dos pronomes *tu/você/cê/ocê* no português brasileiro falado. In *Conquistas e Desafios dos Estudos Linguísticos na Contemporaneidade: Trabalhos do V Congresso Nacional de Estudos Linguísticos—V CONEL*. Edited by Pedro Henrique Witchs, Lucyenne Matos da Costa Vieira-Machado, CláudIA Jotto Kawachi Furlan and Mayara de Oilveira Nogueira. Porto Alegre: Editora Fi, pp. 270–76.

Scherre, Maria Marta Peireira, Edilene Patrícia Dias, Carolina Queiroz Andrade, and Germano Ferreira Martins. 2015. Variação dos pronomes "tu" e "você". In *Mapeamento Sociolinguístico do Português Brasileiro*. Edited by Marco Antonio Martins and Jussara Abraçado. São Paulo: Contexto, pp. 133–72.

Schwenter, Scott A., and Mark Hoff. 2020. Cross-dialectal productivity of the Spanish subjunctive in nominal clause complements. In *Variation and Evolution. Aspects of Language Contact and Contrast across the Spanish-Speaking World*. Edited by Sandro Sessarego, Juan J. Colomina-Almiñana and Adrián Rodríguez-Riccelli. Amsterdam: John Benjamins, pp. 11–31.

Schwenter, Scott A., Mark Hoff, Eleni Christodulelis, Chelsea Pflum, and Ashlee Dauphinais. 2019. Variable past participles in Portuguese perfect constructions. *Language Variation and Change* 31: 69–89. [CrossRef]

Schwenter, Scott A., Mark Hoff, Kendra V. Dickinson, Justin Bland, and Luana Lamberti. 2018. Experimental evidence for 2SG direct object pronoun preferences in Brazilian Portuguese. *Revista LinguíStica* 14: 259–90. [CrossRef]

Silva, Gláucia V. 2001. *Word Order in Brazilian Portuguese*. Berlin: Mouton de Gruyter.

Souza, Christiane Maria Nunes de, and Raquel Gomes Chaves. 2015. A avaliação da concordância verbal com o pronome *tu* em Florianópolis. *UFSC Working Papers em Linguística* 16: 170–89. [CrossRef]

Tagliamonte, Sali, and R. Harald Baayen. 2012. Models, forests, and trees of York English: *Was/were* variation as a case study of statistical practice. *Language Variation and Change* 24: 135–78. [CrossRef]

Tarallo, Fernando. 1996. Turning different at the turn of the century: 19th century Brazilian Portuguese. In *Towards a Social Science of Language: Papers in Honor of William Labov*. Edited by Gregory R. Guy, Crawford Feagin, Deborah Schiffrin and John Baugh. Amsterdam: John Benjamins, vol. 1, pp. 199–220.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.