UNIVERSITY OF MURCIA

International Journal of English Studies

IJES

http://revistas.um.es/ijes

Clause Initial Null Subjects in Web-based Written Language: An Analysis of Eight Varieties of English

IVÁN TAMAREDO*
Complutense University of Madrid (Spain)

Received: 15/09/2023. Accepted: 18/03/2024.

ABSTRACT
Null subjects in English(es) are a phenomenon that has recently received much attention in the specialized literature. However, most studies are based on small datasets and samples of varieties due to the difficulty of extracting null subjects from corpora. The present paper is a first step towards the automatization of the data retrieval process of null subjects and analyzes a much larger sample of cases and varieties than previous research, namely, Australian, Canadian, Jamaican, Singaporean, Nigerian, Indian, Bangladeshi and Pakistani Englishes. By focusing on referential and non-referential third person singular clause initial null and overt subjects, a variationist examination of the data is conducted by means of mixed-effects logistic regression analyses which shows that non-referential null subjects are a much more pervasive and stable phenomenon in World Englishes than their referential counterparts. In addition, a cline of varieties emerges with respect to referential null subjects: these null subjects are more frequent the more advanced varieties are in Schneider's Dynamic Model.

KEYWORDS: Null Subjects; World Englishes; Variationist Linguistics; Probabilistic Grammar; VADIS; Web-based Written Language.

1. INTRODUCTION

The overt/covert realization of pronominal subjects is a phenomenon that has figured prominently in linguistic research from a variety of theoretical frameworks. It has been extensively investigated in the domain of syntax from a generative perspective (e.g., Chomsky, 1981; Radford, 2004), in pragmatics (e.g., Y. Huang, 1992, 2000), from the perspective of

Address for correspondence: Departamento de Estudios Ingleses: Lingüística y Literatura, Facultad de Filología, Universidad Complutense de Madrid, Plaza Menéndez Pelayo s/n, 28040 Madrid; e-mail: itamared@ucm.es

IJES, vol. 24 (2), 2024, pp. 81–105
Print ISSN: 1578-7044; Online ISSN: 1989-6131
doi: 10.6018/ijes.584081

cognitive linguistics (e.g., Ariel, 1994, 2001), and it is considered a “classic sociolinguistic variable” (Bayley et al., 2012: 49; see also, Nagy et al., 2011).

Most formalist research on null subjects has adopted a cross-linguistic perspective, classifying languages according to whether this feature is attested or not, and to what degree (e.g., Biberauer et al., 2010; Jaeggli & Safir 1989). The so-called Null Subject Parameter, first postulated by Perlmutter (1971), distinguishes between canonical null subject languages, such as Spanish or Italian, where null subjects are licensed by the grammar of the language, and canonical non-null subject languages like English or Icelandic. In null subject languages, a null subject must satisfy two conditions: (i) it must be licensed by a governing head, and (ii) its grammatical features must be recoverable from this head. Rich subject-verb agreement is the head that licenses null subjects in canonical null subject languages, whereas in canonical non-null subject languages agreement is poor or non-existent and, therefore, null subjects are not licensed.

While formalist research on null subjects has focused on identifying the syntactic mechanisms that explain their occurrence, it is not entirely clear that such syntactic licensing is even necessary. Cole (2009, 2010) argues that no syntactic mechanism is enough to explain the occurrence of null subjects. Subject-verb agreement, for instance, is clearly a factor favoring null subjects cross-linguistically, but even in rich agreement languages there are verbal endings that are ambiguous and cannot identify the grammatical features of a null subject. Consider the following Spanish examples (adapted from Cole, 2009: 567)[i]:

(1) Juan$_i$ llegaba a casa. Ø$_i$ Tenía las llaves.
Juan$_i$ was arriving home. [He$_i$] had the keys.

(2) Juan$_i$ y yo$_j$ llegábamos a casa. *Ø$_{i/j}$ Tenía las llaves.
Juan$_i$ and I$_j$ were arriving home. [He$_i$/I$_j$] had the keys.

In Spanish, *tenía* is ambiguous, as it can be a first and a third person singular form. In (1), a null subject is allowed because there is only one possible referent in the immediate discourse. In (2), however, a null subject would in principle not occur because there exist two potential antecedents. Therefore, even in rich agreement languages, this is not enough in all cases to license the occurrence of null subjects, which implies that other non-syntactic mechanisms must be at play. Cole argues that the accessibility of antecedents must also be considered, given that, in cases in which agreement morphology is insufficient, only if there is an accessible antecedent in the immediate context can the grammatical features of referential null subjects be successfully recovered[ii].

Empirical research on null subjects has shown that even canonical null subject languages exhibit variation in this respect. Torres Cacoullos and Travis (2014: 22) review studies on subject expression in different null subject languages and show that, even within this group of languages, the frequency of overt/covert subjects varies substantially: for instance, whereas in Finnish subjects are overt almost in 90% of cases, Polish exhibits a much lower percentage of

21%. Such variation cannot possibly be explained by the all-or-none licensing mechanisms proposed in formalist research. In addition, even in canonical non-null subject languages like English null subjects are attested in certain contexts (Biber et al., 1999: 1104–1106; Huddleston & Pullum et al., 2002: 1540–1541; Quirk et al., 1985: 896–898). This suggests that the occurrence of null subjects is not a categorical linguistic phenomenon but instead a continuum, with languages, even canonical non-null subject ones, situated at different points of this cline. In fact, recent research has focused on non-null subject languages, especially English, to uncover the constraints influencing the variable realization on pronominal subjects (e.g., Schröter, 2019; Schröter & Kortmann, 2016; Tamaredo, 2020; Torres Cacoullos & Travis, 2014; Wagner, 2012, 2018). This is also the goal of the present paper.

1.1. Null subjects in Standard English

Null subjects can be defined as follows: the absence of an overt subject in a finite clause that could have been realized by a personal pronoun in subjective form without substantial change in meaning. In English, this definition includes examples in (3)–(5) but excludes those in (6)–(8).

(3) Ø Hope you are right. (Huddleston & Pullum et al., 2002: 1540)

(4) Sue$_i$ found the key and Ø$_i$ unlocked the door. (Huddleston & Pullum et al., 2002: 1348)

(5) Ø Show me your essay. (Quirk et al., 1985: 723)

(6) Ø Glad you think so. (Huddleston & Pullum et al., 2002: 1541)

(7) Ø Want any more beer? (Huddleston & Pullum et al., 2002: 1541)

(8) We would like Ø to stay. (Radford, 2006: 60)

Examples (3)–(5) satisfy all the conditions of our definition: the subject is absent, the clause in which it occurs is finite, and the subject could have been realized by a personal pronoun in subjective form with no major change in meaning. Examples (6) and (7), on the contrary, are not cases of null subjects as, besides subject pronouns, auxiliary verbs are also required to fill the gaps in the clauses: ***I am*** *glad you think so* and ***Do you*** *want any more beer*? Finally, in example (8) the subject gap is in the non-finite clause (*We would like you/him/etc. to stay*) and there is a change in meaning between the null and overt variants.

Furthermore, it is clear that examples (3)–(5) instantiate very different null subject constructions. Example (3), for instance, is a clause initial null subject, which is restricted to informal or casual styles and main clauses (see, for instance, Huddleston & Pullum et al., 2002: 1540). The instances in (4) and (5), on the contrary, can be found in all styles and are subject to other intralinguistic constraints: in (4), the null subject can only occur in the second clause of the coordinate structure and only when it is coreferential with the subject of the first clause; in (5), the null subject is the default option and can only be replaced by a second person pronoun (*you*).

Much empirical research on English null subjects has been recently conducted, both in standard and non-standard dialects. Schröter (2019) and Tamaredo (2020) examine null subjects in British English (GB) and Asian varieties using the *International Corpus of English* (*ICE*). Schröter focuses on spoken informal conversations and finds that GB null subjects are more likely (i) in cases of coreferential coordination, (ii) in utterance initial position, (iii) when the subject is non-referential, (iv) when it is immediately preceded by a null subject, and (v) when it is followed by a lexical verb (as opposed to primary or modal auxiliaries). Other minor effects include (vi) clause type, with main declarative clauses favoring null subjects as opposed to subordinate clauses or questions, (vi) first and third person, which increase the likelihood of a subject being null, and (vii) accessibility, that is, null subjects are more frequent when there is an accessible antecedent in the previous clause. Tamaredo (2020) investigates both spoken and written language and both informal and formal styles but focuses only on referential subjects. The findings confirm most of the pattens uncovered by Schröter (2019) for GB null subjects, with the addition of text type effects: null subjects in GB are disfavored in spoken formal language.

Torres Cacoullos and Travis (2014) examine spoken American English, focusing exclusively on first person singular subjects. They find three significant factors (coreferential coordination, intonation unit position, and persistence) and argue that three constructional schemas explain the occurrence of first person singular null subjects:

(9) Coreferential coordination: [I_i VERB *and* $Ø_i$ VERB]

(10) Intonation unit initial position: [Ø VERB …]

(11) Persistence: [$Ø_i$ VERB (*and*) $Ø_i$ VERB]

Null subjects, therefore, seem to be a relatively restricted phenomenon in standard English, occurring only in some contexts of use, in line with its status as a canonical non-null subject language. There are indications, however, that in non-standard varieties this phenomenon is more widespread. A cursory look at the *Electronic World Atlas of Varieties of English* (*eWAVE*; Kortmann et al., 2020) shows that the four morphosyntactic features included in this atlas that are related to null subjects (i.e., F43, F44, F46, and F47) have an average attestation rate of 39%, that is, 39% of the varieties in *eWAVE* exhibit some type of null subjects. Therefore, the frequency of null subjects may be higher in (some) non-standard dialects of English, and their distribution may differ with respect to standard English.

1.2. Null subjects in world Englishes

As mentioned in Section 1.1, both Schröter (2019) and Tamaredo (2020) investigated null subjects in GB and non-standard Asian varieties. In both cases, most of the factors influencing the alternation between overt and null subjects in GB were also found to have significant effects in the Asian varieties. Differences between the varieties emerged, however, in the overall frequency of null subjects and the relative weight of the factors analyzed. The findings of Schröter showed that null subjects are generally more frequent in Singapore English (SG) than in the other varieties, with Indian English (IN) and Hong Kong English (HK) exhibiting similar omission rates to those of GB, which seems to reflect the endonormative orientation of SG and the more exonormative character of IN and HK.

Tamaredo (2020) also found interesting differences between the varieties examined. Overall, null subjects were found to be more pervasive in SG than in IN and GB, not only because of its higher omission rates but also because null subjects in SG occurred more frequently outside the canonical contexts of coreferential coordination and initial position. Similarly, null subjects were attested more frequently in this variety than in IN and GB in informal styles. The analyses also provided evidence of null subjects being more prevalent in IN than in GB in non-initial positions and spoken language. The findings thus suggest a cline of varieties according to how pervasive null subjects are, with SG situated further towards the null subject pole, GB towards the opposite end, and IN in an intermediate position.

Wagner (2018) examined the occurrence of null subjects in Newfoundland English. She focused exclusively on first person subjects occurring in main clauses and excluded cases in coreferential coordination. In line with previous research, Wagner found effects of persistence, position, and accessibility. Similarly, verb semantics played a role, with perception verbs favoring the overt expression of the subject. In addition, she found interesting complexity effects: the more complex the verb phrase is, the less likely the subject is to occur in null form. Wagner operationalized complexity as a function of the number of sense units in the verb phrase: a simple present tense verb (e.g., *say* or *says*) contains one sense unit; a past tense verb (e.g., *worked*), a negated verb phrase (e.g. *don't go*), or a verb phrase with a modal auxiliary (e.g., *can eat*) contain two sense units; a verb phrase with a negative form of a modal auxiliary (e.g., *cannot come*) or a negated past tense verb (e.g., *didn't go*) contain three sense units; and so on.

Previous variationist research on null subjects in World Englishes has provided important findings regarding the frequency and patterning of null subjects in some non-standard varieties. However, these previous studies suffer from one limitation: not many varieties of English have been investigated to date because retrieving null subjects from a corpus is a highly time-consuming task that has until now been done almost completely in a manual fashion. This has inevitably led to studies in which only a relatively small sample of varieties and cases of null subjects have been examined. The main goal of the present paper is to provide a first step towards the automatization of the data retrieval process of null subjects, thus allowing linguists to gather larger datasets and explore a larger sample of dialects. In fact, eight different varieties of English, many of them varieties in which null subjects have not been investigated before, and a dataset of more than 5,000 observations of null and overt subjects are here examined. The rest of the paper deals with the data retrieval and annotation processes (Section 2), and the results of the study (Sections 3 and 4).

2. DATA AND METHOD

From a methodological perspective, the present paper lies at the crossroads between variationist research on null subjects in English (e.g., Schröter, 2019; Tamaredo, 2020; Wagner, 2018) and studies incorporating the principles of probabilistic grammar into the World Englishes paradigm (e.g., Grafmiller & Szmrecsanyi, 2018; Szmrecsanyi et al., 2016; Szmrecsanyi et al., 2019; Tamaredo et al., 2020). Therefore, the focus is not on how often speakers use a particular construction, but instead on how – that is, subject to which probabilistic constraints – they choose between 'alternate ways of saying "the same" thing' (Labov 1972: 188). Previous research has shown that, overall, varieties have a common probabilistic grammar, since the effect direction of probabilistic constraints is largely stable across varieties, but quantitative differences do exist with respect to the strength of these constraints, a situation for which Szmrecsanyi et al. (2016: 133) coined the term *probabilistic indigenization*. The aim of the present paper, therefore, is to uncover probabilistic indigenization effects in the alternation between null and overt subjects in World Englishes.

2.1. Data retrieval and annotation

The corpus selected was the *Corpus of Global Web-Based English* (*GloWbE*; Davies, 2013), which constitutes the largest resource available for the study of variation in English with almost 1.9 billion words from web pages in 20 anglophone countries. About 60% of the texts in the corpus consist of blogs, while the remaining 40% were extracted from other types of websites (Davies & Fuchs, 2015: 3–4). It is important to note that *GloWbE* is not balanced in size, since some countries contribute more words to the corpus than others. Despite this problem, it is still a very useful resource for the study of World Englishes given its large size and the large number of varieties included.

Eight national components of *GloWbE* were selected for the present study. Two of them, Australia (AU) and Canada (CA), are Inner Circle countries in developmental phase 5 of Schneider's (2007) Dynamic Model (DM). The remaining six countries belong to the Outer Circle: two are in phase 4, namely, Jamaica (JM) and SG; two, Nigeria (NG) and IN, are in phase 3; and Bangladesh (BD) and Pakistan (PK) are in phase 2.

Cases of null and overt subjects were automatically retrieved by inputting a series of search strings in *GloWbE*'s interface:

(12) Search strings for null subjects: [. VVZ] and [. VVD].

(13) Search strings for overt subjects: [. PPH1 VVZ], [. PPH1 VVD], [. PPHS1 VVZ], and [. PPHS1 VVD].

These search strings allowed us to retrieve instances of referential and non-referential third person singular clause initial null and overt subjects immediately followed by a present or past tense lexical verb. The search strings in (12) provided more than 22,000 potential cases of null subjects and those in (13) more than 380,000 hits. Given the large number of potential cases of this type of subjects retrieved from the corpus and, as discussed in Section 1, since different null subject constructions are subject to different extra- and intralinguistic constraints, the present study was restricted to this subset of cases.

In a first step, a sample of 5,000 potential cases of null subjects was randomly selected and analyzed. Out of these, 2,116 were excluded given that they were cases of first/second person and third person plural null subjects[iii]. This large number of false positives was a consequence of the search string [. VVD], which, contrary to [. VVZ], does not distinguish third person singular null subjects from other forms. The remaining 2,884 were all instances of referential and non-referential third person singular null subjects in clause initial position. In the second step, a random sample of 2,884 instances of overt subjects was also extracted from the data, resulting in a final database of 5,768 observations: half of them instances of null subjects (examples (14)–(15)), and the other half instances of overt subjects (examples (16)–(17)).

(14) Referential null subject: He$_i$ escorted her there. Ø$_i$ Told her to go straight home. (*GloWbE*, BD G, thedailystar.net)

(15) Non-referential null subject: Ø Seems to me that the preload will not be enough to prevent the binding from rotating […]. (*GloWbE*, CA B, bomberonline.com)

(16) Referential overt subject: Reilly$_i$ is very proud. He$_i$ goes to bed confident of success […]. (*GloWbE*, CA G, drykids.info)

(17) Non-referential overt subject: It seems that his call to save Pakistan is irrelevant […]. (*GloWbE*, PK B, blog.otherpakistan.org)

The dataset, therefore, contains an artificially equal proportion of null and overt subjects. This is because the analyses conducted here are all based on regression modelling techniques (see Section 2.2), which are susceptible to large class imbalances in the data (Kuhn & Johnson, 2013: 419). Class imbalance refers to situations in which one level of the dependent variable is much more frequent than the other, termed the minority variant, and it results in models which are not sensitive to the minority variant. Given that in the present paper the minority variant, namely, null subjects, is in fact the variant of interest, class imbalance poses a serious problem. Including in the model only a random sample of the majority variant, a process called down-sampling, is a relatively straightforward way of avoiding this problem, and was thus the approach adopted here.

The final dataset was annotated for a series of extra- and intralinguistic variables identified in previous research as significant determinants of null subjects in English. Four groups of factors were examined: (i) extralinguistic variables, (ii) intralinguistic variables pertaining to the target subjects, (iii) intralinguistic variables pertaining to the verbs following the target subjects, and (iv) intralinguistic variables pertaining to the linguistic context preceding the target subjects (see Table 1).

Table 1. Summary of variables analyzed.

Variable group	Variable	Levels
Extralinguistic	Variety	AU, CA, JM, SG, IN, NG, BD, PK
	Genre	General websites, blogs
Intralinguistic - subjects	Reference	Referential, non-referential
	Pronoun	It, s/he
Intralinguistic - verbs	Tense	Present, past
	Verb Semantics	Activity, aspectual, causative, communication, existence, psychological, simple occurrence
	Verb Lemma	-
Intralinguistic - context	Persistence	Null, pronoun, other
	Referential Continuity	Maintenance, partial switch, full switch

The extralinguistic variables code, first, the variety to which each instance belongs and, second, the genre in *GloWbE,* that is, general websites and blogs, the latter supposedly being more informal than the former (Davies & Fuchs, 2015: 3–4; but see Loureiro-Porto, 2017: 455–460).

The following group of variables are of an intralinguistic nature and pertain to the target subjects: while Reference codes the referential status of the subject, Pronoun indicates the pronoun that occurs, or could have occurred instead of the null subject, in subject position; given the nature of the dataset, a binary distinction between *it* and *s/he* pronouns sufficed.

The third set of factors are also intralinguistic and are related to the verbs that null/overt subjects co-occur with. First, the data was annotated for the tense of the verb, which in the present study results again in a binary distinction between present and past tense. The second factor in this group is Verb Semantics. Following Biber et al. (1999: 361–364), verbs are classified into seven semantic types: activity, aspectual, causative, communication, existence, psychological, and simple occurrence verbs. Examples (18)–(24) illustrate the semantic types distinguished.

(18) Activity: Ø **Came** to visit me in lobby various times thru the night […]. (*GloWbE,* AU B, ekilbey.blogspot.com)

(19) Aspectual: Ø **Started** to volunteer in RC organization fifteen years ago. (*GloWbE,* PK G, ifrcmedia.org)

(20) Causative: Ø **Allows** you time to focus on the basics, free the mind. (*GloWbE,* CA G, yarnharlot.ca)

(21) Communication: Ø **Told** them she'd call the doctor […]. (*GloWbE,* IN G, litlive.in)

(22) Existence: Ø **Seems** as if the Caribbean Islands are getting on top of this issue […]. (*GloWbE,* JM B, blogs.jamaicans.com)

(23) Psychological: Ø **Feels** good to have this trophy in my hands after three years […]. (*GloWbE,* BD G, news.priyo.com)

(24) Simple occurrence: Ø **Became** lieutenant-governor of Prince Edward Island in 1847 […]. (*GloWbE*, CA G, canadiana.ca)

The last variable in this group is Verb Lemma, which codes the lemma of the verb co-occurring with the target subject. This variable was included to control for possible collocational preferences of the two competing variants.

The final group of factors pertain to the linguistic context preceding the target subjects. In the case of Persistence, given that the dataset contains both referential and non-referential subjects, the latter without a referent in the preceding discourse, it is here operationalized in purely structural terms as the form of the immediately preceding subject. Finally, Referential Continuity captures accessibility effects. Three levels are distinguished: (i) reference maintenance, that is, when the antecedent of the target subject is found in the subject position of the previous clause, (ii) partial switch, that is, when the antecedent of the target subject is found in the previous clause but not in subject position, and (iii) full switch, which indicates that the antecedent of the target subject is not in the previous clause.

The fact that both referential and non-referential subjects are examined poses a difficulty because the variables Pronoun and Referential Continuity are not applicable to the latter type of subjects. First, since there are no instances of non-referential *s/he* pronouns, there is no variation regarding Pronoun. And second, as non-referential subjects, by definition, do not have a referent in the preceding discourse, Referential continuity is not applicable. These incongruences warrant dividing the dataset in two, one with only referential subjects and one with only non-referential subjects, and conducting separate analyses.

2.2. Data analysis

To uncover the probabilistic effects of the variables examined, the two datasets were analyzed by means of mixed-effects binary logistic regressions (e.g., Baayen, 2008: Ch. 7), which include both so-called random and fixed effects. Variables with repeatable levels, such as all the variables in Table 1 except Verb Lemma, are fixed in that essentially the same levels would be used to annotate a different sample of subjects. Verb Lemma, on the contrary, is a random variable because, in a new sample, some of the lemmas would be repeated but many would not occur (and new ones would).

Different mixed-effects models were fitted. First, two complete models including Verb Lemma as a random effect and all other relevant variables in Table 1 as fixed effects were computed, one for referential subjects and one for non-referential subjects. In these full models, no interactions between variables were incorporated, not even interactions between Variety and other factors. The reason was that a different and more sensitive approach was employed for this purpose, namely, the Variation-Based Distance and Similarity (VADIS) method (Szmrecsanyi et al., 2019). VADIS examines differences between varieties of English along three lines of evidence:

1. Statistical significance: Do the same variables have a statistically significant effect across varieties?
2. Effect size: Are probabilistic constraints similar with respect to the size of their effects across varieties?
3. Constraint ranking: Do the constraints have the same relative importance in all the varieties considered?

VADIS is carried out in three steps. First, a mixed-effects binary logistic regression model is fitted per variety using the same model formula. Second, a similarity score between varieties is calculated for each of the three lines of evidence. These similarity scores range from 0 to 1: the higher the values, the more similar the varieties. The score for the first line of evidence, statistical significance, is calculated as a function of the number of significant and non-significant constraints shared by the varieties. The second score, effect size, is computed as the distance between the coefficient estimates in the per-variety mixed-effects models. Finally, the third score, constraint ranking, is determined based on Spearman's rank correlation coefficients between the factor's variable importance values (see Grafmiller & Szmrecsanyi, 2018). These three similarity scores, therefore, allow us to compare the varieties' probabilistic grammar and uncover subtle probabilistic differences between them.

3. RESULTS

Given that the focus of the present study is not on how often speakers use null and overt subjects but on how they choose between these two competing variants (see Section 2), the absolute and relative frequencies of referential and non-referential null and overt subjects (see Tables A1 and A2 in the appendix) will not be discussed in depth. One important insight that can be extracted from these frequencies, however, is that, while referential subjects are overall more frequent than non-referential subjects, the latter occur much more frequently in null form. Another important point has to do with the variable Verb Semantics: non-referential subjects show a clear preference for existence verbs, particularly in the case of null subjects. For this reason, in the regression model computed based on the non-referential dataset, Verb Semantics will be treated as a binary variable, distinguishing only between existence and non-existence verbs.

3.1. Full Models of Referential and Non-referential Subjects

As mentioned in Section 2.2, two mixed-effects binary logistic regression analyses were first carried out, one for referential subjects and one for non-referential subjects, including all the relevant factors but no interactions between them. The model computed based on the referential dataset (henceforth MR) included all the variables in Table 1, with Verb Lemma as a random effect and the rest as fixed effects. Out of the seven fixed variables included, only Genre was not statistically significant. Table 2 summarizes the goodness-of-fit statistics of MR. The *C* index of concordance is a measure of how well the model discriminates between the two levels of the dependent variable: values higher than 0.8 indicate that a model has a strong predictive capacity. With a *C* value of 0.91, this is in fact the case of MR. The accuracy of the model reflects the percentage of correct predictions, that is, how many times the predictions of the model match the observed data. In this case, MR's predictions are correct in 83.17% of the cases. This accuracy value is significantly better ($p < 0.001$) than the baseline accuracy of 56.46%, that is, the percentage of the most frequent level of the dependent variable.

Table 2. Model summary: referential subjects.

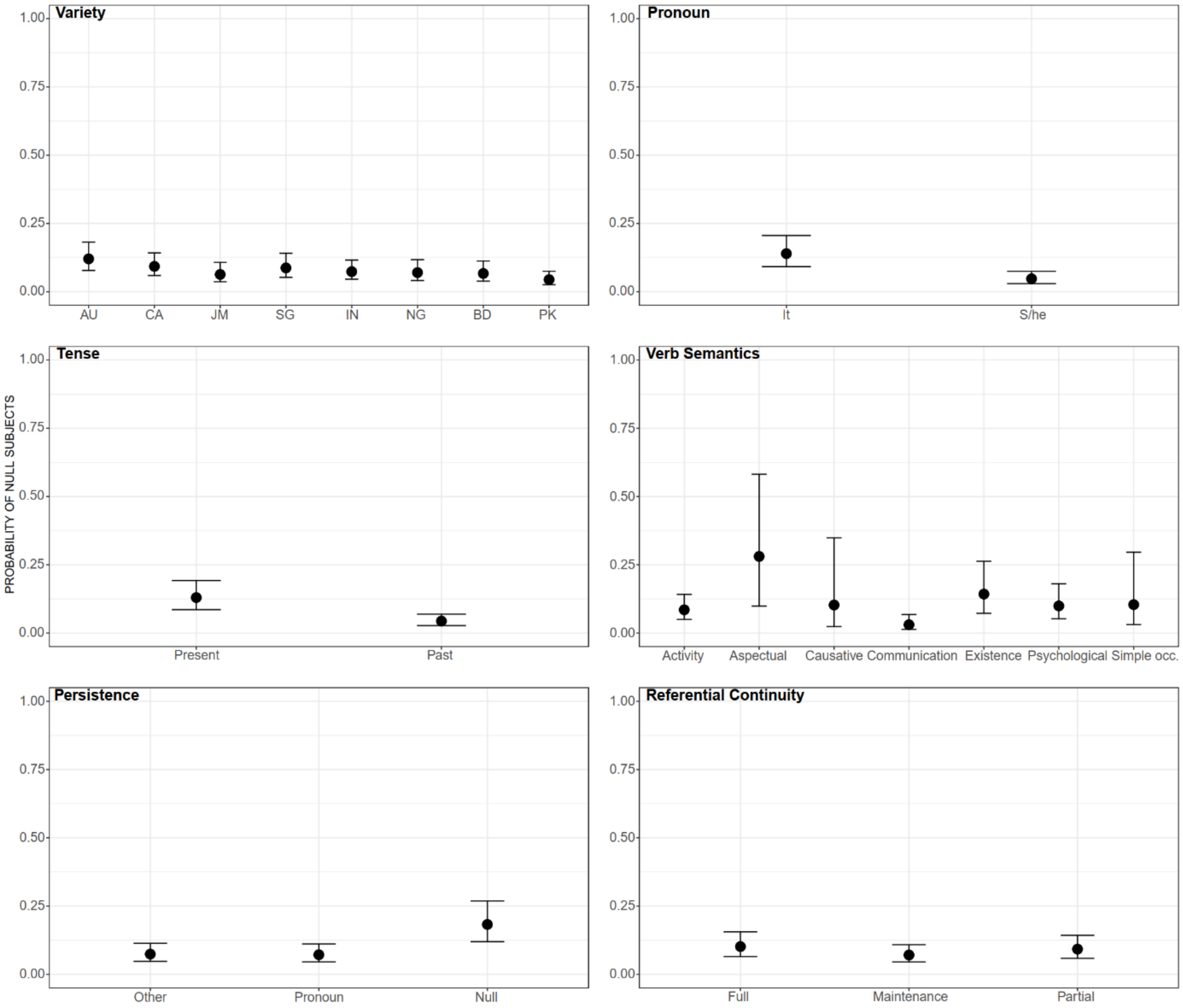
C	0.91
Accuracy	83.17%

Figure 1 shows the effects of the statistically significant factors on the choice between referential null and overt subjects (see also Table A3 in the appendix). As can be observed, referential null subjects are in all cases less probable than their overt counterparts, as in none of the graphs is the likelihood of null subjects higher than 0.5.

The factor Variety exhibits an almost linear effect, with AU at one end of the continuum and PK at the other end. In fact, the probability of a subject being null is significantly higher in AU than in the other varieties, and it is significantly lower in PK than in all other varieties except JM. CA also exhibits a significantly higher probability of null subjects than JM, IN, BD, and PK. Therefore, in varieties in phase 5 of Schneider's DM subjects are more likely null than in varieties in earlier phases of development. PK, a phase 2 variety, is the dialect where null subjects are less probable, and in between we find varieties in phase 4 and 3, as well as BD.

Regarding the intralinguistic variables, the effect of Pronoun is clear: *it* null subjects are significantly more probable than *s/he* null subjects. Tense also has a significant effect, with null subjects being more likely with present tense than with past tense verbs. Verb Semantics has an influence as well: as can be seen in Figure 1, the likelihood of a subject being null is similar with all types of verbs except two: aspectual verbs seem to favor null subjects more strongly than other verbs, while communication verbs seem to inhibit the omission of the subject. Persistence shows the expected effect, with null subjects being more probable when the subject of the previous clause is also null. Finally, Referential Continuity exhibits an unexpected distribution: null subjects are significantly more likely when there is a partial or full referential switch from the previous clause. However, the size of the effect is weak, as shown by the similar probabilities of null subjects in all three levels of this variable.

Figure 1. Fixed effects: referential subjects.



The formula of the model computed based on the non-referential dataset (henceforth MNR) includes Verb Lemma as a random effect and all other factors except Pronoun and Referential Continuity as fixed effects. In this case, none of the extralinguistic variables emerged as significant, which means that there are no differences between varieties or genres with respect to the probability of null subjects. Table 3 summarizes the goodness-of-fit statistics of MNR. With a *C* value of 0.85, MNR also has a strong predictive capacity. The accuracy of the model, 83.31%, is also significantly better than the baseline of 78.29 ($p <$ 0.001).

Print ISSN: 1578-7044; Online ISSN: 1989-6131

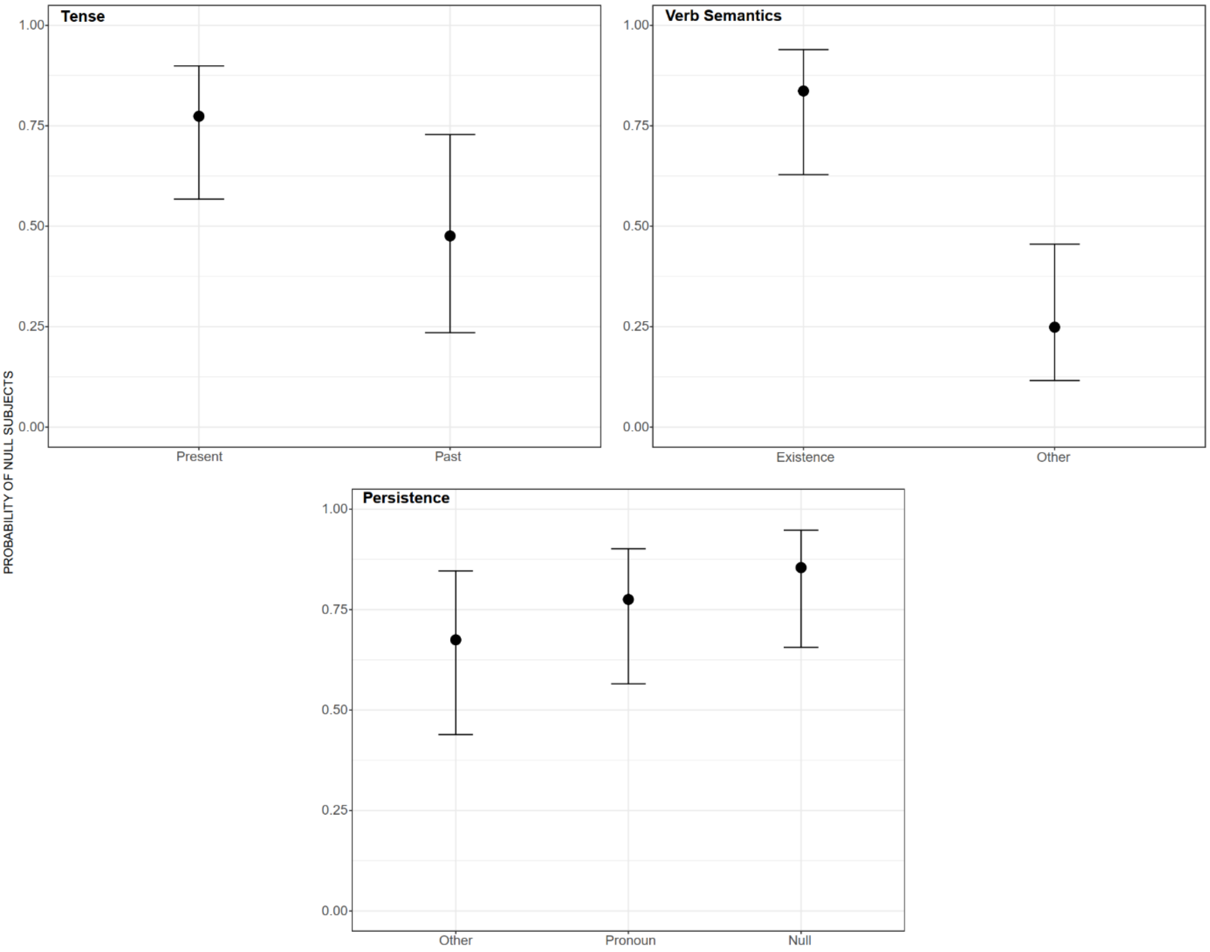
Table 3. Model summary: non-referential subjects.

C	0.85
Accuracy	83.31%

Figure 2 plots the effects of the significant factors of MNR (see also Table A3 in the appendix). In this case, the predicted probability of null subjects is overall much higher than that of their referential counterparts.

With respect to the intralinguistic factors, the effect of Tense is the same as in MR, with non-referential null subjects being favored by present tense verbs. Verb Semantics also plays a significant role: as suggested by the frequencies discussed above, null subjects are much more likely when they co-occur with existence verbs. Finally, Persistence also has a similar effect in MNR than in MR, with a slight difference: in the case of non-referential subjects, both a preceding null subject and a preceding pronominal subject significantly increase the likelihood of the subject being null. However, the differences are minor, particularly considering the large confidence intervals.

Figure 2. Fixed effects: non-referential subjects.



The models described in this section confirm many of the tendencies found in previous research (see Section 1). However, some interesting, unexpected patterns were uncovered (see Section 4 for a discussion of the results). Among them, it is worth noting the insignificance of Variety in MNR, which suggests that varieties do not differ in terms of the choice between non-referential null and overt subjects. This finding is particularly relevant for the VADIS analyses: there seems to be no point in applying the VADIS method to non-referential subjects, as they appear to be stable across varieties. Therefore, Section 3.2 will focus exclusively on the referential dataset.

3.2. VADIS analysis of referential subjects

The output of the VADIS method, shown in Table 4, consists of three similarity scores that summarize the probabilistic differences between the varieties examined[iv].

Table 4. Similarity scores.

Statistical significance	Effect size	Constraint ranking
0.801	0.280	0.455

The similarity scores indicate that, while varieties differ substantially as regards the ranking of constraints and, especially, the size of their effects, they do not differ much in which constraints have a statistically significant effect in the choice between the competing variants. The first line of evidence, however, allows us to restrict the effects found in MR to some varieties. First, the inhibiting effect on null subjects of communication verbs is limited to AU, IN, and NG, as only in these three varieties does it reach statistical significance. Similarly, the facilitating effect of aspectual verbs seems to be restricted to CA. Second, the VADIS method identifies one variety in which null subjects are not favored by present tense verbs, namely JM. Likewise, persistence effects do not play a role in SG or PK. Third, Referential Continuity was significant in MR, but the direction of the effect was unexpected, with reference maintenance inhibiting the occurrence of null subjects; this unexpected effect is, however, restricted only to IN. Finally, in NG and BD Pronoun does not influence the choice between null and overt pronouns.

Regarding the second line of evidence, aspectual verbs have a much stronger facilitating effect in CA than in the other varieties. In the same vein, past tense verbs inhibit null subjects more strongly in BD than in the other varieties and less so in SG. Regarding Referential Continuity, in IN reference maintenance has a much stronger hindering effect on the occurrence of null subjects than in the other varieties and, in the case of Persistence, a preceding null subject has a much stronger facilitating effect in JM as compared to the other varieties, while in CA the effect is significantly weaker. Finally, *s/he* subjects are substantially less likely in null form in JM and SG and more likely in AU than in the other varieties.

With respect to the last line of evidence, constraint ranking, the differences between varieties can be summarized as follows. Tense has a much stronger impact on the variation between referential null and over subjects in BD than the average and, similarly, Persistence is more important in IN and NG and less so in CA. Finally, Pronoun splits the varieties in two groups: in CA, JM, SG, and PK it has a stronger impact than in AU, IN.

4. DISCUSSION

One of the main findings of the present study has to do with the incidence of referential and non-referential null subjects in World Englishes. Even though the current data does not allow us to extract conclusions about the overall frequency of these phenomena, it does allow us to compare their pervasiveness in relative terms. The results clearly show that non-referential null subjects are a more pervasive phenomenon in English than referential null subjects in all the varieties examined, a finding shared for the most part by Schröter (2019). Therefore, it seems that speakers of English(es) prefer to omit subjects when they are empty from a semantic point of view, rather than when they have a referential function. Moreover, this appears to be a stable feature in English rather than a characteristic of one or two varieties.

The variable Genre, which distinguished between general websites and blogs, does not play a role in the alternation between null and overt subjects. This finding was unexpected, given that Tamaredo (2020) found that null subjects were preferred in informal genres, and blogs are supposedly more informal than general websites. However, Loureiro-Porto (2017: 460) showed that blogs and general websites in *GloWbE* are in fact very similar in terms of their degree of informality, a conclusion supported by the findings of the present study.

One of the main differences found between referential and non-referential subjects, besides their frequency, was the role of Variety. Whereas Variety was a significant factor in MR, and a series of probabilistic differences between the varieties emerged in the VADIS analysis, non-referential subjects seemed to be stable in this respect. In the case of referential subjects, a cline of varieties emerged: the phase 5 variety AU favored referential null subjects as compared to the other varieties, while the contrary was true in the case of the phase 2 variety PK. In between these two, CA and SG displayed higher probabilities of null subjects than the remaining varieties. Therefore, it can be preliminarily concluded that the more advanced varieties are in Schneider's DM, the higher the probability of referential null subjects. In the specialized literature on linguistic complexity, it has long been recognized that transparency, or a one-to-one mapping between forms and meanings, results in simpler linguistic systems and that simplicity, other things being equal, is indeed favored by L2 users of a language (e.g., Steger & Schneider, 2012). Therefore, the fact that more advanced varieties in Schneider's DM exhibit a higher frequency of referential null subjects than less advanced varieties should not come as a surprise: referential subject omission results in less transparent structures in which there is not a one-to-one mapping between forms and meanings since the subject is not overtly expressed. The more equal distribution across varieties of non-referential null subjects can also be understood as a consequence of complexity considerations, given that non-referential subjects are semantically empty and their omission does not entail a violation of the one-meaning-one-form principle.

Tense exhibited similar effects in both MR and MNR: present tense verbs favored null subjects as compared to past tense verbs. Two different mechanisms may account for the preference of null subjects for present tense verbs. As argued by Cole (2009, 2010; see Section 1), subject-verb agreement plays a role in the alternation between null and overt subjects in those languages in which such agreement exists by aiding in the recovery of the antecedent. Therefore, the fact that null subjects are favored by present tense verbs could be a result of the facilitating effect of subject-verb agreement, given that all present tense verbs in the dataset contain the English third person singular present -(*e*)*s* suffix. On the other hand, Wagner (2018; see Section 1.2,) found that in Newfoundland English null subjects are inhibited by complex verb phrases. In her analysis, past tense lexical verbs are more complex than present tense verbs since the former contain two sense units (root + past) and the latter only one. Therefore, the tense effects uncovered in the present study could also reflect complexity effects. The nature of the present datasets, however, does not allow us to test these two competing explanations.

Persistence effects were also found in both MR and MNR, in line with previous research: a preceding null subject increases the likelihood of a subject being null, as this is the case with both referential and non-referential subjects. It seems, therefore, that in most varieties a purely structural type of persistence influences the choice between null and overt subjects. In those varieties in which persistence does not have a significant role, namely, SG and PK in the case of referential subjects, it could be that a preceding null subject increases the probability of a subsequent subject being null only if the two subjects are coreferential. SG, in particular, is known for establishing topic chains in which, once the topic of a stretch of discourse is established, all subsequent references to the topic are made by means of reduced forms, including null elements (Schröter, 2019: 211–213). In topics chains, therefore, what determines the form of a subject is not the form of the immediately preceding subject (especially if they are not coreferential), but instead whether the subject refers to the previously established topic: if it does, then the subject is most likely in null form. Example (25) shows a topic chain in SG:

> (25) [Justin Lee]$_i$ # justintech65.org # A weird, passionate geek for technology with an undying love to pick at every flaw that will hopefully improve technology all around. Ø$_i$ Owns a lovely MacBook Pro with 8gigs of ram, Ø$_i$ involved intimately with Linux and Ø$_i$ works closely with Microsoft technologies. Ø$_i$ Wants to own an iPhone 4, Ø$_i$ owns an

HTC Hero Android phone, Nexus One, $Ø_i$ once owned an iPhone 3g, $Ø_i$ still owns a 1st gen Sony Ericsson W800i. (*GloWbE*, SG G, tech65.org)

In (25), the topic is clearly established at the beginning of the chain, which makes it unnecessary for all subsequent subjects referring to this topic to be in overt form. Topic chains may account for a substantial number of cases of null subjects, not only in SG but also is other varieties. However, it is difficult to operationalize topic chains as a variable for quantitative analysis, so they are not commonly included in quantitative studies of null subjects. In fact, to the best of the author's knowledge, there are no studies which have quantified the influence of topic chains on the alternation between null and overt subjects; topic chains have so far been dealt with in a qualitative manner by describing them and showing examples but it is still unclear how important they are in the occurrence of null subjects.

The semantics of the verb co-occurring with the subject also emerged as a determinant of variation in both MR and MNR, although the patterns differed. In the case of referential subjects, aspectual verbs facilitated the occurrence of null subjects in CA and communication verbs inhibited them in AU, IN, and NG. Wagner (2018) found that, in Newfoundland English, perception verbs (in which she included communication verbs and psychological verbs, among others) favored the overt expression of the subject. The findings of the present study suggest that this is also the case in AU, IN, and NG, although here only communication, but not psychological, verbs exhibited such inhibiting effects, a tendency that is probably best explained as different idiosyncratic lexical preferences for the null or overt variants in different varieties. On the other hand, non-referential null subjects occurred essentially with existence verbs. This preference of non-referential null subjects for existence verbs like *seem*, *turn* (*out*), *look* (*like*), and *sound* (*like*) is not surprising, as many of these verbs are undergoing grammaticalization processes on their way to become parenthetical expressions (e.g., López-Couso & Méndez-Naya, 2014; Serrano-Losada, 2017). Examples (26) and (27) show two cases of non-referential null subjects with the existence verbs *turn* (*out*) and *look* (*like*):

(26) […] I was testing out all sorts of volumising hair products in an attempt to get some oomph into my fine, flat hair. Ø Turns out, cutting it all off made it fuller and bouncier than ever before […]. (*GloWbE*, AU B, theplasticdiaries.com)

(27) There were at least 3 interesting news in yesterday's press. Ø Looks like nobody is at all bothered about these new developments. (*GloWbE,* BD G, rumiahmed.wordpress.com)

The parenthetical function of these expressions is particularly clear in (26), where *turns out* is separated from the rest of clause by a comma. In both cases, the subject pronoun *it* is omitted, as it is non-referential and, therefore, semantically empty.

Finally, the variables Pronoun and Referential Continuity were only relevant in the case of referential subjects. Referential null subjects were more common when they could have been replaced by *it* than by *s/he*. Although further research is needed to clarify this issue, the fact that referential *it* subjects are more commonly null than other subjects may be due to the influence of non-referential *it* subjects, which occur very frequently in null form. Lastly, Referential Continuity exhibited an unexpected distribution, with null subjects being less common in cases in which the antecedent of the subject was the subject of the immediately preceding clause. As in previous studies, however, the effect strength of this variable was rather weak, and the VADIS analysis showed that, in fact, it was only significant in IN. As mentioned in Section 2.1, Referential Continuity was included in the analysis to capture accessibility effect. Accessibility, however, is a complex cognitive notion influenced by numerous factors, for instance, the saliency of the antecedent, its animacy, its syntactic status (i.e., subjects are more accessible than other constituents, precisely the type of accessibility effects that Referential Continuity captures), or the number of potential antecedents in the surrounding linguistic context, among others (e.g., Ariel, 2001). Referential Continuity is the accessibility factor most commonly investigated in studies on null subjects because it is the easiest one to annotate and quantify, but it may well be the case that this is not enough to account for accessibility effects, thus its commonly reported weak influence.

5. CONCLUSIONS AND FUTURE STUDIES

The goals of the present paper were twofold. First, we aimed to contribute to the automatization of the data extraction process of research on null subjects in English, given that until now the most common approach had been that of manually identifying the instances of null subjects in a corpus. A series of search strings were proposed, which served to automatically extract from *GloWbE* a relatively large number of instances of this phenomenon. The precision of these search strings was relatively high in most cases. If research on English null subjects aims to be based on representative data samples, future studies should employ an approach such as the one proposed in the present paper, thus first delimiting the contexts in which null subjects occur and then translating those contexts into search strings that automatically retrieve instances of this linguistic feature from corpora.

A second goal of the study was to uncover the probabilistic grammar underlying the alternation between null and overt subjects in a balanced set of varieties of English and if differences existed between the varieties in this respect. In addition, this linguistic phenomenon was examined in written web-based language, a text type in which null subjects had so far not been investigated. The results confirmed most of the tendencies uncovered in previous research, but it also revealed additional interesting patterns. One of the main findings of the present study was the relative heterogeneity of referential subjects across varieties as compared to the homogeneity of non-referential subjects. It seems that, as shown by Szmrecsanyi et al. (2016), syntactic alternations are not equally sensitive to probabilistic indigenization effects, not even in the case of highly related alternations such as the ones analyzed here (see also Tamaredo et al., 2020).

In comparison with previous studies on English null subjects, the dataset analyzed in the present paper contained a much larger sample of varieties and observations. However, precisely because of the number of varieties analyzed, the size of the dataset should still be enlarged to achieve a more representative sample. Two ways in which the dataset can be expanded are (i) by including subjects in other persons (and numbers) besides third person singular subjects, which would incidentally increase the precision of the search strings employed, and (ii) by extracting also examples of null and overt subjects followed by non-lexical verbs. Even though the range of variables analyzed was substantial and in line with previous studies on null subjects, enlarging the dataset in these two ways would also allow us to test the effects of other constraints. Including also subjects followed by non-lexical verbs, for instance, would enable us to replicate previous findings in the literature regarding the inhibiting effect of modal and non-modal auxiliaries on the occurrence of null subjects.

Finally, in the mixed-effects binary logistic regression analyses described in Section 3, Verb Lemma was included in the random effect structure of the models to control for the possible influence of idiosyncratic collocational preferences of the null and overt variants. An in-depth analysis of these collocational preferences, however, was not attempted but it would be interesting to examine how influential these are in the choice between null and overt subjects. In fact, the author is currently in the process of comprehensively analyzing the role of individual lexical items in the alternation between null and overt subjects by resorting to collostructional analytical techniques that measure the degree of association between words and constructions (e.g., Gries & Stefanowitsch, 2004). This line of research is expected to shed light on the degree of lexical specificity of the two competing variants across varieties, thus achieving a more complete understanding of the phenomenon of subject omission in Englishes.

ACKNOWLEDGEMENTS

This research was conducted with the financial support of the Spanish Ministry of Science, Innovation and Universities (grants PID2020-114604GB-100 and PID2023-146887NB-I00). Thanks are also due to two anonymous reviewers for their fruitful comments.

NOTES

[i] The symbol Ø is used throughout the paper to represent the position of the null subject in the clause.

[ii] In example (2), in particular, the low accessibility of the antecedent of the subject of the second clause is caused by the fact that there are two potential antecedents in the immediate context (*Juan* and *yo*) and, especially, the change of reference from one clause to the next: if the subject of the second clause had been the same as that of the first clause (first person plural), then the second subject would have most likely occurred in null form. The author thanks an anonymous reviewer for pointing out this issue.

[iii] Other minor false positives included duplicates, fairly common in *GloWbE*, and spelling mistakes, as in *[...] That's puts us on a high plane and . Gives. Peace a d healing* (*GloWbE*, AU B, blogs.crikey.com.au)

[iv] All the per-variety models computed have a strong discriminatory power, with *C* values and percentages of correct predictions higher than 0.90 and 85% in most cases (see Table A5 in the appendix).

REFERENCES

Ariel, M. (1994). Interpreting anaphoric expressions: A cognitive versus a pragmatic approach. *Journal of Linguistics, 30(1)*, 3–42. doi: https://doi.org/10.1017/S0022226700016170

---. (2001). Accessibility theory: An overview. In T. Sanders, J. Schliperoord & Wilbert Spooren (Eds.), *Text Representation: Linguistic and Psycholinguistic Aspects* (pp. 29–87). Amsterdam/Philadelphia: John Benjamins.

Chomsky, N. (1981). *Lectures on Government and Binding*. Dordrecht/Providence: Foris Publications.

Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge/New York: Cambridge University Press.

Bayley, R., Cardenas, N. L., Trevino Schouten, B. & Velez Salas, C. M. (2012). Spanish dialect contact in San Antonio, Texas: An exploratory study. In K. Geeslin & M. Diaz-Campos (Eds.), *Selected Proceedings of the 14th Hispanic Linguistics Symposium* (pp. 48–60). Somerville, CA: Cascadilla Proceedings Project.

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Biberauer, T., Holmberg, A. & Roberts, I. (Eds.). (2010). *Parametric Variation: Null Subjects in Minimalist Theory*. Cambridge: Cambridge University Press.

Cole, M. D. (2009). Null subjects: A reanalysis of the data. *Linguistics, 47(3)*, 559–587. doi: https://doi.org/10.1515/LING.2009.019

---. (2010). Thematic null subjects and accessibility. *Studia Linguistica, 64(3)*, 271–320. doi: https://doi.org/10.1111/j.1467-9582.2010.01172.x

Davies, Mark. (2013*). Corpus of Global Web-Based English: 1.9 Billion Words from Speakers in 20 Countries (GloWbE)*. Retrieved 21 August, 2023 from https://corpus.byu.edu/glowbe/.

Davies, M. & Fuchs, R. (2015). Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-based English Corpus (GloWbE). *English World-Wide, 36(1)*, 1–28. doi: https://doi.org/10.1075/eww.36.1.01dav

Grafmiller, J. & Szmrecsanyi, B. (2018). Mapping out particle placement in Englishes around the world: A case study in comparative sociolinguistic analysis. *Language Variation and Change, 30(3)*, 385–412. doi: https://doi.org/10.1017/S0954394518000170

Gries, S. Th., & Stefanowitsch, A. (2004). Extending collostructional analysis: A corpus-based perspective on ‘alternations’. *International Journal of Corpus Linguistics, 9(1)*, 97–129. doi: https://doi.org/10.1075/ijcl.9.1.06gri

Huang, C.-T. J. (1984). On the distribution and reference of empty pronouns. *Linguistic Inquiry, 15(4)*, 531–574.

Huang, Y. (1992). Against Chomsky’s typology of empty categories. *Journal of Pragmatics, 17(1)*, 1–29. doi: https://doi.org/10.1016/0378-2166(92)90026-8

---. (2000). *Anaphora: A cross-linguistic approach*. Oxford/New York: Oxford University Press.

Huddleston, R., Pullum, G. K., Bauer, L., Birner, B., Briscoe, T., Collins, P., Denison, D., Lee, D., Mittwoch, A., Nunberg, G., Palmer, F., Payne, J., Peterson, P., Stirling, L. & Ward, L. (2002). *The Cambridge Grammar of the English Language*. Cambridge/New York: Cambridge University Press.

Jaeggli, O. & Safir, K. J. (1989). The null subject parameter and parametric theory. In O. Jaeggli & K. J. Safir (Eds.), *The Null Subject Parameter* (pp. 1–44). Dordrecht/Boston/London: Kluwer Academic Publishers.

Kortmann, B., Lunkenheimer, K. & Ehret, K. (Eds.) (2020). *The Electronic World Atlas of Varieties of English*. Zenodo. Retrieved 21 August, 2023 from http://ewave-atlas.org.

Kuhn, M. & Johnson, K. (2013). *Applied Predictive Modelling*. New York: Springer-Verlag New York.

López-Couso, M. J. & Méndez-Naya, B. (2014b). From clause to pragmatic marker: A study of the development of *like*-parentheticals in American English. *Journal of Historical Pragmatics, 15(1)*, 66-91. doi: https://doi.org/10.1075/jhp.15.1.03lop

Loureiro-Porto, L. (2017). ICE vs GloWbE: Big data and corpus compilation. *World Englishes, 36(3)*, 448–470. doi: https://doi.org/10.1111/weng.12281

Nagy, N. G., Aghdasi, N., Denis, D. & Motut, A. (2011). Null subjects in heritage languages: Contact effects in a cross-linguistic context. *University of Pennsylvania Working Papers in Linguistics, 17(2)*, 135–44.

Perlmutter, D. (1971). *Deep and Surface Constraints in Generative Grammar*. New York: Holt, Rinehart and Winston.

Quirk, R., Greenbaum, R., Leech, G. & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London/New York: Longman.

Radford, A. (2004). *Minimalist Syntax: Exploring the Structure of English*. Cambridge: Cambridge University Press.

---. (2006). *Minimalist Syntax Revisited*. Retrieved 21 August, 2023 from http://courses.essex.ac.uk/lg/lg514.

Schneider, E. W. (2007). *Postcolonial English: Varieties around the World*. Cambridge/New York: Cambridge University Press.

Schröter, V. (2019). *Null Subjects in English: A Comparison of British English and Asian Englishes*. Berlin/Boston: De Gruyter Mouton.

Schröter, V. & Kortmann, B. (2016). Pronoun deletion in Hong Kong English and Colloquial Singaporean English. *World Englishes, 35(2)*, 221–241. doi: https://doi.org/10.1111/weng.12192

Serrano-Losada, M. (2017). Raising *turn out* in Late Modern English: The rise of a mirative predicate. *Review of Cognitive Linguistics, 15(2)*, 411–437. doi: https://doi.org/10.1075/rcl.15.2.05ser

Steger, M. & Schneider, E. W. (2012). Complexity as a function of iconicity: The case of complement clause constructions in New Englishes. In B. Kortmann & B. Szmrecsanyi (Eds.), *Linguistic Complexity: Second Language Acquisition, Indigenization, Contact* (pp. 156–191). Berlin/Boston: De Gruyter.

Szmrecsanyi, B., Grafmiller, J., Heller, B. & Röthlisberger, M. (2016). Around the world in three alternations: Modeling syntactic variation in varieties of English. *English World-Wide, 37(2)*, 109–137. doi: https://doi.org/10.1075/eww.37.2.01szm

Szmrecsanyi, B., Grafmiller, J. & Rosseel, L. (2019). Variation-based distance and similarity modeling: A case study in World Englishes. *Frontiers in Artificial Intelligence, 2*, 23. doi: https://doi.org/10.3389/frai.2019.00023

Tamaredo, I. (2020). Complexity, Efficiency, and Language Contact: Pronoun Omission in World Englishes. Bern: Peter Lang.

Tamaredo, I., Röthlisberger, M., Grafmiller, J. & Heller, B. (2020). Probabilistic Indigenization Effects at the Lexis–Syntax Interface. English Language and Linguistics, *24(2)*, 413–440. doi: https://doi.org/10.1017/S1360674319000133

Torres Cacoullos, R. & Travis, C. E. (2014). Prosody, priming and particular constructions: The patterning of English first-person singular subject expression in conversation'. *Journal of Pragmatics 63*, 19–34. doi: https://doi.org/10.1016/j.pragma.2013.08.003

Wagner, S. (2012). *Null subjects in English: Variable rules, variable language?* Postdoctoral Dissertation, Chemnitz University of Technology, Germany.

---. (2018). *Never Saw One* – First-person null subjects in spoken English. *English Language and Linguistics, 22(1)*, 1–34. doi: https://doi.org/10.1017/S1360674316000216

APPENDIX

Table A1. Absolute and relative frequencies of referential null and overt subjects.

Factor		**Null**		**Overt**		**Total**
		N	**%**	**N**	**%**	
Variety	AU	674	56.31	523	43.69	1,197
	CA	512	46.21	596	53.79	1,108
	JM	94	33.45	187	66.55	281
	SG	180	48.65	190	51.35	370
	IN	316	40.15	471	59.85	787
	NG	97	32.55	201	67.45	298
	BD	86	31.62	186	68.38	272
	PK	85	22.25	297	77.75	382
Genre	General	1,378	42.17	1,890	57.83	3,268
	Blogs	666	46.67	761	53.33	1,427
Pronoun	It	1,631	66.52	821	33.48	2,452
	S/he	413	18.41	1,830	81.59	2,243
Tense	Present	1,697	60.85	1,092	39.15	2,789
	Past	347	18.21	1,559	81.79	1,906
V. Semantics	Activity	773	43.85	990	56.15	1,763
	Aspectual	68	40	102	60	170
	Causative	81	52.60	73	47.40	154
	Communication	122	14.61	713	85.39	835
	Existence	449	64.05	252	35.95	701
	Psychological	493	53.24	433	46.76	926
	Simple occ.	58	39.73	88	60.27	146
Persistence	Null	399	56.52	307	43.48	706
	Pronoun	687	41.14	983	58.86	1,670
	Other	958	41.31	1,361	58.69	2,319
R. Continuity	Maintenance	683	32.04	1,449	67.96	2,132
	Partial	557	51.53	524	48.47	1,081
	Full	804	54.25	678	45.75	1,482
Total		2,044	43.54	2,651	56.46	4,695

Table A2. Absolute and relative frequencies of non-referential null and overt subjects.

Factor		**Null**		**Overt**		**Total**
		N	**%**	**N**	**%**	
Variety	AU	291	82.44	62	17.56	353
	CA	251	78.68	68	21.32	319
	JM	49	74.24	17	25.76	66
	SG	76	80.85	18	19.15	94
	IN	94	74.60	32	25.40	126
	NG	19	67.86	9	32.14	28
	BD	25	69.44	11	30.56	36
	PK	35	68.63	16	31.37	51
Genre	General	514	77.29	151	22.71	665
	Blogs	326	79.90	82	20.10	408
Tense	Present	796	82.92	164	17.08	960
	Past	44	38.94	69	61.06	113
V. Semantics	Activity	71	44.10	90	55.90	161
	Aspectual	0	0	3	100	3
	Causative	3	33.33	6	66.67	9
	Communication	0	0	0	0	0
	Existence	747	86.86	113	13.14	860
	Psychological	19	57.58	14	42.42	33
	Simple occ.	0	0	7	100	7
Persistence	Null	93	83.04	19	16.96	112
	Pronoun	435	82.08	95	17.92	530
	Other	312	72.39	119	27.61	431
Total		840	78.29	233	21.71	1,073

Table A3. Results of the MR model.

Fixed effects				
Regressor	**Estimate**	**Std. error**	***Z***	***p***
Intercept	-0.875	0.311	-2.812	0.005
Variety = CA	-0.292	0.119	-2.446	0.014
Variety = JM	-0.705	0.197	-3.581	0.001
Variety = SG	-0.361	0.168	-2.148	0.032
Variety = IN	-0.546	0.132	-4.130	0.001
Variety = NG	-0.595	0.186	-3.207	0.001
Variety = BD	-0.649	0.197	-3.294	0.001
Variety = PK	-1.083	0.185	-5.852	0.001
Verb Semantics = Aspectual	1.433	0.681	2.105	0.035
Verb Semantics = Causative	0.206	0.791	0.261	0.794
Verb Semantics = Communication	-1.087	0.409	-2.661	0.008
Verb Semantics = Existence	0.582	0.423	1.375	0.169
Verb Semantics = Psychological	0.170	0.363	0.467	0.640
Verb Semantics = Simple occ.	0.222	0.639	0.348	0.728
Tense = Past	-1.182	0.103	-11.458	0.001
Ref. Continuity = Maintenance	-0.395	0.105	-3.782	0.001
Ref. Continuity = Partial	-0.106	0.118	-0.899	0.369
Persistence = Pronoun	-0.034	0.095	-0.358	0.721
Persistence = Null	1.026	0.126	8.169	0.001
Pronoun = s/he	-1.179	0.111	-10.661	0.001
Random effects				
Predictor		**Variance**		
Verb Lemma		3.807		

Table A4. Results of the MNR model.

Fixed effects				
Regressor	**Estimate**	**Std. error**	***Z***	***p***
Intercept	1.413	0.573	2.466	0.014
Verb Semantics = Other	-2.739	0.630	-4.348	0.001
Tense = Past	-1.326	0.292	-4.541	0.001
Persistence = Pronoun	0.509	0.191	2.664	0.008
Persistence = Null	1.041	0.349	2.984	0.003
Random effects				
Predictor		**Variance**		
Verb Lemma		2.413		

Table A5. Summary of per-variety models.

Variety	***C***	**Accuracy**
AU	0.90	82.2%
CA	0.90	83.8%
JM	0.97	90.1%
SG	0.92	84.9%
IN	0.91	85.5%
NG	0.94	88.6%
BD	0.92	86.4%
PK	0.91	88%