Null Subjects in Colloquial English: A Corpus Study

Michael George Phillips

Syracuse University
1. Introduction

A common observation regarding the world’s languages is the fact that some languages can lack overt subject arguments, particularly when the subject argument is pronominal. Such languages are typically identified as null-subject languages (henceforth NSLs). Some examples of NSLs include Italian, Spanish, Greek, Korean, and Japanese, among many others. An example of subject pronoun drop (SPD) from Italian is shown in (1):

(1) Italian pro-drop

Parlo francese.
speak-1SG.PRS French
“I speak French.”

We can observe in (1) that the pronoun subject argument io “I” is not overtly present, which is consistent with the classification of Italian as an NSL. English, on the other hand, is known as a canonical example of a non-NSL, as it is observed that all English sentences must have a subject overtly present. Even in cases where there is no logical subject with respect to semantic roles, a so-called expletive pronoun\(^1\) is inserted, as demonstrated in the examples in (2):

(2) Expletive pronouns in English

a. There appears to be something in your coffee.
b. It seems that no one was interested.
c. It’s been raining for a few hours.\(^2\)

Though seemingly underdiscussed, English can in fact have null subjects, at least in colloquial contexts. For example, in consideration of an informal style, it would be grammatical to restate the examples in (2), but without the expletive pronouns omitted (ø) , as demonstrated in (3):

(3) English expletive SPD

a. ø appears to be something in your coffee.
b. ø seems that no one was interested.
c. ø been raining for a few hours.

The nature of null subjects in English is not limited to expletive pronouns, however, as there are many instances where informal English also allows SPD of thematic subjects, as in (4):

(4) Null thematic subjects in English (data modified from Weir (2012)):

(i) ø\(_{1SG}\) won’t be in the office tomorrow.

\(^1\) Also occasionally referred to as “non-thematic” subjects or “non-referential” subjects.

\(^2\) Examples of pronouns such as those in (2c) are sometimes considered to be quasi-arguments rather than fully expletive pronouns, and as such there is controversy whether such pronouns correspond to a thematic role or not.
(ii) A: Am I invited to the party?
B: $\emptyset_{2\text{SG}}$ must be, I’m sure.

(iii) A: Why didn’t {he/she/they} come to the party?
B: $\emptyset_{3\text{SG}}$ didn’t want to, I guess.

Thus, it would appear that informal English can allow dropping of both non-thematic and thematic subject pronouns. But what empirical evidence beyond the suggestions in (3) and (4) exist to support this claim?

For this study, a corpus of colloquial English text from Twitter was collected. From the corpus, instances of sentences with null subjects were sought out and identified. In what follows, the purposes of this paper are three-fold. In section 2, the creation of the colloquial corpus and the process of null subject detection is detailed. In section 3, the resulting findings from the dataset are discussed, including the frequency of null subjects as well as any observed syntactic generalities. Finally, in section 4, the results from section 3 are considered with respect to prior literature on English null subjects in order to observe if the findings of the present study are consistent with past observations.

2. Collecting the Data

For this study, Twitter was selected as the source from which to build the colloquial English corpus. In looking for examples of ubiquitously available informal English text, where prescriptive pressures of grammatical “correctness” are minimalized, online chat rooms and social media websites are obvious ideal candidates. In this section, the details of corpus creation and null subject identification are discussed.

2.1 Corpus Creation Process

Twitter has a public API which makes downloading thousands of tweets in an instant rather straightforward. An API wrapper in the Python programming language called Twython was used to collect the tweets. The initial strategy of tweet collection considered was to collect a stream of completely random tweets, however, this approach proved to be problematic, resulting in too many non-sentences, non-English sentences, sentences with typos, etc. Therefore, to ensure that the tweets being collected would contain meaningful English sentences, a search filter was applied, such that only tweets containing a specified list of search terms were collected, shown here in (5):

(5) Specified search terms:

<table>
<thead>
<tr>
<th>appears</th>
<th>might</th>
<th>too many</th>
</tr>
</thead>
<tbody>
<tr>
<td>appears to be</td>
<td>must</td>
<td>too much</td>
</tr>
<tr>
<td>been</td>
<td>mustn’t</td>
<td>try</td>
</tr>
<tr>
<td>not been</td>
<td>need</td>
<td>trying</td>
</tr>
<tr>
<td>feel</td>
<td>should</td>
<td>want to</td>
</tr>
<tr>
<td>feels like</td>
<td>shoulda</td>
<td>wanna</td>
</tr>
<tr>
<td>has</td>
<td>should’ve</td>
<td>will</td>
</tr>
<tr>
<td>hasn’t</td>
<td>seem</td>
<td>would</td>
</tr>
<tr>
<td>have</td>
<td>seems like</td>
<td>woulda</td>
</tr>
<tr>
<td>haven’t</td>
<td>seems to be</td>
<td>would’ve</td>
</tr>
</tbody>
</table>
In the case of each token searched for, a limit of 500 tweets was set (this doesn’t mean that a full 500 tweets always surfaced for each term, but no more than 500 were collected for each). These particular search terms were selected, as it was hypothesized that they would be able to elicit null subjects. An attempt was made to include a variety of verbal predicates which take thematic subjects (want, need, try), ones which can take non-thematic subjects (appear, feel, seem), as well as auxiliaries, which can participate in clauses with both thematic and non-thematic subjects. Certain variant forms of verbs/auxiliaries were included ({going to, gonna}, {should’ve, shoulda}, ...) to see if null subjects were more prevalent in tandem with the more informal forms. The search terms too much/many were used to evoke expletive there (e.g., “There's too much drama”).

This methodology of corpus creation has its advantages and disadvantages. Unfortunately, it limits our ability to measure the organic prevalence of null subjects on Twitter, as we are purposefully seeking out search terms which are likely to evoke null subjects. It also generally limits the overall diversity of tweets we can collect, as they must all contain these specified search terms. However, on the plus side, it also drastically reduces the number of tweets which we might consider to be completely irrelevant to the task, and helps to bring forthright tweets featuring null subjects.

2.2 Null subject detection

In order to determine which tweets contained null subjects and which did not, a syntactic parser was used. SpaCy, an open-source software library for natural language processing, currently offers the fastest and most accurate dependency parser in the world – according to the findings of Honnibal and Johnson (2015) and Choi, Stent, and Tetreault (2015). An example of dependency relations, as produced by the parser, is shown in (6):

(6) Dependency parsing results using SpaCy³:

a. John kicked the ball.

b. I will see you tomorrow.

---
³ Visuals of the dependencies such as those in (6) were produced online at https://demos.explosion.ai/displacy/
In the examples in (6), we see that each token has been tagged with its part of speech, and each dependency relation has been labeled. The “root” of the sentence is always the primary verbal predicate. We can identify the root as the only token in the sentence which is not dependent on another token (and thus has no arrows pointing to it). In the case that a subject argument is present, the dependency relation NSUBJ is seen, meaning “nominal subject”. For example, in (6a), we see that the proper noun (PROPn) John is dependent on the root verb kicked, with the dependency arc between them labelled as NSUBJ. In the case that no subject argument is present, the dependency parser will simply lack any NSUBJ relation, as in (7):

(7) Dependency parse trees lacking NSUBJ

(a) Have a nice day.

(b) Sounds like a plan.

---

4 For a full glossary of dependency labels and their meanings as used by SpaCy, see https://github.com/clir/clearnlp-guidelines/blob/master/md/specifications/dependency_labels.md
As is canonically the case in English, this imperative sentence in (7a) lacks the overt presence of the second person subject argument you, and as such no subject dependency relation exists (NSUBJ is absent). It should be noted at this point that imperative constructions such as the example in (7a) are not the kind of “null subject” which we are trying to detect here. As such, any collected examples of imperatives will be considered false positives and removed from the dataset.

In the case that a single tweet consisted of multiple sentences to parse, the process of sentence tokenization was first applied to each tweet, using the Natural Language Toolkit (NLTK), another software library for natural language processing.

3. Results and Discussion

3.1 False Positives

All sentences detected by the syntactic parser as lacking a nominal subject were evaluated for false positives. False positives can occur for a variety of reasons. First, as stated above, we might collect English imperative sentences, since these lack overt subject arguments. We must also keep in mind that the parser itself is not, nor claims to be, 100% accurate, and as such false positives collected due to the parsers failure to identify the subject must be removed. This is made more difficult by the fact that the data source is Twitter, where rules of spelling, grammar, and punctuation are not always followed. For example, consider the sentence in (8):

(8) False positive: improper sentence tokenization

“thank you i really tried”

The NLTK sentence tokenizer relies on certain clues such as punctuation marks to correctly identify sentence boundaries in text form. However, examples of informal text such as the tweet in (8) sometimes lack said clues (it should be “Thank you. I really tried”, with a period between the two sentences). As such, the above example is treated by the dependency parser as a single phrase, and the subject argument I is erroneously treated as a direct object of the root verb, identified as thank.

Another common difficulty was that proper nouns were not always capitalized: the parser uses capitalization of unfamiliar tokens as the primary hint that the word is indeed a proper noun. This difficulty is highlighted in the example in (9):

(9) False positive: improper capitalization

“i am a student and you are not”
(9) False positive: lowercased proper nouns

a. why must twitch lie

b. Why must Twitch lie?

In this case, Twitch is the name of a company, and as such should be uppercased. The lack of uppercasing in the tweet shown in (9a) results in *twitch* being identified as the root verb. If we make the change of applying proper casing to the sentence, as in (9b), the parser can now correctly identify *lie* as the root verb, and creates an NSUBJ dependency relation with the proper noun *Twitch*.

The examples shown here in (8) and (9) are just a few of the difficulties created by attempting to perform dependency parsing on informal text. The other primary type of false positive, as mentioned above in section 2.2, is that of imperative constructions. These were most prolific with respect to the search terms from (5) which contain present tense verbal predicates that can take thematic subjects, such as *have* and *try*.

3.2 Frequency of Null Subjects

In total, after the manual removal of collected false positives, 7,509 parsed sentences remained, among which 748 had null subjects – 10.1% of the total corpus. The search terms identified in (5) are restated here in (10), along with how many results there were for each, how many sentences had null subjects, and the computed percentage of SPD for that search term.
## (10) Frequency of Null Subjects by Search Term

<table>
<thead>
<tr>
<th>Search term</th>
<th># of SPD</th>
<th>Total # of Sentences</th>
<th>% of SPD</th>
</tr>
</thead>
<tbody>
<tr>
<td>appears to be</td>
<td>6</td>
<td>140</td>
<td>4.29%</td>
</tr>
<tr>
<td>appears</td>
<td>5</td>
<td>183</td>
<td>2.73%</td>
</tr>
<tr>
<td>been</td>
<td>21</td>
<td>351</td>
<td>5.98%</td>
</tr>
<tr>
<td>feel</td>
<td>15</td>
<td>294</td>
<td>5.10%</td>
</tr>
<tr>
<td>feels like</td>
<td>21</td>
<td>201</td>
<td>10.45%</td>
</tr>
<tr>
<td>going to</td>
<td>15</td>
<td>219</td>
<td>6.85%</td>
</tr>
<tr>
<td>gonna</td>
<td>28</td>
<td>254</td>
<td>11.02%</td>
</tr>
<tr>
<td>has not</td>
<td>2</td>
<td>27</td>
<td>7.41%</td>
</tr>
<tr>
<td>has</td>
<td>10</td>
<td>296</td>
<td>3.38%</td>
</tr>
<tr>
<td>hasn’t</td>
<td>2</td>
<td>110</td>
<td>1.82%</td>
</tr>
<tr>
<td>have not</td>
<td>1</td>
<td>17</td>
<td>5.88%</td>
</tr>
<tr>
<td>have</td>
<td>13</td>
<td>329</td>
<td>3.95%</td>
</tr>
<tr>
<td>haven’t</td>
<td>20</td>
<td>161</td>
<td>12.42%</td>
</tr>
<tr>
<td>might</td>
<td>35</td>
<td>303</td>
<td>11.55%</td>
</tr>
<tr>
<td>must</td>
<td>21</td>
<td>239</td>
<td>8.79%</td>
</tr>
<tr>
<td>mustn’t</td>
<td>28</td>
<td>176</td>
<td>15.91%</td>
</tr>
<tr>
<td>need</td>
<td>31</td>
<td>281</td>
<td>11.03%</td>
</tr>
<tr>
<td>not been</td>
<td>1</td>
<td>16</td>
<td>6.25%</td>
</tr>
<tr>
<td>not seen</td>
<td>3</td>
<td>40</td>
<td>7.50%</td>
</tr>
<tr>
<td>seem</td>
<td>12</td>
<td>212</td>
<td>5.66%</td>
</tr>
<tr>
<td>seems like</td>
<td>78</td>
<td>225</td>
<td>34.67%</td>
</tr>
<tr>
<td>seems to be</td>
<td>10</td>
<td>128</td>
<td>7.81%</td>
</tr>
<tr>
<td>seen</td>
<td>23</td>
<td>254</td>
<td>9.06%</td>
</tr>
<tr>
<td>should</td>
<td>18</td>
<td>295</td>
<td>6.10%</td>
</tr>
<tr>
<td>shoulda</td>
<td>47</td>
<td>152</td>
<td>30.92%</td>
</tr>
<tr>
<td>should’ve</td>
<td>13</td>
<td>64</td>
<td>20.31%</td>
</tr>
<tr>
<td>too many</td>
<td>48</td>
<td>196</td>
<td>24.49%</td>
</tr>
<tr>
<td>too many</td>
<td>48</td>
<td>196</td>
<td>24.49%</td>
</tr>
<tr>
<td>too much</td>
<td>23</td>
<td>217</td>
<td>10.60%</td>
</tr>
<tr>
<td>tried</td>
<td>19</td>
<td>216</td>
<td>8.80%</td>
</tr>
<tr>
<td>try</td>
<td>7</td>
<td>200</td>
<td>3.50%</td>
</tr>
<tr>
<td>trying</td>
<td>57</td>
<td>270</td>
<td>21.11%</td>
</tr>
<tr>
<td>wanna</td>
<td>42</td>
<td>316</td>
<td>13.29%</td>
</tr>
<tr>
<td>want to</td>
<td>13</td>
<td>201</td>
<td>6.47%</td>
</tr>
<tr>
<td>wanted</td>
<td>19</td>
<td>249</td>
<td>7.63%</td>
</tr>
<tr>
<td>will</td>
<td>2</td>
<td>294</td>
<td>0.68%</td>
</tr>
<tr>
<td>would</td>
<td>13</td>
<td>289</td>
<td>4.50%</td>
</tr>
<tr>
<td>woulda</td>
<td>26</td>
<td>252</td>
<td>10.32%</td>
</tr>
<tr>
<td>would’ve</td>
<td>2</td>
<td>33</td>
<td>6.06%</td>
</tr>
</tbody>
</table>
A number of observations can be made regarding the data in (10). The search term which resulted in the highest frequency of SPD was *seems like* at a rate of 34.67%, the majority of cases of which involved the dropping of expletive *it*.\(^5\) Also, consider that all instances of informally contracted forms (e.g., wanna, shoulda, etc.) resulted in higher frequencies of SPD than their counterparts, as highlighted in (11):

(11) Prevalence of informally contracted terms

(a) going to → gonna: 6.85% → 11.02%
(b) should’ve → shoulda: 20.31% → 30.92%
(c) want to → wanna: 6.47% → 13.29%
(d) would’ve → woulda: 6.06% → 10.32%

A similar pattern occurs when we consider a comparison of verbs with their past tense forms (or auxiliaries which include past perfect aspect), even if slightly so:

(12) Prevalence of past tense/aspect

(a) should → should’ve: 6.10% → 20.31%
(b) try → tried: 3.5% → 8.8%
(c) want to → wanted: 6.47% → 7.63%
(d) would → would’ve: 4.5% → 6.06%

3.3 Syntactic observations

3.3.1 Representation of Pronoun types

Among these data, several patterns surfaced regarding subject pronoun drop. First of all, we can confirm that SPD in English is not limited to any particular pronouns, such as expletives or first person: it can apply to all pronouns. Consider the data in (13) as an example of each:

(13) SPD examples from the data

(a) \(\emptyset_{1SG}\) Need a good movie to watch
(b) \(\emptyset_{2SG}\) shoulda studied harder for quizzes 7 and 9
(c) \(\emptyset_{3SG}\) hasn't watched Game of Thrones yet because it's 'not really his thing.'
(d) \(\emptyset_{1PL}\) might as well have a huge comet crash on us while we’re at it
(e) \(\emptyset_{2PL}\) Want to really throw your PC's for a loop?
(f) \(\emptyset_{3PL}\) might as well. Soon they will be totally lawless.

\(^5\) Some examples of *seems like* involved a thematic subject, such as “Seems like a pretty well-adjusted strong young man” and “Seems like a sweet girl”. Since the subject argument is absent, the only way to verify what the missing argument is (it, she, he, …) is to consider the context of each tweet individually – a tedious process not performed in the current study.
(g) $\emptyset_{\text{EXPL}}$ feels like my back has been set on fire.
(h) $\emptyset_{\text{EXPL}}$ seems to be some major earth shaking going on!

However, that is not to say that the various pronouns shown in (13a-h) were evenly distributed with respect to their frequency, although the prevalence of each was not precisely measured in the current study. It appears, however, that the most frequent null thematic subject is the first person singular, which is not surprising since Twitter provides a context for discourse which is quite conducive to the expression of one’s own thoughts, feelings, and observations. Note that in all cases in (13), the dropped pronoun can (and must) be recovered by the context.

3.3.2 Omission of adjacent auxiliaries/copular verbs

One important observation in the data is that in many cases, when the subject dropped was adjacent to an aspectual auxiliary or a copular verb, the auxiliary or verb would also be dropped. Examples of each are shown in (14), where the omitted pieces have been crossed out:

(14) Auxiliary/Copular omission from the data

(a) Do you want to help mike title his next novel?
(b) I've been so happy lately it's crazy
(c) I'm watching SNL and there are too many live mistakes.
(d) I'm trying not to overthink it.
(e) It's gonna be a long week.
(f) There's so much great music out here it's almost overwhelming.
(g) I'm not sure I can continue now.
(h) I'm bored as hell tonight.

One logical observation regarding the elided pieces is that all of these auxiliaries/verbs can undergo cliticization with the adjacent subject argument, such that they treat the subject as their phonological host. Even if orthographic contractions are rare for examples like “do you” and “there are”, they are nevertheless commonly contracted in spoken informal English. The intuitive understanding of auxiliary deletion, then, would be to assume that the cliticized auxiliaries, in sacrificing their prosodic independence, are dropped when their host, the subject, is dropped.

The parallel between phonological cliticization and syntactic omission holds even across different dialects of English. For example, in American English, the pattern *I’ve not as a contraction of “I have not” is not attested (one would say “I haven’t”), but in British English, this is allowed (though not super common). This pattern shows up in the data as well. Consider the examples in (15), which take sentences from the Twitter data:

(15) Dialectal variation of SPD

(a) I’ve not been able to get on for the last 3 weeks.
(b) I haven’t been myself recently
The tweet in (15a) comes from an individual in England and thus the cliticized auxiliary ‘ve gets omitted along with the subject, while the tweet in (15b) comes from an individual in the United States, and the aspectual auxiliary have survives. Thus, it seems that the phonological contractions, as they are dialectally attested, inform which pieces can be omitted in SPD.

3.3.3 SPD and Syntactic blocking

One final observation, and perhaps the most important, is the distribution of null thematic subjects based on the search terms at hand. For example, when tweets containing try were searched, many results with no identified subject surfaced, however, all of them were imperative constructions, and were thus removed as false positives. Meanwhile, when try was put in the past tense as tried, then true positives of SPD surfaced. Consider the 3 tweets from the data in (16a-c), compared with their infelicitous transformation into present tense in (16d-f):

(16) Felicitous nature of Ø_{1SG} in present/past tense

(a) Ø_{1SG} tried to be positive
(b) Ø_{1SG} tried to bring them together
(c) Ø_{1SG} tried explaining that to them
(d) *Ø_{1SG} try to be positive
(e) *Ø_{1SG} try to bring them together
(f) *Ø_{1SG} try explaining that to them

The obvious explanation for why (16a-c) are attested but their present tense counterparts in (16d-f) are not is because the latter three examples seem like imperative constructions. In other words, a sentence like try to be positive matches the syntactic configuration of you omission in imperative commands, and as such the reading of (16d-f) with an assumed null first person singular subject argument (Ø_{1SG}) is not possible. That is not to say that present tense constructions with null first person subjects were never attested, but it was only attested with verbal predicates which are rare or disallowed with the imperative mood, such as need or want (e.g., “Ø_{1SG} need some help!”), although the most common null subject paired with want is actually you in interrogative constructions, as in “Do you want to try #glutenfree bread products?”.

This observation seems reminiscent of the concept of lexical blocking in morphology, where a prototypical formation of a word is “blocked” by a pre-existing form. For example, a common procedure for nominalizing adjectives ending with -ous is to add the suffix “-osity”, such as curious → curiosity or animous → animosity, but the pre-existence of the word glory prevents glorious → *gloriosity (Lieber 2016). In the case of English null subjects, the grammaticality of null subjects in present tense is “blocked” by the fact that present tense configurations which lack overt subjects are already canonically associated with imperative constructions, at least if the verb assigns an agent role to its subject argument, as with try.

A similar observation occurs with the fact that auxiliaries such as will or have never appear along the left-edge of a phrase in spoken informal English, such as *will go there later or *have already tried that, because in English, beginning a sentence with an auxiliary is a standard
configuration for forming interrogatives, thus blocking the occurrence of null subjects (it is much more standard in these cases, as the data would suggest, that these cliticisable verbs would be deleted along with the subject). However, it is noted that occasionally in written colloquial English, these configurations of null subjects with pronounced auxiliaries can occur, as will be seen in section 4.1.

4. Consideration of Prior Insights

One interesting consideration is whether the observations from the Twitter corpus regarding null subjects in English are consistent with prior literature on the subject. First, the most current developments on English null subjects is very briefly summarized, followed by a discussion of the Twitter data with respect to these insights.

4.1 Summarizing the literature

4.1.1 Phonological account of informal spoken English

Weir (2012) notes that particular written registers of English allow dropping of various types. These registers include “diaries, text messages, telegrams, newspaper headlines, recipes, and forms of electronic communication such as e-mails or messages on online forums or social networking sites” (Weir 2012). Among these registers of so-called “reduced” language, the data in (17) show some of those most studied examples of omission phenomena:

(17) Element omission in reduced registers (omitted pieces shown in ‹›):

(a) ‹I› Left the party exhausted (Haegeman 1990)
(b) ‹A› Man bites ‹a› dog (Weir 2009b)
(c) Let ‹it› cool for 30 minutes.

In these data, (17a) shows an example of subject drop in diaries (“diary drop”), (17b) shows an example of article drop in newspaper headlines, and (17c) shows object drop in recipes. Not only does the type of omission seem specific to certain kinds of registers, but Weir (2009a) also suggests that the phenomenon of null subjects in written versus spoken colloquial English merit separate analyses, since the following examples in (18) would be grammatical only in colloquial written forms of English (not spoken):

(18) Grammatical null subjects in written English (from Weir 2009a:120)

(a) ø have been feeling a bit ill lately.
(b) ø will go to the gym tomorrow.
(c) Tomorrow ø won’t be in the office.
(d) So ø shall now stop writing for a day

(18a-b) show examples with cliticisable auxiliaries that have not undergone contraction, and (18c-d) show examples where the null subject is preceded by preposed material. Weir also makes note of certain configurations of null subjects in English which are always forbidden, regardless of register. This includes (a) after preposed arguments, (b) in yes/no interrogatives, and (c) in wh-questions. Examples of each are shown in (19):
(19) Forbidden configurations of English SPD

(a) **preposed argument:** *More problems, φ don’t need.

(b) **yes/no interrogative:** *Should φ go to the gym on Saturday?

(c) **wh-questions:** *Who did φ see last night?

A variety of phonological, syntactic, and pragmatic theories have been proposed to explain when a null subject in English can be licensed. With respect to spoken informal English, the primary solution offered in Weir (2012) is a phonological one. Namely, subject omission is regarded as left-edge deletion of phonologically weak material. This phenomenon is not unique to null subjects, and can apply to many examples of left-edge material, as shown in (20):

(20) Left-edge omission (Weir 2012, his (15))

(a) Have you seen Tom?

(b) Have you seen Tom?

(c) The paper boy’s here.

(d) Has the professor arrived yet?

(e) It’s a good thing you decided to come along

In each example in (20), we can observe that the first non-omitted constituent is prosodically stronger than the weaker syllable(s) along the left edge of the sentence. In Weir (2012), the primary motivation of this phenomena is the optimal-theoretic constraint STRONGSTART, which requires prosodic constituents to start with a prosodically strong element. This analysis not only explains the phenomenon of null subjects as a unified understanding with other kinds of left edge omission, but can also why the forbidden configurations of (19) are not attested in spoken English: in each case, the placement of the null subject is not at the left edge, but is rather sentence-internal. As such, null subjects would not be predicted in these cases.

4.1.2 Syntactic Accounts of informal written English

Remember, however, that this analysis cannot account for the differences between spoken and written English seen in (18). For example, if we should delete not just subject arguments but any weak phonological left-edge material, then the cliticizable auxiliaries in (18a-b) should be omitted. Also, the grammaticality of (18c-d) with preposed adjuncts would not be predicted. In order to capture the difference between spoken and written English, a number of syntactic proposals have been made in order to predict the phenomena of written English as in (18). One such proposal, set forth by Haegeman (1990), is that null subjects in written English (particularly examined for the register of diaries) can be viewed as a form of topic drop. An example is shown in (21):

(21) Topic drop analysis

∅₁₅G saw the new Star Wars movie today.
[TOP\textsubscript{1} [t\textsubscript{i} saw the new Star Wars movie today.]]

Thus, the null subject argument is an A'-trace which is bound by a null topic operator, which itself is phonologically null but receives its referent from the discourse. Note that this is the same proposal present in Jaeggli and Safir (1989) regarding how languages with morphological uniformity like Chinese (lacking verbal inflection) identify their empty categories. This is of course much more acceptable than an analysis that regards English null subjects as pro, as English lacks the rich agreement morphology by which pro is said to be identified. The primary advantage of this analysis is that it explains the ungrammaticality of preposing arguments (as in (19a)) but supposed grammaticality of preposing adjuncts in diary English (as in (18c)). This is because the preposed argument and the null subject would both be fighting to occupy the same TOP\textsubscript{0} position, but not a preposed adjunct. This account also explains the ungrammaticality of English null subjects in wh-questions and yes/no-interrogatives, as overt topics are incompatible with these syntactic structures, as shown in (22):

(22) Ungrammaticality of topics in wh-/yes/no questions
   
   (a) *That course, did you take t\textsubscript{i}?  
   (b) *My pen, who did you let borrow t\textsubscript{i}?  

However, there are two obvious problems with this account. First of all, this account says nothing about expletive pronouns, as non-referential elements cannot be topicalized. Second, this account would seem to overpredict, as we can also topicalize object arguments, and yet null objects are not allowed in the same registers as null subjects. Consider (23):

(23) Ungrammaticality of null objects
   
   (a)  \textemptyset\textsubscript{1SG} didn’t see you  

   [TOP \textemptyset\textsubscript{1SG,i} [CP t\textsubscript{i} didn’t see you]]  

   (b)  *I didn’t see \textemptyset\textsubscript{2SG}  

   [TOP \textemptyset\textsubscript{2SG,i} [CP I didn’t see t\textsubscript{i}]]  

The other syntactic proposal, also set forth by Haegeman, is a phase based analysis of English SPD, which depends on an understanding of phase-based spellout, per Chomsky (1998). Put briefly, only the complements of phrases identified as “phases” (usually vP and CP) are spelled out. Haegeman (2007), following Rizzi and Shlonsky (2007), postulates the existence of a “Subject phrase” (SubjP) in between CP and TP, and subject arguments are said to move to the specifier position of SubjP. If CP is designated as the highest phase (as is normally the case), then this does not affect spell-out of sentences with overt subject arguments. However, perhaps SubjP is the highest phase in certain written registers of English (such as diaries). If so, then since only the complements of phases are spelled out, the subject argument in spec SubjP is not spelled out. An example is shown in (24):

(24) Spellout of SubjP
   
   [CP [SubjP t\textsubscript{i} [SubjP\textsubscript{0} [TP t\textsubscript{i} [VP need a new hobby]]]]]
In (24), if CP is selected as the highest phase, then the entire complement of $C^0$ is spelled out – “I need a new hobby”. However, if a register of English can select SubjP as the highest phase, then only the complement of SubjP$^0$ is spelled out, namely “need a new hobby”, resulting in a phonologically null subject argument. This proposal also accounts for the ungrammaticality of the forbidden configurations in (19), as any preposed material such as fronted wh-items and topicalized elements would not be spelled out if SubjP is the highest phase head. The case of the grammaticality of sentences in written informal English such as (19c) would still be problematic, as preposed adjuncts would not be spelled out, as they would appear above SubjP$^0$. Haegeman’s proposal is that such adjuncts do not move above the SubjP head, but rather merges in a TP-adjoining position. She supplies evidence of this structure as a possibility based on example data taken from British journalistic writing, and deduces that it is also an attested structure in the diary register of written informal English.

An example from Haegeman (2007) is shown here in (25) (her (45b-c)):

(25) Merge of modifiers in phase-based analysis

a. $[\text{SubjP The Prime minister [yesterday [TP the prime minister had made …]]]}$

b. $[\text{SubjP Jean, [la semaine prochaine, [TP Jean partira en Italie]}].$

4.2 Reconsidering the Twitter data

In light of the summary of theoretical insights in section 4.1, we might wonder how the Twitter data collected for this study fits into the pre-established picture of null subjects in English. It seems clear that the assertion about English having different registers which license different kinds of omission of material seems correct. Weir (2009a) notes that the reduced register of “telegramese” (reduced English as it appears in telegrams) is motivated by constraints on length. We must imagine that a similar constraint contributes to the omission of elements in tweets, as a single tweet is limited to 140 characters. This is an important consideration: we cannot extrapolate the examples of null subjects collected from Twitter to be equally attestable to all registers of English, including spoken informal English.

Since the work of Weir and Haegeman summarized in section 4.1 primarily concerns research into the diary register of written colloquial English, it is perhaps first worth comparing this register with that of the “Twitter register” of written English. We must reconsider the examples from (18), which Weir takes as evidence that written English allows pronoun drop in tandem with preposed adjuncts (18c-d) and that written English can allow pronoun drop before a cliticisable verb, per (18a-b). We can note an immediate problem with respect to an example sentence like (18c) Tomorrow $\emptyset$ will go to the gym: such a sentence would not have been detected as containing

---

6 It is worth noting that Weir (2009a) dismisses this pragmatic constraint on the discourse as a potential explanation or motivation for null subjects. This is because, in the case of tweets, there are many examples of tweets well under the 140-character limit which still disallow certain configurations, such as null subjects with a topicalized argument.
a null subject by the dependency parser, because in this case *tomorrow* would simply be identified as the subject argument, as in (26):

(26) Preposed adjuncts as false negatives

Tomorrow φ will go to the gym.

Thus, while no examples of such preposed adjuncts occurred in the Twitter data, they also likely would not have been detected. Despite this fact, by my grammatical intuitions as a native speaker, I find the examples with preposed adjuncts to be ungrammatical. (18d) is taken from the diary of British writer Virginia Woolf (1882 – 1941), and I would hypothesize that similar constructions would not make themselves present in the Twitter register, if we could detect them, but of course data is needed. Meanwhile, examples of cliticisable auxiliaries such as (18a-b) were present in the data, such as the tweets shown in (27):

(27) Cliticisable auxiliaries from the Twitter corpus

(a) will do my best tomorrow!

(b) will watch it as soon as I wrap the shoot

(c) am back home.

(d) am lost here.......help!!!

So while these cliticisable auxiliaries are present in the data, it should be noted that the “am” search term only contained a null subject 1.19% of the time, and the “will” tweets only contained a null subject 0.68% of the time: two of the lowest frequencies for null subjects among search terms in the analysis. It should also be noted that in the cases of (27), none of the individuals were from the United States; most were from India, meaning there’s still a lack of solid evidence that cliticisable auxiliaries are attestable in the dialect of written American English. Thus, the diary register of British English and the Twitter register don’t seem completely compatible, although more data is needed – it would be particularly helpful if a method could be devised to attempt to detect cases of preposed adjuncts such as (18c).

Every instance of SPD as it has occurred in the data collected in this study has been on the left edge of the sentence. As such, among the theoretical proposals discussed in section 4.1, the phonological account of Weir (2012) seems to be the most compatible with the Twitter register of informal written English. As we saw in the tweets from (15), the spoken dialect of the English

---

7 Though not included in the initial study, ‘am’ was added as a search term for the purposes of exploring the facts concerning cliticisable auxiliaries. The results for ‘am’ were 4 null subjects across 337 total parsed sentences (1.19% SPD frequency).
speaker seems to strongly inform what can be omitted in the written register, but no matter the dialect, it is always the weak leftmost syllable which is omitted. Meanwhile, for the cases of (18), while the unique grammatical acceptance of cliticisable auxiliaries in written registers must be accounted for, we can at least notice that it is consistent with the phonological explanation proposed here. That is, the null subject is always the weak leftmost syllable which is omitted.

Meanwhile, for the cases of (18), while the unique grammatical acceptance of cliticisable auxiliaries in written registers must be accounted for, we can at least notice that it is consistent with the phonological explanation proposed here. That is, the null subject is always the weak leftmost syllable which is omitted.

The preposed adjuncts of (18c-d) must also be considered, even if none were attested in the current dataset. Weir (2012) makes a tentative proposal that, if the syntax of (24) as proposed by Haegeman is correct (that subjects can move above preposed adjuncts), then these structures would violate the STRONGSTART constraint, as in (28) (his (55)):

(28) Phonological account of preposed adjuncts

(a) \[CP [\text{SubjP} \ I_1 [\text{SubjP}^0 [\text{tomorrow} [TP \ t_1 [TP \ [VP \ \text{go to the gym}]]]]]]\]
(b) \[\Phi [\sigma \ I_1 [\phi \ [\phi \ [\phi \ \text{tomorrow} [\phi \ \text{will go to the gym}]]]]]\]

In (28b), since the subject argument \(I\) is weaker on the prosodic hierarchy than \(\text{tomorrow}\) (\(\sigma < \phi\): syllables are weaker than phonological phrases), the STRONGSTART constraint is violated which motivates the deletion of the subject argument \(I\).

Therefore, the phonological analysis of left edge omission seems to be not only consistent with the observations of the Twitter data in (13), (14), (15), and (16) but also predicts the ungrammaticality of (19) (these syntactic configurations do not have the null subject at the left edge), and it may also be able to provide an account for the special cases of informal English specific to the written register(s), as in (18). Since the STRONGSTART constraint is not specific to subject arguments, we would also predict to find other parallels between phonologically null elements and omitted elements in written text, such as tweets which omit leftmost articles as in (20c), for example.

Thus, among the various theoretical explanations discussed in 4.1, Weir’s phonological account seems to be most compatible with the observations.\(^8\) This approach is not without faults, however. Weir himself admits that it can overpredict deletion. Consider (29) (also his (29)):

(29) Overprediction of deletion (Weir 2012)

(a) Don’t think he’ll go.

(b) \[\Phi [\sigma \ \text{don’t} [\phi \ \text{think he’ll go}]]\]

In this case, \(\text{don’t}\) is a weak start (\(\sigma < \phi\)), but it cannot drop, and this is due to its unrecoverability under deletion (i.e., deleting \(\text{don’t}\) creates an opposite meaning to (29a)). There is also no discussion on how left-edge material is identified in cases where it is recoverable. With respect to the null subjects identified in the Twitter data, we can most certainly say that identification occurs through the context of the discourse, as with languages like Chinese, since there is a lack of any agreement morphology in most cases. However, in the absence of any syntactic explanation for identification, as with the null topic analysis, the precise procedure explaining how left-edge deleted material is identified is left a mystery.

---

\(^8\) The compatibility of the phase-based analysis of Haenegan (2007) with the observations from the Twitter data will be left up to debate, as phase theory is still a highly contested subject in the field of syntax. See Boskovic (2016) for a discussion on how not just phase complements, but rather entire phases, are sent to spell-out.
One major observation from the Twitter data not captured by any theory presented here is the configuration restrictions on SPD noted in section 3.3.3. It seems that any ideal and complete explanation of the phenomena of English SPD needs to incorporate this observation of “syntactic blocking” into its analysis. Perhaps, following the example of Weir (2012), an optimal-theoretic style account can be crafted, such that we propose a BLOCKED constraint which would be violated any time a subject pronoun drop would result in a configuration already canonically associated with another configuration (as was observed with imperatives in (16)). A full sketch of such a proposal will not be attempted here, but it would be interesting future work.

A final consideration is the need to assign null subjects in English to some empty category for the purposes of computational identification of null subjects in dependency parsers. Consider the fact that the collection method for instances of English SPD in the current study depended on the fact that the dependency parser would fail to identify a subject argument, even if one is underlyingly present. Work has been done in computational linguistics to devise methods of recovering null material, such as Yang and Xue (2010), who describe a framework for detecting empty categories in Chinese, such as pro. The syntactic identification of empty categories can be crucially important for certain tasks such as accurate machine translation, as Chun and Gildea (2010) discuss. If we should ever hope to create syntactic parsers which can recover null subjects as they occur in informal English, it seems a purely phonological account would not suffice, as there would be a need to annotate such null subjects in the same way that other types of empty categories are annotated in various parsed text corpora, such as the Penn Treebank.

5. Conclusion

Though not a null subject language, it has been observed through collected data of colloquial English that subject pronouns can be omitted in certain cases. In this study, a corpus consisting of thousands of tweets from Twitter was collected to support this fact. A method of dependency parsing was used to identify tokenized sentences which lacked nominal subjects, and some of the challenges of this method were briefly discussed. Employing the methodology of data collection detailed in this paper, approximately 10% of sentences contained null subjects. A number of observations were made, including the fact that both thematic and non-thematic pronouns were able to be omitted. Another observation concerned the fact that the licensing of null subject pronouns seemed restricted in cases where the resulting syntactic configuration was strongly associated with another syntactic form, such as imperatives and interrogatives. The exact details of this “blocking” phenomenon and the general distribution of different pronouns in English SPD require further investigation.

Among the previous theoretical analyses of English SPD mentioned here, the phonological account of Weir (2012), which postulates English SPD as a type of a more general phenomenon of metrical left-edge omission, seemed to be quite compatible with the data collected from Twitter. However, a purely phonological account doesn’t seem to explain the observation of the phenomenon of syntactic restrictions mentioned above. In any case, it seems that the ideal theoretical approach to explaining the English SPD, at least in the informal registers of current investigation, would at some level require a unique interfacing of syntax and phonology.
References


Weir, A. (2009b) *Article drop in English headlines*. University College London
