

Imposters: An Online Survey of Grammaticality Judgments*

Chris Collins, Stephanie N. Guitard and Jim Wood

1 Introduction

In this paper we develop a simple online survey technique for collecting grammaticality judgments. Our primary target audience is the set of formal syntacticians who are interested in expanding their repertoire of tools for collecting syntactic data.

We illustrate this technique by applying it to a subset of the data on imposters in Collins and Postal (2008). For the purposes of this paper, we can define an imposter as a 3rd person DP that is used to refer to the speaker, as illustrated below (see Collins and Postal for much elaboration of this definition):

- (1) Daddy will buy you an ice cream cone.
("Daddy" refers to speaker)

We will investigate two main empirical claims made in Collins and Postal (2008). First, even though imposters are used to refer to the speaker, Collins and Postal claim that they cannot be the local antecedent (in the same clause) of a 1st person reflexive:

- (2) a. I am enjoying myself
b. *Daddy is enjoying myself
("Daddy" refers to speaker)

Second, Collins and Postal claim that there is a word order effect whereby an imposter can antecede a 1st person pronoun more easily when it follows that pronoun:

- (3) a. *Daddy will put on suntan lotion to keep myself from getting sunburned.
b. ? To keep myself from getting sunburned, Daddy will put on suntan lotion.

We will investigate these contrasts in this paper and show that they hold up (with some interesting twists) when data is collected from 15 NYU undergraduates using an online survey.

In linguistics, there are various acceptable methods for gathering data. In the tradition of generative grammar, judgments are elicited from one or a small number of speakers. The linguist notes the judgments as OK, ?, ??, ?*, *, **. In this paradigm, the linguist usually engages the speakers in a discussion of the individual sentences, asking them questions such as: How could the sentence be made better or more natural? Describe a situation where the sentence could be used? Is the sentence still acceptable if the order of certain words is changed? These questions and others allow the linguist to probe deeply into the data.

On the other end of the spectrum are the methods developed in psycholinguistics and psychophysics, such as magnitude estimation (Bard et al. 1996). Magnitude estimation involves

*The authors' names appear in alphabetical order. We would like to thank Frank Lopresti, Richard Kayne, Jon Brennan and two anonymous reviewers for helpful suggestions and comments on this paper.

comparison to a yardstick sentence. Other psycholinguistic methods involve use of a 7 point scale. Speakers are presented with a large number of examples, and asked to give judgments on them. The possibility for discussion with the linguist is limited or non-existent. The psycholinguistic methods have the advantage that they allow one to establish the judgments of a large number of speakers for a large number of sentences in a relatively short period of time. Also, these methods introduce various controls in the data (e.g., speakers are presented data in random order, so one can reduce the effects of order of presentation on grammaticality judgments).

However these methods are too cumbersome and foreign to be adopted on a large scale with working syntacticians. In some generative work, judgments of larger groups of speakers are presented in a slightly more systematic fashion. For example, the following is taken from Sigurðsson's (2008) study of control in Icelandic (see also Sigurðsson & Holmberg (2008)):

(4)					OK	?	*
a.	Hún bað Ólaf	[að PRO fara bara	einn	í veisluna].	9	4	2
	she asked Olaf.ACC	[to go just	alone.NOM	to party.the]			
b.	Hún bað Ólaf	[að PRO fara bara	einan	í veisluna].	12	2	1
	she asked Olaf.ACC	[to go just	alone.ACC	to party.the]			

Small scale data tabulations of this sort are popular with syntacticians. While this method of presentation alone gives a helpful overview of how speakers judge the sentences in question, one relevant empirical question might be whether we should take the differences in patterns of acceptability between (4a) and (4b) to be reliable ("statistically significant"). That is, while the presentation in (4) shows that a large number of speakers find both sentences acceptable, it does not tell us whether 9 vs. 12 is reliable or accidental. If we sent the same sentences to 15 other speakers, would there still be more in the 'ok' column of (4b) than (4a), or would it be reversed? To the extent that this question is important (depending on the theoretical question at hand, it might not be), it can only be answered using statistical tools. Understanding whether or not there is a reliable difference between (4a) and (4b) in how speakers rate them could be one of the first steps in understanding the syntactic principles underlying the formation of these sentences.

The goal of the current project is to find a middle ground between the psychophysics methodology employed in Bard et al. (1996) and the informal methods of generative grammar. We will present a technique for obtaining grammaticality judgments over the internet, and some simple statistical tools for evaluating these judgments. We will also discuss how the statistical results can be interpreted in a way that makes sense to a working syntactician.

Note we are not claiming that the only source or best source of evidence in generative grammar is grammaticality judgments. There are many different sources of data, which ultimately must yield converging results if we are to have confidence in what we are doing (cf. Clifton et al. 2006:56). Corpus studies are of immense value in scoping out the character of a construction and the range of variation among speakers. The use of Google (a particular kind of natural language corpus), especially in combination with native speaker intuitions, has become a tool in every syntactician's toolbox. In work on less studied languages, the researcher is wise to consult texts (e.g., oral texts that have been transcribed) to get the lay of the land, and establish the basic generalizations, before approaching more specific questions using grammaticality judgments. Another particularly rich source of data is translations of sentences from language X to language Y by speakers who are

completely bilingual in language X and Y (that is, speakers who are native speakers of both X and Y). Corpus data, translation data, and grammaticality judgments are all important sources of data. In this paper, we focus on grammaticality judgments.

It is also important to note that we are not advocating that all grammaticality judgments be obtained in the way described in this paper. Rather, there are different ways to obtain grammaticality judgments, each with their own strengths and weaknesses (see the conclusion for a summary of these).

2 Grammaticality Judgements

In this section we will discuss how grammaticality judgments were elicited in our survey, and how these results were processed statistically. Participants were asked to judge each of the sentences on the questionnaire using the rating system in (5).

- (5) Sounds completely natural and it is something I would say;
 Sounds kind of odd, but I wouldn't be surprised to hear someone else say it;
 Sounds completely wrong and no one would say this.

This system was adapted from Sobin (1987), who used the ratings in (6) to study judgments of *that*-trace effects.

- (6) Something you might say;
 Something you wouldn't be surprised to hear others say;
 Something that would sound odd to hear;

One important advantage to this type of system is that it corresponds closely to the ratings of OK, ? and * used by syntacticians. The use of a seven point scale, often found in psycholinguistic experiments (Arregui et al. 2006, Gordon and Hendrick 1997), does not naturally correspond to the rating system used by syntacticians, even if the full range of markings is used (OK, ?, ??, ?*, *, **).

Although some authors have advised against such rating systems (e.g. Cowart 1997:71), we feel that using a rating system such as that in (5) made our survey maximally understandable to the experimental subjects. Making it understandable was especially important, considering we were collecting data via an internet survey where we had to rely completely on written instructions (cf. Phillips 2008, to appear). There were no provisions in the instructions that allowed the subjects to ask for clarification. Further, Sobin (1987) has demonstrated that this type of scale is able to provide accurate, useful results.

Strictly speaking, (5) is what is known as an ordinal scale (Cowart 1997). That is, there is a meaningful order to the categories, such that (5a) is more acceptable than (5b), which itself is more acceptable than (5c). However, although order is meaningful, the intervals are not. We cannot say, for example, that the difference between (5a) and (5b) is the same as the difference between (5b) and (5c). In fact, in some sense, (5b) is closer to (5a) than to (5c), since, intuitively speaking, saying that one would not be surprised to hear someone else utter a sentence seems closer to saying that one would say it oneself than to saying that "no one would ever say this."

Nevertheless, we chose such a scale because it corresponds closely to traditional generative categories, it is immediately intuitive, and it requires little to no training to use. The problem,

then, becomes how to transform these data for statistical analysis. We translated these answers into numbers, as shown in (7).

- | | | | |
|-----|----|--|---|
| (7) | a. | Sounds completely natural and it is something I would say; | 3 |
| | b. | Sounds kind of odd, but I wouldn't be surprised to hear someone else say it; | 2 |
| | c. | Sounds completely wrong and no one would say this. | 1 |

These numbers can then be averaged into means which can be subjected to standard statistical measures such as the ANOVA and the *t*-test (which we will describe below). While this approach could be criticized as making inappropriate assumptions about the interval between 1 and 2 and that between 2 and 3, as discussed above (see discussion under (6)), Cowart (1997:120) notes that “where a particular contrast is numerically large compared with variability around the relevant means, any statistical problems deriving from a failure to achieve interval level measurement are not likely to be consequential.”¹

Note that giving the number 1 to “sounds completely wrong” in no way implies that ungrammatical sentences are being awarded points. It would not have made a difference if we had used 2, 1, 0 rather than 3, 2, 1. The reason is that what is important is the difference between the scores, not the scores themselves. The numbers chosen are essentially arbitrary. We could have used a scale of 9, 8, 7, or any other set of arbitrary numbers. As long as the difference between the middle score and the other two scores is equal, it would make no difference in the statistical tests.

Moreover, the numbers were so chosen to bias in the opposite direction from our intuitions. Since we feel that 2 might be closer to 3 than to 1, if we gave it a value of, say, 2.5, it might be argued that any significant differences between the marginal cases and the unacceptable cases were due to our decision to give extra weight to (7b). By making the distance equal (i.e. the difference between 1 and 2 is numerically the same as that between 2 and 3), we can be confident that significant differences between marginal and unacceptable cases are in fact real. As shown in the next three sections, the significant differences we report are robust enough that scaling issues are unlikely to be problematic for the substantive conclusions we draw from our data.

It should be pointed out that while we take averages in order to be able to run the statistics (e.g., to see if two sets of grammaticality judgments are significantly different within a group of native speakers of English), we do not place any theoretical importance on the averages themselves. For example, suppose that a set of grammaticality judgments (each of which is 1, 2 or 3) are found for a particular sentence, and the average of these judgments over the entire set of speakers is calculated. We do not think that this average corresponds to some real linguistic measure in the minds of the speakers. Most importantly, there are different types of variation across speakers in grammaticality judgments. One type of variation is a threshold for what counts as acceptable. It is widely known that certain speakers judge sentences more leniently than others, accepting lots of sentences that would be marginal to others. Another type of variation is that different speakers might have significant differences in their grammars (even though they are in some sense speaking the same “language”). So one speaker might judge a particular sentence as good, and another

¹Strictly speaking, the kind of data we use should be subjected to what are known as “non-parametric” statistical tests (Siegel and Castellan Jr. 1988). Unlike the tests we used, a non-parametric test does not make the assumption that the difference between 1 and 2 is the same as that between 2 and 3, nor that the results will generalize beyond the participants to the population from which they are drawn (i.e. other speakers of their dialect). Sprouse (2007:67) observes that “unfortunately, there is no generally accepted non-parametric version of factorial ANOVA, and in fact, standard ANOVA is often reported for ordinal data in the psychological literature.”

speaker might judge it as bad, because they have genuinely different linguistic systems. Yet another type of variation might be termed noise: errors in understanding the test materials, etc. It is clear that averaging grammaticality judgments, given these three very different types of variation, could never claim to correspond to some real linguistic measure within speakers (on rejecting the use of means/averages of group data we agree here with Den Dikken et. al. 2007).

These conclusions do not, however, mean that means are useless. If a contrast between two sentences does happen to be statistically reliable across a group of speakers (e.g., the number of people who rate sentence 1 as grammatical is significantly different than the number of people who rate sentence 2 as grammatical), we can be reasonably sure that the contrast at hand is a real one for many individual speakers. That being the case, it could prompt us to look for deeper linguistic explanations of the data. However, if there is no statistically reliable contrast across a group of speakers, the contrast may still be real in the grammars of individual speakers. This is why, as we emphasize in this paper, various methodologies should be used, among which averaging over the judgments of a group of speakers is only one.

As mentioned earlier, we needed to make sure that our participants were judging the sentences based on whether or not two terms within those sentences could be coreferential. To ensure this, we indicated coreference by surrounding the terms with ‘+’ signs, as shown in (8).

(8) +John+ wants to buy +himself+ an ice cream.

Though we would have preferred to use boldface to indicate this, our software platform (Google Docs, see below) made this impossible. Since ‘+’ does not signify anything in linguistics, we thought this was a good alternative to the other methods that we entertained (including boldface, coindexing, underlining, etc.), and the instructions along with the practice questions made it clear what the ‘+’ signs meant. Furthermore, nothing in the survey results indicated that subjects were confused by this notation for coreference.

3 Design of Questionnaire

Our questionnaire consisted of 48 sentences with an approximate time of completion of 20 minutes. The subjects were not timed, and they were encouraged in the instructions to take as much time as they wanted (see Appendix B). Furthermore, the sentences were presented together as a questionnaire (not one at a time), so subjects could go back and forth and compare sentences.

This feature of our experiment differs from traditional psycholinguistic studies where sentences (or words) are usually presented one at a time with no possibility to go back and check. One problem with the latter is that it creates more noise; subjects often, after experiments, report that they “think they messed up on a bunch of them.” Our methodology gives the subjects a chance, once they realize that they are uncomfortable with one of their choices, to go back and choose a different answer. Judging a sentence is a deeply introspective task, and not always an easy one. Offering the possibility of comparing sentences makes the task much easier (see the conclusion for more discussion). Cornips & Poletto (2005:947) cite Schütze (1996) for the observation that “speakers usually feel more confident about relative judgments than absolute ones.”²

²A reviewer remarked that “. . . the items that form the frame of comparison determine how you will evaluate the given sentence. This pertains both to the other experimental items and to the fillers!” We fully agree with this (see the

There were two main designs in our study, intended to test two different aspects of imposter syntax. The first design involved an imposter DP (e.g., “Daddy”, “this reporter”, “yours truly”) and a co-referential 1st or 3rd person pronoun or reflexive (1st vs. 3rd) contained in an adjunct clause (a purpose clause, a *because* clause, or temporal adverbial clause). The pronoun was in a subordinate clause which would either be preposed or postposed (pre vs. post). We will refer to this design as “Order of Adjunct Clause.” Sample sentences of this type are shown in (9). See Appendix C for the full sentence set.

- (9) a. Daddy will put on suntan lotion to keep myself from getting sunburned. **Post-1**
 b. To keep myself from getting sunburned, Daddy will put on suntan lotion. **Pre-1**
 c. Daddy will put on suntan lotion to keep himself from getting sunburned. **Post-3**
 d. To keep himself from getting sunburned, Daddy will put on suntan lotion. **Pre-3**

The other design involved a subject which was either an imposter or a 1st person pronoun (Impost vs. Pron) followed by a coreferential 1st or 3rd person pronoun or reflexive (1st vs. 3rd). We will refer to this design as “Imposter versus Pronoun as Antecedent.” Sample sentences of this type are shown in (10). Again, the full sentence set is presented in Appendix C.

- (10) a. Daddy is enjoying myself. **Impost-1**
 b. I am enjoying myself. **Pron-1**
 c. Daddy is enjoying himself. **Impost-3**
 d. I am enjoying himself. **Pron-3**

Within each of these two designs were five sets of sentences, making a total of 40 sentences. We presented all sentences to all subjects in the same order. In other words, the test sentences were not randomized across subjects. Once again, randomizing test sentences across subjects is an idea that does not fit naturally in the generative paradigm for judgment elicitation. Usually, when a syntactician works with an informant, they go over small related groups of sentences (called “paradigms”). Randomizing across subjects would break up these sub-groups of sentences.

However, we did “randomize” the order within each set of related sentences, such that the order for one set might be [a, b, c, d], while another might be something like [b, d, a, c], etc. This was to prevent participants from guessing or predicting which kinds of sentences would be in which position in the group (and therefore trying to guess the judgments, instead of actually judging the sentences). Given the lack of a time limit on the survey, it is unclear whether even this tiny bit of randomization is really needed.

The first eight questions of the spreadsheet were warm-up sentences that did not contain imposters, but were generally similar to our test sentences. We then alternated between the two main

conclusion for discussion, see also Cornips and Poletto (2005)). Our use of fillers was by no means intended to prevent subjects from comparing across or within paradigms— we used them to avoid the monotony of having subjects look at many nearly identical paradigms one after the other. The reviewer goes so far to cast doubt on “. . . the usefulness of comparative and non-randomized judgments that this experiment elicited,” citing Clifton et al. (2006). In the latter study, however, subjects were asked to pick the better of two sentences, which is not the same thing as rating two sentences next to each other. This shows up especially where two sentences might be roughly equal in acceptability; if half the subjects pick each, we have no idea if both sentences were bad or both were good. At any rate, if the results turn out to be statistically reliable with respect to the relevant contrasts, and they do not reveal something wildly unexpected from a research design perspective, then why should they not be useful? As Singler (2001:275) points out, “statistical programs are always tools, never analyses.”

sentence types such that the four sentences in (9) would be presented, followed by the four sentences in (10). In some sense, then, each quadruplet was functioning as a set of filler sentences for the next, and vice-versa, in two simultaneous experiments.

In psycholinguistic experiments, the main reason to use filler sentences is to prevent the subject from guessing what kind of contrast is being manipulated. As mentioned above we did not feel this was important in our study, since we did not want to prevent participants from comparing similar (grouped together) sentences and noticing what was being manipulated. However, judging the same contrast over and over, with no intervening sentences may give rise to fatigue. In order to give variety to the materials being judged, we alternated the two main sentence types (see Cowart (1997: 97) and Gordon and Hendrick (1997: 336) on filler sentences).

4 Procedure

We recruited 15 participants by e-mailing an advertisement to all subscribers of NYU's undergraduate linguistic department listserv. Our participants all self-reported as being English speaking, linguistics major undergraduates. The flyer that announced the experiment is given in Appendix A.

Once our participants e-mailed us that they were interested in the study, we would send a reply e-mail with all of the necessary information. The text of this e-mail contained the instructions to fill out the questionnaire and the link to access it (See Appendix B for full instructions.) An NYU UCH AIS (University Committee on Activities Involving Human Subjects) debriefing form, which explained the participants' rights, was attached to this e-mail. Participants completed the questionnaire over the internet.

We created the questionnaire using Google Documents Beta. This is a new, easily shareable, free software package.³ Once the form was created, the information provided by participants was automatically recorded in a Google spreadsheet when they clicked the submit button. This spreadsheet was only a place to collect our data, and we did not use it to calculate our results.

To keep track of the data, our participants were provided with an ID code. This allowed us to make sure that people only took the survey once and that only the participants who signed up were taking the survey. It also allowed us to maintain anonymity of the participants.

5 Results

5.1 Order of Adjunct Clause

In the following section, the results are presented. Recall that there were two main designs. First, there is the order of the adjunct clause where preposing (pre vs. post) and anaphor person (1st vs. 3rd) were manipulated. This was presented in (9). Another example of adjunct clause order is provided below in (11).

³Keller and Asudeh (2001) also conducted their experiments over the internet. Instead of using Google however, they used WebExp 2.1, a java applet designed specifically to conduct psychological experiments.

- (11) a. This reporter apologized after losing my cool. **Post-1**
 b. After losing my cool, this reporter apologized **Pre-1**
 c. This reporter apologized after losing his cool. **Post-3**
 d. After losing his cool, this reporter apologized. **Pre-3**

The sentences in (11) fit into a 2 x 2 cell design as follows. For the first design, there were five sentences in each cell judged by 15 subjects.

Table 1– Preposed x Person

	Post	Pre
First Person	a	b
Third Person	c	d

The means are reported in Table 2, averaged across subjects and items.

Table 2

	Post	Preposed
First Person	1.573	1.787
Third Person	2.76	2.507

These means were submitted to a two-way ANOVA.⁴ There was a significant main effect of Anaphor Person $F(1, 14) = 52.472$, $MSE = 1.3$, $p < .0001$, but no main effect of Preposing. There was, however, a significant interaction between the Person and Preposing conditions $F(1, 254) = 14.003$, $MSE = 74.07$, $p = .0002$. Scheffe post hoc tests revealed all pairwise comparisons to be significant at $p < .001$, except Pre-3 x Post-3, which was significant at $p = .004$, and Pre-1 x Post-1, which was significant at $p = .016$.

Before moving on, a brief discussion is in order of what the ANOVA tells us that the means alone do not. The first is just significance. Looking at the means in Table 2, we can see that ‘Third Person’, in each column, is higher than ‘First Person’ in that column. But what we cannot tell is how likely this is to be an accident. (Recall the discussion in the introduction.) The ANOVA tells that although there is variation in the judgments, a lot of it will be predictable when we take ‘Person’ into account. In this case, the ANOVA tells us that it is highly unlikely that the difference between ‘Third Person’ and ‘First Person’ is a random accident. The lower the p value (probability), the less likely it is that the contrast is an accident.⁵

The second thing this ANOVA tells us is that the numbers we see in the columns ‘Post’ and ‘Preposed’ very well could be a random accident. The fact that it did not come out significant

⁴For all ANOVAs reported, we used a by-subjects analysis of variance, collapsing across items. This means that for each subject x , we average x ’s response to each sentence type. For example, in the present case we would average subject A’s response to the five Post-1 sentences, to the five Post-3 sentences, etc. In this way, the variance that we analyze is the variance between subject A’s average response to Post-1 sentences, subject B’s response to Post-1 sentences, and so on. A by-items analysis, in contrast, averages all of the subjects responses to each particular sentence. We chose the former option because we were especially interested in maintaining any differences between subjects, comparing for each subject their contrasts across the different sentences types (where we were less concerned about variation across the tokens themselves).

⁵The standard benchmark is a p of less than .05. As we can see, p , i.e. the probability that our results are an accident, is less than .0001 for ‘Person’.

means that there is no statistical reason to rule out the possibility that these numbers are random. A closer inspection of the table provides a clue as to why. Notice that in the ‘First Person’ row, the number 1.573 in the ‘Post’ column goes up to 1.787 in the ‘Preposed’ column. But in the ‘Third Person’ row, the number 2.76 in the ‘Post’ column goes down to 2.507. In neither case does the number change very much, and each case, it goes in a different direction. When the ANOVA looks at the effect of ‘Preposing’, it ignores ‘Person’ completely, and as a result sees just minor unreliable variation in the effect of preposing.

This leads, however, to the third, and arguably most important thing the ANOVA tells us. It tells us whether there is a significant interaction. An interaction is precisely what we discovered looking at the effect of preposing: it causes acceptability to increase in one case (‘First Person’), and decrease in another (‘Third Person’). Thus, the effect of preposing “interacts” with the effect of ‘Person’. The ANOVA can tell us what the chances are that this interaction is an accident. In this case, the interaction has a *p* value of .0002, a highly significant interaction.

The means and ANOVA results are highly suggestive and relevant for linguistic theorizing. Take the “significant main effect of Anaphor Person”. We can explain this data by assuming that it is more difficult for a 1st person pronoun to have an imposter antecedent than for a 3rd person pronoun, and furthermore that this difference holds for most, if not all speakers (all I-Languages). Under such an assumption, we would in fact predict the significant effect of Anaphor person, without necessarily assuming that the means in the ANOVA Table 2 directly reflect some psychologically real quantity (because, as commented on earlier, these means are averaging over different types of variation, see the discussion following 7). Whether or not such an explanation holds for all speakers can only be revealed by an individual level analysis of the data, which we have not completed so far.

Similar remarks go for the other significant effects (discussed under Table 2). In particular, consider the “significant interaction between the Person and Preposing conditions”. This interaction is the result of two patterns. For third person, a preposed adjunct clause has a lower mean than a postposed one. This effect is clearly due to the well known difficulty some people have with backwards anaphora. For 1st person pronouns, a preposed adjunct clause has a higher mean than a postposed adjunct clause. This implies, surprisingly, that in general (with a few exceptions for particular speakers and particular examples) there is no backwards anaphora effect for 1st person pronouns (compare, for example, the data from [25] and [27]).

Finally, pairwise comparisons, often performed after an ANOVA, compare each cell to each other cell, one by one, to see which cells are significantly different from each other. In this case, all pairwise comparisons were significant. Therefore, the acceptability hierarchy can be summarized as in (12), where > means ‘more acceptable than’ (following the notation of Kluender (1998)).

$$(12) \quad \text{Post-3} > \text{Pre-3} > \text{Pre-1} > \text{Post-1}$$

These results, especially the interaction of the Person and Preposing conditions, thus support the claim made in Collins and Postal (2008) that linear order plays a role in the acceptability of imposters antecedent pronouns.

A more careful consideration of the data shows that purpose clause adjuncts behave differently from the others. Consider first a case where the adjunct is a (finite) temporal clause (some of the adjunct clauses in our materials are finite and some non-finite; we did not control for this factor in the testing materials):

- (13) a. Mommy needs to have a drink before I give you your bath. **Post-1**
 b. Before I give you your bath, Mommy needs to have a drink. **Pre-1**
 c. Mommy needs to have a drink before she gives you your bath. **Post-3**
 d. Before she gives you your bath, Mommy needs to have a drink. **Pre-3**

The mean judgments for these sentences (no others), averaged across 15 subjects, were as follows.

Table 3

	Post	Preposed
First Person	2	2.133
Third Person	2.667	2.533

These means were submitted to a two-way ANOVA. The main effect of Person only approached significance $F(1, 14) = 4.513$, $MSE = .945$, $p = .0519$.⁶ The main effect of Preposing was not significant $F(1, 14) = 0$, $MSE = .107$, $p > .05$. The interaction between the two only approached significance $F(1, 14) = 3.027$, $MSE = .088$, $p = .1$. Scheffe post hoc tests revealed Pre-3 x Pre-1 to be significant at $p = .002$. Neither Pre-3 x Post-3 nor Pre-1 x Post-1 were significant ($p > .05$). All other pairs were significant at $p < .001$.

The actual judgements for these sentences are given below (see Appendix C for all judgements):

Test Sentences	Bad	Marg	Good
[25] +Mommy+ needs to have a drink before +I+ give you your bath	3	9	3
[26] +Mommy+ needs to have a drink before +she+ gives you your bath	2	1	12
[27] Before +I+ give you your bath, +Mommy+ needs to have a drink	2	9	4
[28] Before +she+ gives you your bath, +Mommy+ needs to have a drink	2	3	10

Focusing first on the third person pronouns in [26] and [28], there seems to be a very weak effect regarding backwards anaphora. If the anaphoric pronoun precedes the antecedent, the sentence is degraded for some people (5 people find [28] marginal or bad versus 3 people for [26]). Focusing on [25] versus [27], the general pattern is that most people consider an imposter antecedent to a first person pronoun as marginal. It is not clear that preposing the temporal clause has any effect on the acceptability of the 1st person pronoun, although it might be significant that backwards anaphora effect found with third person pronouns does not surface with first person pronouns (as noted above).

Consider the variation in judgments for sentence [25] alone. Given the high number of marginal judgments (and even distribution of the Good and Bad judgments), we believe that it is highly likely that the variation with regard to this sentence has only to do with the thresholds different people have (not different grammars). In other words, [25] is basically marginal, because there is a grammatical principle barring an imposter as antecedent of a 1st person pronoun, and violation of this principle yields a marginal judgment. However, the way particular informants interpret this marginality differs. A particularly picky person might judge it as unacceptable, whereas

⁶Even though the means for 3rd person seem robustly higher than the means for 1st person, the effect for Person only approached significance. The reason for this is most likely the fact that each participant only judged one sentence in each cell, rather than five sentences as in Table 2 above. In general, the fewer the items, the harder it is to get a significant result.

a more lenient person could judge it as grammatical. This explanation could be tested by seeing if the Good-raters in [25] were the same as the Good-raters for other marginal examples, and the Bad-raters for [25] were the same as the Bad-raters for other marginal examples, an individual level micro-variationist study we have not completed so far.

Now consider a different example involving an (infinitival) purpose structure (once again, some of the adjunct clauses in our materials are finite and some non-finite, we did not control for this factor in the testing materials):

- (14) a. Daddy will put on suntan lotion to keep myself from getting sunburned. **Post-1**
- b. To keep myself from getting sunburned, Daddy will put on suntan lotion. **Pre-1**
- c. Daddy will put on suntan lotion to keep himself from getting sunburned. **Post-3**
- d. To keep himself from getting sunburned, Daddy will put on suntan lotion. **Pre-3**

The mean judgments for these sentences, averaged across 15 subjects, were as follows.

Table 4

	Post	Preposed
First Person	1.2	1.533
Third Person	2.8	2.467

These means were submitted to a two-way ANOVA. The main effect of Person was significant $F(1, 14) = 138.466, MSE = 24.067, p < .001$. The main effect of Preposing was not significant $F(1, 14) = 0, MSE = .179, p > .05$. The interaction between the two, however, was significant $F(1, 14) = 12.727, MSE = 1.667, p = .003$. Scheffe post hoc tests revealed Post-1 x Pre-1 to be significant at $p = .016$. Post-3 x Pre-3 only approached significance at $p = .055$. All other pairs were significant at $p < .001$.

The actual judgements are given below:

Test Sentences	Bad	Marg	Good
[9] To keep +myself+ from getting sunburned, +Daddy+ will put on suntan lotion	8	6	1
[10] To keep +himself+ from getting sunburned, +Daddy+ will put on suntan lotion	0	8	7
[11] +Daddy+ will put on suntan lotion to keep +myself+ from getting sunburned	12	3	0
[12] +Daddy+ will put on suntan lotion to keep +himself+ from getting sunburned	0	3	12

Overall, first person pronouns in purpose clauses controlled by an imposter are worse than first person pronouns in finite temporal clauses. Only one person judged either [9] or [11] as completely acceptable, whereas 7 people judged the temporal clause as completely acceptable, see [25] and [27]. Furthermore, it is clear that the use of a first person reflexive is worse overall than a third person reflexive ([9/11] vs. [10/12]).

As can be seen from the judgments and the statistics, in the case of purpose clauses, preposing has a much more significant impact on the acceptability of an imposter anteceding a first person pronoun than in the previous example involving a temporal adjunct. Seven people judged sentence [9] as marginal or good, but only three judged sentence [11] so. Furthermore, preposing made the sentences with a third person reflexive worse ([10] vs. [12]), a backwards anaphora effect which is opposite of the effect found in the first person reflexive case.⁷

⁷In the third person, backwards anaphora made the average worse, though not significantly so, while in the first person, backwards anaphora made the average better, and significantly so, as shown above.

The basic conclusion is that there is an effect with respect to adjunct order that is only significant for purpose clauses. How this is to be explained is unclear at the moment.

A methodological point on variation is in order. The above data is quite valuable in outlining what syntactic factors (adjunct order, pronoun person) have an effect on the grammaticality judgments. We should expect that the ultimate theory of imposters would allow us to account for the data (including which effects are significant, etc.). The drawback of the ANOVA data is that it does not allow one to track the linguistic behavior of individuals. Of course, we can unpack the data, and see if there are such correlations, but the ANOVA test themselves do not answer such individual level micro-parametric questions. The usefulness of the means and the ANOVA results is that they suggest which micro-parametric correlations to look for, as we will show in the next set of data.

5.2 Imposters versus Pronouns as Antecedents

The second design manipulated antecedent noun form (Imposter vs. Pronoun) and anaphor person (1st vs. 3rd). Examples are shown below along with a 2 x 2 table. (See also the examples in 15 and the Appendix C.)

- (15) a. Mommy is not proud of myself. **Impost-1**
- b. I am not proud of myself. **Pron-1**
- c. Mommy is not proud of herself. **Impost-3**
- d. I am not proud of herself. **Pron-3**

Table 5– Noun Form x Person

	Imposter	Pronoun
First Person	a	b
Third Person	c	d

In this design, there were four sentences in each cell judged by 15 subjects. There was a fifth set of sentences within this design which were analyzed separately for reasons described below. The mean judgments are reported in Table 6.

Table 6

	Imposter	Pronoun
First Person	1.294	3
Third Person	2.817	1

These means were submitted to a two-way ANOVA. There was a significant main effect of Anaphor Person $F(1, 14) = 5.69, MSE = .625, p = .032$ but not for Noun Form $F(1, 14) = 1.515, MSE = .092, p > .05$. There was also, however, a significant Person x Noun Form interaction $F(1, 192) = 1488.8, MSE = .125, p < .0001$. Scheffe post hoc tests revealed all pairwise comparisons to be significant at $p < .0001$, with the exception of Impost-3 x Pron-1, which was significant at $p = .002$.

All differences being significant, the acceptability hierarchy can be summarized as in (16), where > means ‘more acceptable than’ (as above, following the notation of Kluender (1998)).

(16) Pron-1 > Impost-3 > Impost-1 > Pron-3

Consider the data from a particular example:

	Test Sentences	Bad	Marg	Good
[13]	+Daddy+ is enjoying +himself+	1	1	13
[15]	+I+ am enjoying +himself+	13	1	0
[14]	+Daddy+ is enjoying +myself+	15	0	0
[16]	+I+ am enjoying +myself+	0	0	15

The data is nearly categorical. Each of the four sentences is either nearly unanimously good or bad. Without interviewing the subjects, it is impossible to know why “I am enjoying himself” was rated as marginal instead of bad by one speaker. Similarly for “Daddy is enjoying himself”, which was rated bad by one person and marginal by one person. One possibility is that these sentences represent “noise”. For some reason the subject misinterpreted the sentences or instructions, and responded inappropriately (that is, in a way not consistent with their own grammar). An alternative is that the bad rating of “Daddy is enjoying himself” might have correlated with similar bad ratings for the same subject for other sentences involving an imposter anteceding a third person pronoun. We have yet to explore these possibilities.

Collins and Postal (2008) claim that there is a dialect where singular imposters can antecede first person reflexive pronouns in the same clause, so that sentences like (15a) would be acceptable. The results in [14] do not confirm this claim. But a look at all the individual sentences where an imposter antecedes a first person reflexive is revealing:

	Test Sentences	Bad	Marg	Good
[38]	+Mommy+ is not proud of +myself+	11	3	1
[46]	+Daddy+ just bought +myself+ a new car	12	2	1
[32]	+This reporter+ sees +myself+ as managing editor in the future	9	5	1

While everybody rejected [14], for each of the other sentences where an imposter antecedes a first person reflexive, there was one person who accepted them. Also, there were a number of people who rated those other sentences as marginal. Breaking the data down by informant, we have the following table:

Sentence	125	130	139	143	124	132	137
38 +Mommy+ is not proud of +myself+	Good	Bad	Bad	Marg	Marg	Bad	Marg
46 +Daddy+ just bought +myself+ a new car	Good	Marg	Marg	Bad	Bad	Bad	Bad
32 +This reporter+ sees +myself+ as managing editor in the future	Good	Marg	Marg	Bad	Marg	Marg	Marg

This table shows that one speaker [125] was responsible for all the good judgements on [32], [38] and [46]. This suggests that there is a dialect where imposters can locally antecede first person reflexives, supporting the claim made in Collins and Postal (2008). Note that this speaker does not seem to be responding idiosyncratically. (S)he rejected test sentences such as +I+ *try to keep +himself+ from getting sunburned* and +I+ *am not proud of +herself+*, and accepted test sentences such as +I+ *am not proud of +myself+* and +I+ *try to keep +myself+ from getting sunburned*.

Consider now the marginal judgments. As can be seen from the table, there is significant overlap in the people giving the marginal judgments. Six speakers account for all instances of first person reflexives anteceded by imposters being judged as marginal (a total of 10 marginal judgments). Four of these speakers account for all but two instances. This suggests that the ‘marginal’ judgments of these sentences are not just noise. If they were noise, we would not expect the same speakers to judge the same kinds of sentences in the same way.⁸ It is possible that that these six, or a subset of them, also has the dialect postulated by Collins and Postal (2008).

The analysis of individual speaker data reveals an important aspect of ANOVA data, such as that in Table 6. The mean for the sentences where an imposter antecedes a first person reflexive is 1.294 (slightly better than bad which is 1). But from this data it is impossible to tell that it is the same speaker (informant 125) who rated three out of four of the cases of an imposter anteceding a 1st person reflexive as good. Only an individual level, micro-parametric study can reveal such clear idiolects (which are probably governed by slightly different grammatical principles). The one outstanding question is why speaker 125 did not accept [14], and whether speaker 125’s data on [32], [38], and [46] correlates with other judgments.

There was one set of four sentences within the Imposter x Person design where the imposter was in a Coordination Phrase (CoP). This set was not included in the above analysis because pre-theoretically, it is thought to have a different level of acceptability than normal DP imposters (see Collins & Postal 2008, Cattaneo 2007). The quadruplet is given in (17) below.

- (17) a. We are enjoying ourselves on the beach.
 b. Daddy and Mommy are enjoying themselves on the beach.
 c. Daddy and Mommy are enjoying ourselves on the beach.
 d. We are enjoying themselves on the beach.

The mean judgments for this quadruplet (averaged across 15 subjects) are reported here in Table 7.

Table 7

	Imposter	Pronoun
First Person	2.2	3
Third Person	2.8	1

The biggest change between Table 6 and Table 7 is the imposter anteceding a first person reflexive, as hypothesized. The mean of this cell in Table 6 is 1.29, whereas in Table 7 it is 2.2. The means in Table 7 were submitted to a two-way ANOVA. There were significant main effects of both Person $F(1, 14) = 30.265$, $MSE = .243$, $p < 0.0001$ and Noun Form $F(1, 14) = 26.25$, $MSE = .143$, $p = .0002$. There was also a significant interaction between the two $F(1, 14) = 104.38$, $MSE = .243$, $p < .0001$. Scheffe post hoc tests revealed all pairs to be significant except Pron-1 x Impos-3 ($p > .05$). Impos-3 x Impos-1 was significant at $p < .005$. Pron-1 x Impos-1 was significant at $p < .0006$. All other pairs were significant at $p < .0001$.

As can be seen from this data there is a clear difference between a singular imposter and a coordinated imposter in so far as pronominal antecedence goes. The actual data for a coordination case is given below:

⁸That is, if it were noise, there would probably be more than 6 out of 15 people contributing to it.

Test Sentences	Bad	Marg	Good
[21] +Daddy and Mommy+ are enjoying +themselves+ on the beach	1	1	13
[22] +We+ are enjoying +ourselves+ on the beach	0	0	15
[23] +We+ are enjoying +themselves+ on the beach	15	0	0
[24] +Daddy and Mommy+ are enjoying +ourselves+ on the beach	2	8	5

Whereas every single person rejected “Daddy is enjoying myself”, only two people rejected outright “Daddy and Mommy are enjoying ourselves on the beach”. A direct comparison of the singular and coordinated example demonstrates the difference statistically:

- (18) a. Daddy is enjoying myself. 1.07
 b. Daddy and Mommy are enjoying ourselves on the beach. 2.2

A matched *t*-test revealed this difference to be significant $t(13) = 5.491, p < .0001$.

Bringing the discussion down to the level of individual I-languages, all five speakers who rated (18b) as Good rated (18a) as Bad. The two who rated (18b) as bad also rated (18a) as bad. There was only one speaker who rated both as marginal. For the other 7 marginal speakers, (18a) was still bad. One thing is clear from this summary of individual results: there is no idiolect (I-language) that allows (18a) but rejects (18b). This fact (about I-languages) underlies the significance of the statistical test in (18), and it needs to be explained by a theory of imposters.

6 Conclusion

In this paper, we have presented a simple online survey tool for collecting grammaticality judgments and we have applied it to some data on imposters found in Collins and Postal (2008).

We have demonstrated that there is a difference between purpose clauses and other types of adjuncts as far as the effect of linear order on the acceptability of imposters anteceding first person pronouns. We have also suggested that there is no backward anaphora effect in the case of first person pronouns in adjunct clauses. Since the use of first person pronouns with imposter antecedents is often marginal, such survey results are useful in illuminating what is going on. We have also given some preliminary evidence for a dialect that accepts singular imposters as the antecedents of first person singular pronouns. Lastly, we have demonstrated that there is a difference between singular imposters and coordinate imposters as antecedents for first person reflexives.

The survey data in this paper has to be interpreted alongside of traditional grammaticality judgments and corpus studies (both kinds of evidence used in Collins and Postal 2008). Compared to traditional methods of eliciting grammaticality judgments, the online survey technique prevents any communication at all between informant and linguist, and hence makes it harder for the linguist to communicate his or her expectations about the data (and hence makes it harder for the linguist to influence the results on a subconscious level). Furthermore, the online survey technique allows one to survey a relatively large number of people in a short period of time, which in turn allows one to get a feel for the range of judgments across a large number of speakers. One drawback of the online survey is that it creates more possibilities for misunderstanding on the part of the informants (if they do not understand the instructions). Also, an online survey does not allow the linguist to ask follow up questions.

The strength of our study compared to more traditional psycholinguistic approaches is that we have stuck fairly close to traditional generative methodologies, a summary of which is listed below:

- (19) a. Use of a three point scale of grammaticality judgements corresponding to “good”, “bad” and “marginal”.
- b. No time limit on test items.
- c. Items are not presented one at a time, but rather all together, so they can be compared to one another.
- d. No randomization across subjects (difficult to implement and not clearly needed).
- e. Related sentences are clumped together in “paradigms”.

(19c) (together with (19b), (19d), and (19e)) is particularly important. It is widely known in syntax that it is easier to fix one’s judgments for a given sentence X relative to a closely related sentence Y. This is why in teaching syntax, one often asks the students to give relative judgments, not absolute ones (e.g., a sentence with a “weak subjacency violation” is better than a sentence with a “strong ECP violation”, even though to untrained ears both are really terrible). We are aware of no systematic studies of this effect (although see Clifton, Fanselow and Frazier 2006 and Cornips and Poletto 2005 for some discussion), although every good syntactician is familiar with it. As far as naïve informants go, similar considerations apply. If one asks an informant whether a sentence X is acceptable, there are many things about X that the informant could focus on in making a decision (whether the sentence sounds acceptable, the lexical choices, the interpretation, register, its length, etc.). If X is presented along with Y (as part of a paradigm, or closely related set of sentences), it is likely that this helps the informant focus on those aspects of X relevant for the particular study.

Ultimately, all the results presented (means and ANOVA results) above must be explained in terms of the I-languages of the particular speakers in the study (see for example the discussion following (18)). If each subject were treated as a particular I-language and each sentence as a particular grammatical property, it would be possible to correlate properties (grammaticality judgements on particular sentences) across I-languages and see what the interrelations are. Database technology seems particularly well suited for this task.

7 Appendices

The advertisement for the study is given in 7.1, the instructions given to participants are presented in section 7.2, and the full set of test sentences and the judgements on them are presented in section 7.3.

7.1 Appendix A: Advertisement

We want to know what YOU think!

We are interested in your judgments concerning the status of certain sentences. We will ask you to judge around 40 sentences, and tell us whether or not you think they sound natural. This is a quick survey which should only take 20 minutes or so and which you can take online from anywhere you can access the Internet.

If you are interested please contact Stephanie at **Stephanie.n.guitard@gmail.com**

Please remember all participation is voluntary and you may withdraw from the study at anytime.

7.2 Appendix B: Instructions

Thank you for taking our survey! Your ID Code is XXXXX. Attached is the project summary statement required by UCAIHS. Please read it at your leisure.

Please carefully read the sentences in the questionnaire and judge them based on the following criteria:

Sounds completely natural and it is something I would say.

Sounds kind of odd, but I wouldn't be surprised to hear someone else say it.

Sounds completely wrong and no one would say this.

In the following sentences, when two expressions are surrounded by plus signs (+), it means that they refer to the same individual. For example, consider the sentence below:

+John+ said +he+ is leaving.

“John” and “he” are both surrounded by plus signs, because they refer to the same individual.

In this experiment, we are looking at phrases like “Daddy”, “Mommy” and “this reporter” that refer to the speaker. For example, in the following sentences, the phrases “Daddy”, “Mommy” and “this reporter” are used instead of the pronoun “I”:

Daddy is going to get you an ice cream cone.

Mommy is enjoying herself.

This reporter believes the election will be close.

Remember there is no right or wrong answer. Though some of the sentences you read may not be acceptable in formal writing, please evaluate them based only on whether or not they sound natural to you.

The questionnaire contains 48 questions and should take about 20 minutes, but you can take as much time as you want. Please feel free to return to any question you have already completed if you have changed your mind on a judgment.

To begin click on the link below. Remember to enter the ID Code that you received at the beginning of the e-mail when it is requested.

<http://spreadsheets.google.com/viewform?key=p8nXByAuPbkYkRYhSAtB1HA&email=true>

7.3 Appendix C: Test Sentences and Responses

Warm up sentences		Good	Marg	Bad
[1]	+I+ try to keep +myself+ from getting sunburned	17	2	0
[2]	+My brother+ tries to keep +himself+ from getting sunburned	17	2	0
[3]	+I+ try to keep +himself+ from getting sunburned	0	0	19
[4]	+My brother+ tries to keep +myself+ from getting sunburned	0	2	17
[5]	To keep +myself+ entertained, +I+ watch television on Sunday	18	0	1
[6]	+I+ watch television on Sunday to keep +myself+ entertained	17	1	1
[7]	To keep +himself+ entertained, +my brother+ watches television on Sunday	17	0	2
[8]	+My brother+ watches television on Sunday to keep +himself+ entertained	18	0	1
Test Sentences		Bad	Marg	Good
[9]	To keep +myself+ from getting sunburned, +Daddy+ will put on suntan lotion	8	6	1
[10]	To keep +himself+ from getting sunburned, +Daddy+ will put on suntan lotion	0	8	7
[11]	+Daddy+ will put on suntan lotion to keep +myself+ from getting sunburned	12	3	0
[12]	+Daddy+ will put on suntan lotion to keep +himself+ from getting sunburned	0	3	12
[13]	+Daddy+ is enjoying +himself+	1	1	13
[15]	+I+ am enjoying +himself+	13	1	0
[14]	+Daddy+ is enjoying +myself+	15	0	0
[16]	+I+ am enjoying +myself+	0	0	15
[17]	After losing +my+ cool, +this reporter+ apologized	7	8	0
[18]	+This reporter+ apologized after losing +his+ cool	0	3	12
[19]	+This reporter+ apologized after losing +my+ cool	8	6	1
[20]	After losing +his+ cool, +this reporter+ apologized	2	4	9
[21]	+Daddy and Mommy+ are enjoying +themselves+ on the beach	1	1	13
[22]	+We+ are enjoying +ourselves+ on the beach	0	0	15
[23]	+We+ are enjoying +themselves+ on the beach	15	0	0
[24]	+Daddy and Mommy+ are enjoying +ourselves+ on the beach	2	8	5
[25]	+Mommy+ needs to have a drink before +I+ give you your bath	3	9	3
[26]	+Mommy+ needs to have a drink before +she+ gives you your bath	2	1	12
[27]	Before +I+ give you your bath, +Mommy+ needs to have a drink	2	9	4
[28]	Before +she+ gives you your bath, +Mommy+ needs to have a drink	2	3	10
[29]	+I+ see +himself+ as managing editor in the future	15	0	0
[30]	+I+ see +myself+ as managing editor in the future	0	0	14
[31]	+This reporter+ sees +himself+ as managing editor in the future	1	2	12
[32]	+This reporter+ sees +myself+ as managing editor in the future	9	5	1

Test Sentences		Bad	Marg	Good
[33]	+Daddy+ has to go back to the office because +I+ forgot something	5	7	3
[34]	Because +he+ forgot something, +Daddy+ has to go back to the office	1	7	7
[35]	Because +I+ forgot something, +Daddy+ has to go back to the office	7	4	4
[36]	+Daddy+ has to go back to the office because +he+ forgot something	1	0	14
[37]	+I+ am not proud of +herself+	15	0	0
[38]	+Mommy+ is not proud of +myself+	11	3	1
[39]	+Mommy+ is not proud of +herself+	1	0	14
[40]	+I+ am not proud of +myself+	0	0	15
[41]	To cover +myself+ in case of an investigation, +this reporter+ is going to keep +himself+ out of the newspapers	4	8	3
[42]	To cover +himself+ in case of investigation, +this reporter+ is going to keep +himself+ out of the newspapers	1	3	11
[43]	+This reporter+ is going to keep +himself+ out of the newspapers, to cover +myself+ in case of investigation	11	4	0
[44]	+This reporter+ is going to keep +himself+ out of the newspapers, to cover +himself+ in case of investigation	1	3	11
[45]	+Daddy+ just bought +himself+ a new car	1	0	14
[46]	+Daddy+ just bought +myself+ a new car	12	2	1
[47]	+I+ just bought +himself+ a new car	15	0	0
[48]	+I+ just bought +myself+ a new car	0	0	15

References

- Arregui, Ana, Charles Jr. Clifton, Lyn Frazier and Keir Moulton. 2006. Processing elided verb phrases with flawed antecedents: The recycling hypothesis. *Journal of Memory and Language* 55:232-246.
- Bard, Ellen Gurman, Dan Robertson and Antonella Sorace. 1996. Magnitude Estimation of Linguistic Acceptability. *Language* 72 (1):32-68.
- Cattaneo, Andrea. 2007. Imposters and Subject Clitics: Four different types of Imposters in Bellinzonese. Manuscript, New York University.
- Clifton, Charles Jr., Gisbert Fanselow and Lyn Frazier. 2006. Amnestying Superiority Violations: Processing Multiple Questions. *Linguistic Inquiry* 37 (1):51-68.
- Collins, Chris and Paul Postal. 2008. Imposters. Manuscript, New York University. Available on lingBuzz/000640.
- Cornips, Leonie and Cecilia Poletto. 2005. On standardising syntactic elicitation techniques (part 1). *Lingua* 115 (7):939-957.
- Cowart, Wayne. 1997. *Experimental Syntax: Applying Objective Methods to Sentence Judgments*. Thousand Oaks, CA: Sage Publications.
- Den Dikken, Marcel, Judy B. Bernstein, Christina Tortora and Raffaella Zanuttini. 2007. Data and grammar: Means and individuals. *Theoretical Linguistics* 33 (3):335-352.
- Gordon, Peter C. and Randall Hendrick. 1997. Intuitive knowledge of linguistic co-reference. *Cognition* 62:325-370.
- Kluender, Robert. 1998. On the distinction between strong and weak islands: A processing perspective. In *Syntax and Semantics 29: The Limits of Syntax*, edited by Peter W. Culicover and L. McNally, 241-279. New York: Academic Press.
- Phillips, Colin. 2008, to appear. Should we impeach armchair linguists? *Japanese/Korean Linguistics* 17.
- Schütze, Carson T. 1996. *The Empirical Base of Linguistics. Grammaticality Judgments and Linguistic Methodology*. Chicago: The University of Chicago Press.
- Siegel, Sidney and N. John Castellan Jr. 1988. *Nonparametric Statistics for the Behavioral Sciences*. New York: McGraw-Hill.
- Sigurðsson, Halldór Ármann. 2008. The Case of PRO. *Natural Language and Linguistic Theory* 26 (2):403-450.
- Sigurðsson, Halldór Ármann and Anders Holmberg. 2008. Icelandic Dative Intervention: Person and Number are separate probes. In *Agreement Restrictions*, edited by Roberta D' Alessandro, Susann Fischer and Gunnar Hrafn Hrafnbjargarson, 251-280. Berlin: Mouton de Gruyter.
- Singler, John Victor. 2001. Why you can't do a VARBRUL study of quotatives and what such a study can show us. *University of Pennsylvania Working Papers in Linguistics* 7 (1):257-278.
- Sobin, Nicholas. 1987. The variable status of Comp-trace phenomena. *Natural Language and Linguistic Theory* 5:33-60.
- Sprouse, Jon. 2007. *A program for experimental syntax*. Doctoral Dissertation, University of Maryland.

Chris Collins
New York University
Department of Linguistics
726 Broadway, 7th Floor
New York, NY 10003
cc116@nyu.edu

Stephanie Guitard
502 E. 89th Street Apt 4D
New York, NY 10128

Jim Wood
New York University
Department of Linguistics
726 Broadway, 7th Floor
New York, NY 10003
jim.wood@nyu.edu