Draft of July 13, 2011

On the Grammatical Status of PP-Pied-Piping in English: Results from Sentence-Rating Experiments*

Seth Cable
UMass Amherst

Jesse A. Harris
UMass Amherst

Keywords: massive pied-piping, locality, acceptability ratings, crowdsourcing

1 Introduction

We report a series of sentence-rating experiments that were designed to test recent claims about the grammaticality of sentences formed with certain movement operations related to question formation in matrix and subordinate clauses in English. Specifically, we tested the claim that so-called 'pied piping' of prepositional phrases in English is largely restricted to main clauses, and is rather marginal in subordinate clauses.

In English question formation, so-called *wh*-elements like *who*, *what*, and *how* are commonly analyzed as having moved to a position other than that which they were subcategorized for. For example, the question word who in (1b) is commonly viewed as having originated in object position t_1, the position occupied by John in (1a). Its surface appearance at the left periphery of the clause is attributed to a movement operation that displaces the *wh*-word into its observed position.

(1) **Simple Wh-Movement in English**
   a. She left John
   b. Who$_1$ did she leave t_1 ?

In addition to the basic case in (1), in which only a *wh*-element is subject to movement, the moved *wh*-element can co-occur with another phrasal element, making for a more complex moved constituent. Such cases, illustrated below, are commonly referred to as 'pied-piping' (Ross 1967), a term that metaphorically suggests the *wh*-word has lead the other phrasal material along as it moves to the left periphery.

* For Experiment 1, we wish to thank Margaret Grant for her assistance in creating items, as well as, collecting data, for which we also thank Aynat Rubinstein. For Experiment 2, we extend our thanks to Lyn Frazier for discussing our materials and design with us, to Kyle Johnson and Peggy Speas for allowing us to recruit participants from their courses, to John J. McCarthy for allocating the funds to pay subjects on Amazon's Mechanical Turk, and to Alex Drummond for developing the Ibex farm presentation software and his help implementing our experiment. We are grateful to Brian Dillon for comments on an earlier draft. All errors are naturally our own.

(2) **Pied-Piping in English**

a. She left that guy.

b. [Which guy]$_1$ did she leave t_1 ?

Interestingly, when a *wh*-element is complement to a preposition, there appears to be some optionality in whether the preposition is 'pied-piped' when the *wh*-element undergoes movement. That is, it would appear that English freely permits both the structures in (3). In sentence (3a), the *wh*-word has not pied-piped the higher PP. Such structures are commonly referred to as 'preposition stranding' or 'P-stranding'. In sentence (3b), the *wh*-word has pied-piped the PP, a structure referred to as 'PP-pied-piping'.

(3) **The Optionality of Pied-Piping PP in English**

a. Who$_1$ did she leave [$_{PP}$ with t_1]?

b. [$_{PP}$ With whom]$_1$ did she leave t_1 ?

Although both the structures in (3) are commonly reported in the syntactic literature as 'acceptable', the grammatical status of PP-pied-piping structures is somewhat unclear, especially when compared to their preposition stranding counterparts. While speakers recognize structures like (3b) as English, such structures are not particularly colloquial. It is sometimes said that such structures are limited to particular registers, but often what is meant by 'register' in this context is unclear. After all, structures like (3b) are no longer a regular occurrence in formal written English either.

Perhaps not coincidentally, English PP-pied-piping poses analytic difficulties for current theories of pied-piping. Both the theories of Heck (2008) and Cable (2010) are at first glance challenged by the seemingly free alternation between (3a) and (3b). As noted by both Heck (2008) and Cable (2010), these two theories predict that (3a) should serve to 'block' (3b).[1]

In response to this challenge, both Heck (2008) and Cable (2010) argue that PP-pied-piping in English possesses a special status, that of 'massive pied-piping'. In brief, the term 'massive pied-piping' refers to the ability for the locality conditions on pied-piping to be marginally violated in non-embedded (matrix) clauses. For example, English typically does not allow a *wh*-word to pied-pipe a noun phrase (NP) it is contained within (4). This condition can be understood as a locality condition on the pied-piping relation itself: a lexical category like NP cannot intervene between the *wh*-word and the edge of the pied-piped phrase (Heck 2008; Cable 2010). Structures that violate this constraint contain an NP in the pied-piped phrase, leading to the impression that the size of the pied-piping is larger than usual, or otherwise

1 Heck (2008) and Cable (2010) differ in the exact way in which this alleged 'blocking' occurs. The details here would take us too far afield in the present work. The interested reader is referred to Heck (2008: 117–127) and Cable (2010: 166–173).

'massive'. For example, in the ill-formed (4b.i), the NP *a picture* in the pied-piped phrase *a picture of whom* intervenes between the left phrase boundary and the *wh*-word *whom*, whereas movement of *who* to a higher clausal position in (4b.ii) appears to obviate the restriction.

(4) **Massive Pied-Piping of NPs in English**

a. *In Matrix Clause*

i. ?? [A picture of whom] did Bill take at the party?

ii. Who did Bill take [a picture of] at the party?

b. *In Embedded Clause*

i. * I wonder [a picture of whom] Bill took at the party.

ii. I wonder who Bill took [a picture of] at the party.

As noted above, a defining feature of massive pied-piping is that it is restricted to non-embedded clauses. Put differently, the locality constraints governing pied-piping structures are (in some cases) weaker in main clauses than they are in subordinate clauses. Although Heck (2008) and Cable (2010) explain these contrasts differently, both share the view that embedded clauses reveal the true, general conditions on pied-piping in a given language. The seemingly broader range of pied-piping structures in main clauses is merely due to independent factors that can mitigate the severity of the grammatical violations in question.[2]

With this in mind, both Heck (2008) and Cable (2010) note the relative ill-formedness of English PP-pied-piping in subordinate clauses.[3]

(5) **Pied-Piping of PPs in English Limited to Main Clauses**

a. (?) [With whom] did she leave?

b. * I wonder [with whom] she left.

This contrast suggests that PP-pied-piping structures in English should be categorized as massive pied-piping structures. Note that a similar contrast does not seem to exist for 'determiner pied-piping' structures (DP-pied-piping) like those in (6), structures that all theories of pied-piping find unproblematic.

2 According to Heck (2008: 297–315), these 'independent factors' are the possibility of feature movement from the *wh*-word in main clauses. According to Cable (2010: 190–198), they are the ability for Q-particles in main clauses not to bear the [*wh*]-features that otherwise force Agreement with *wh*-words.

3 As noted by (Heck 2008: 123), this contrast has been observed by many other scholars. Heck cites at least eight other works noting this contrast, the first of which appears to be Allen (1980).

(6) **Pied-Piping by Wh-Determiners in English is Not Massive Pied-Piping**

a. [Which book] did she borrow?

b. I wonder [which book] she borrowed.

The difference between the patterns in (5) and (6) reinforces the notion that pied-piping of PPs has a distinct grammatical status in English, one that separates it from pied-piping by determiners and possessors.[4]

Given these facts, Heck (2008) and Cable (2010) indeed conclude that PP-pied-piping in English is a case of 'massive pied-piping'. Thus, as revealed in embedded clauses (5b), such structures do indeed violate the locality conditions on English pied-piping; this violation is simply mitigated in main clauses like (5a).[5] Consequently, the analytic problem raised by the seemingly free variation between (3a) and (3b) disappears, as only (3a) is entirely licit.

While this line of analysis rescues Heck (2008) and Cable (2010) from the challenge posed by the data in (3), one is right to criticize its empirical basis. The key notion – that pied-piping of PP is 'massive pied-piping' in English – is supported by only a few crucial example sentences.[6] Further complicating things is the fact that that the key comparison between (5) and (6) is not truly a minimal one. Since one is comparing PP-pied-piping to determiner pied-piping, various extraneous changes in lexical choice and syntactic structure separate the sentences to be compared. Consequently, one might rightly worry whether the putative contrast between (5) and (6) is simply an artifact of the particular examples chosen.

For these reasons, we sought to more rigorously test the proposal that pied-piping of PPs in English is 'massive pied-piping'. We developed a series of sentence-rating experiments to test the following two predictions, which follow from accounts like Heck (2008) and Cable (2010), albeit for different reasons.

4 Citing prior literature, Heck (2008) notes that pied-piping of PPs in subordinate clauses greatly improves in sentences where P-stranding is not permitted. As illustrated below, such cases of 'forced pied-piping' do seem to be licit for PPs in subordinate clauses.

(i) I wonder [[in what sense] he was responsible]

(ii) * I wonder [[what sense] he was responsible in]

In this paper, we put aside such cases of 'forced pied-piping', restricting our attention to cases where PP-pied-piping appears to optionally vary with P-stranding. That is, all our items in Experiment 1 were instances of optional (non-forced) pied-piping; see Appendix A for details.

5 Of course, it must be noted that speakers find PP-pied-piping in main clauses is much better than other cases of massive pied-piping, such as that in (4a.i). Neither Heck (2008) nor Cable (2010) account for this. However, one explanation may be that speakers find PP-pied-piping more 'familiar', through their exposure to literature from earlier stages of the English language where PP-pied-piping was more prevalent. Also see the discussion in Heck (2008: 123–124).

6 Again, however, the key empirical observation has been informally corroborated by several scholars, as cited by Heck (2008: 123).

(7) **Experimental Predictions to be Tested**

a. Pied-piping of PPs is less acceptable in English subordinate clauses than it is in main clauses.

b. The contrast between pied-piping in main clauses vs. subordinate clauses is seen only for PP pied-piping. A comparable contrast is not found for pied-piping of DPs by *wh*-determiners.

The following sections describe the structure of these experiments in more detail. In Section 2, we review our first experiment, which supported prediction (7a). Section 3 presents the second experiment, which was found to support prediction (7b).

2 Experiment 1

Prediction (7a) states that pied-piping of PPs is less acceptable in English subordinate clauses than in main clauses. That is, if pied-piping of PPs in English is truly 'massive pied-piping', then we should find the judgment paradigm in (8) can be observed for PP-pied-piping. Specifically, we should find a greater contrast in the acceptability ratings between the embedded pairs displayed in (8b.i-ii) than their matrix counterparts displayed in (8a.i-ii). In other words, the effect of PP-pied-piping should be greater for embedded clauses than for matrix clauses.

(8) **Judgment Pattern, If Pied-Piping of PPs is Massive Pied-Piping**

a. In Matrix Clause

i. ?? [With whom] did he dance?

ii. Who did he dance with?

b. In Embedded Clause

i. * I wonder [with whom] he danced.

ii. I wonder who he danced with.

Our first experiment sought to test this prediction. The details of its design and results are reported below.

2.1 Design and materials

Experiment 1 was developed as a 2x2 factorial design, in which we crossed Clause type (Matrix or Subordinate clause) with Dependency type (PP-pied-piping or Preposition stranding). In (9) below, we illustrate the paradigm with an example quartet from the experiment:

(9) a. With whom did he dance? (Matrix: PP-pied-piping)

b. I wonder with whom he danced. (Subordinate:PP-pied-piping)

c. Who did he dance with? (Matrix: Prep stranding)
d. I wonder who he danced with. (Subordinate: Prep stranding)

All of the Pied-piping cases contained the prescriptive relative pronoun whom, a potential confound that was eliminated in Experiment 2. The materials consisted of 8 quartets following the pattern in (9); see Appendix A for a complete list of items.

2.2 Participants and procedures

Forty undergraduates from the University of Massachusetts Amherst participated in the study for course credit. The experiment took less than 1/2 hour on average to complete. Participants were invited to take a break between trials at any point during the experiment, and were given a planned break halfway through the experiment to prevent fatigue. The items were presented in randomized counterbalanced order with items from 5 sub-experiments, unrelated to the manipulation presented here, and 10 non-experimental filler sentences, for a total of 102 experimental trials per session. Items were shown in a Latin Square design so that each participant saw one and only one condition from each experimental quartet.

Participants were instructed to read each sentence carefully and then rate its acceptability on a 5-point Likert scale. Each sentence item and the acceptability scale, as in (10), were simultaneously presented on a computer screen with the program Linger,[7] which recorded responses and reaction times. Prior to the experiment, participants were instructed verbally and then again in writing with a guided practice. Participants were invited to ask the experimenter for assistance at any point during the experiment.

(10) a. I wonder with whom he danced?
b. How acceptable was that sentence?
i. Terrible
ii. Pretty bad
iii. So so
iv. Pretty good
v. Perfect

2.3 Results

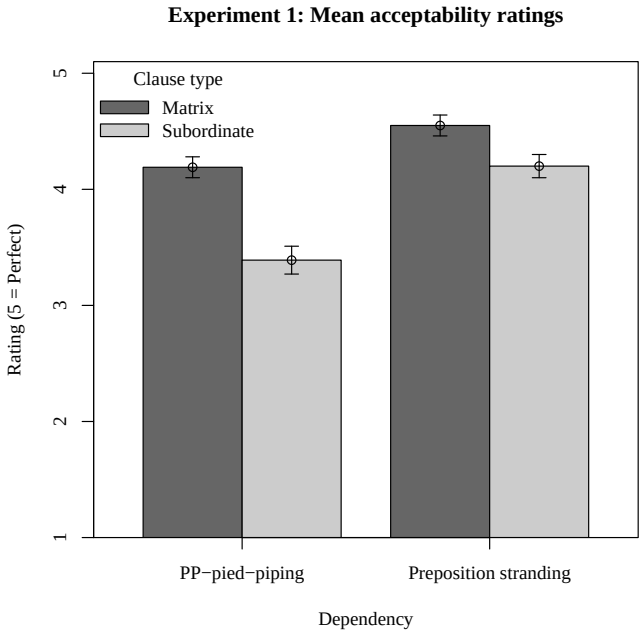
Prior to analysis, the data were cleaned of outliers on the basis of reaction time. Scores with reaction times registering over 3 standard deviations were removed from

7 Linger is available at http://tedlab.mit.edu/∼dr/Linger/

the data. Removing these outliers resulted in less than 3% data loss for any one condition, and less than 2% data loss for the entire data set. All data analysis was computed using the R statistical language suite R Development Core Team (2008).

2.3.1 Acceptability judgments

We first present the mean acceptability ratings and standard errors in Figure 1.



	PP-pied-piping	Preposition stranding	Mean
Matrix	4.19 (0.09)	4.55 (0.09)	4.37 (0.07)
Subordinate	3.39 (0.12)	4.20 (0.10)	3.80 (0.08)
Mean	3.79 (0.08)	4.38 (0.07)	

Figure 1: Mean acceptability judgment ratings (standard errors in parentheses) by condition for Experiment 1.

We used a linear mixed-effects regression model to analyze the data, treating subjects and items as simple random effects (for discussion of this method in the context of psycholinguistic designs see, e.g., Baayen, 2008; Baayen et al., 2008). In our model, we crossed the planned factors of Clause type and Dependency, as illustrated in (11) below; any statistical results we report were obtained from this

model, using the lmer package (Bates & Maechler 2009) to compute significance values. All significant results are reported.

(11) Rating $\sim$ Clause$\times$Dependency + (1|Subject) + (1|Item)

First, as expected, there was a significant main effect of Clause type: items with subordinate clauses were rated significantly lower (M = 3.80, SE = 0.08) than their matrix counterparts (M = 4.37, SE = 0.07), t = -6.37, p $<$ 0.001. This cost for subordinate clauses is not surprising; as subordinate clauses are more complex, they should be more difficult to process, and hence be judged as less acceptable. Second, we observed a main effect of Dependency: items with Preposition stranding (M = 4.38, SE = 0.07) were rated as significantly more acceptable than items with PP-pied-piping (M = 3.79, SE = 0.08), t = 2.75, p $<$ 0.01. Again, this effect is not particularly surprising, as we expect that the PP-pied-piping construction is probably far less frequent than constructions with preposition stranding in modern English.

Crucially, however, we observed a significant interaction between Clause type and Dependency, such that Clause type affected the acceptability rating of PP-pied-piping structures more than it affected those of Preposition stranding structures. In particular, the extent to which a subordinate clause lowered acceptability ratings as compared to matrix counterparts was significantly greater for PP-pied-piping structures (d = 0.8) than it was for Preposition stranding structures (d = 0.35), t = 2.65, p $<$ 0.01. The direction of the interaction is particularly clear in Figure 1. To determine whether this effect was spurious, we removed the interaction term from the model and fit the data to this simpler model. Again, we observed an effect of Clause type and an effect of Dependency. However, comparing the two models revealed that removing the interaction term did not result in a better fitting model. Indeed, the first model, which included the crucial interaction term, was a significantly better fit, $\chi^2 = 6.99$, $p < 0.001$, even though it was a more complex model.

Finally, although we did collect reaction time data from the experiment, we do not present it in detail here. We choose to omit the data, because it is of only marginal interest. Recall that our reaction times include the time spent reading the sentence, as well as the time spent rating the sentence, as subjects viewed both the target sentence and the rating scale simultaneously (10). As the subordinate condition was longer across the board, we would expect that conditions with subordinate clauses would elicit longer reading times – regardless of dependency type – for relatively uninteresting reasons. That is indeed the only effect regarding reaction times that we found in this experiment.

2.4 Discussion

As noted above, the results of Experiment 1 support the predictions in (7a); there was a significant difference between the effect of PP-pied-piping in main clauses and subordinate clauses. Furthermore, as the difference between matrix and subordinate structures was significantly greater for PP-pied-piping than for Preposition stranding, the effect cannot be reduced to a contrast between matrix and subordinate clauses. Thus, Experiment 1 confirms that the key judgment pattern in (8) does hold for PP-pied-piping, which supports the claim that such pied-piping should be categorized as 'massive pied-piping'.

However, since only PP-pied-piping was examined in Experiment 1, there remains the question of whether the effects observed are specific to that construction. After all, if all pied-piping structures exhibited these effects, then the conclusion that PP-pied-piping deserves special categorization is defeated. For this reason, we sought in Experiment 2 to examine the possible existence of these effects in 'non-massive' pied-piping.

3 Experiment 2

Prediction (7b) states that the effect of clause type observed for PP-pied-piping should not be found for instances of 'non-massive' pied-piping, such as pied-piping by *wh*-determiners. That is, if the effect observed in Experiment 1 is truly due to PP-pied-piping having a special status within pied-piping constructions, then the judgment paradigm in (8) should not be observed for those other constructions.

Unfortunately, due to independent features of English grammar, one cannot make a perfect comparison between the pied-piping of PPs and 'non-massive' pied-piping. The problem is that all cases of 'non-massive' pied-piping in English are also not optional. For example, while both (3a) and (3b) are possible in English, only (12b) is grammatically licit.

(12) **Non-Massive Pied-Piping (in English) is Not Optional**

a. * Which did she write book.

b. [Which book] did she write?

Thus, one cannot perfectly adapt Experiment 1 for, e.g., pied-piping by *wh*-determiners. Instead, the best one can do is to compare the interaction between clause-type and pied-piping. That is, if the effect of clause type is greater for PP-pied-piping than for DP-pied-piping, then this lends support to prediction (7b), and thus the special status of PP-pied-piping in English, in support of analyses like Heck (2008) and Cable (2010). Of course, such an effect would also be open to alternative analyses, which

will be discussed in Section 3.4 below. Before we come to that, however, we lay out the experimental design and results.

3.1 Design and materials

Experiment 2 was also developed as a 2x2 factorial design, crossing Clause type (Matrix or Subordinate) with Pied-piping type (DP-pied-piping or PP-pied-piping), as illustrated in (13). There were 12 such quartets in all.

(13) a. Which book did she write? (Matrix: DP-pied-piping)
b. It's obvious which book she wrote. (Subord: DP-pied-piping)
c. About what did she write? (Matrix: PP-pied-piping)
d. It's obvious about what she wrote. (Subord: PP-pied-piping)

The complete materials may be found in Appendix B.

3.2 Participants and procedures

We conducted our experiment twice. In the first instance, 91 members from the general population participated in Experiment 2; they were recruited on Amazon's Mechanical Turk, an internet-based service where workers anonymously answer short questions or perform small tasks in exchange for payment. Workers who elected to participate in our experiment were given $5 as compensation. Only workers who had performed at least 50 previous questions and received a 98% approval rating or above were permitted to participate in the experiment. Prior to taking the questionnaire, participants were asked to fill out a brief survey on their linguistic history. Self reported non-native English speakers were excluded from the analysis. In the second instance, we sought to replicate our results with undergraduates from the University of Massachusetts Amherst.

Since we did not have direct contact with our participants, we enacted several stringent controls to reduce the possibility of deceptive or disengaged participants. First, we recorded an MD5 hash of the IP address of all our participants to reduce the possibility that participants were repeating the experiment. If identical IP addresses were recorded, all of that participant's data were removed. Second, immediately after reading the instructions, participants were asked to answer three comprehension questions from Harris & Potts (to appear), designed to be particularly difficult for non-native speakers. Incorrectly answering any one of these questions was grounds for removal. Third, we included 4 undeniably ungrammatical or non-sensical sentences with failures of number, case agreement, or sensicality (see Appendix B for these items). Any participant who rated any one of these items greater than a 3 was excluded from analysis. Twenty-five subjects were eliminated

from the data on this basis. Fourth, we looked at the responses over the course of the experiment. Participants who showed little to no variation in their responses – either in their ratings or the speed by which they responded – were excluded on the presumption that they were not genuinely performing the task. In order to create a balanced dataset, we excluded 18 participants from the experiment.[8] Finally, we removed observations with reaction times that were greater than 3 standard deviations from the mean, which, as before, resulted in less than 3% data loss for any condition, and less than 2% data loss across the entire experiment.

After a short, guided practice, participants rated the items for acceptability on a 7-point Likert scale as in (14) below, in which 1 was designated as "Completely unacceptable" and 7 as "Completely acceptable."

(14) Which book did she write?
How acceptable is this sentence? (Please use entire scale.)
(Completely unacceptable) 1 2 3 4 5 6 7 (Completely acceptable)

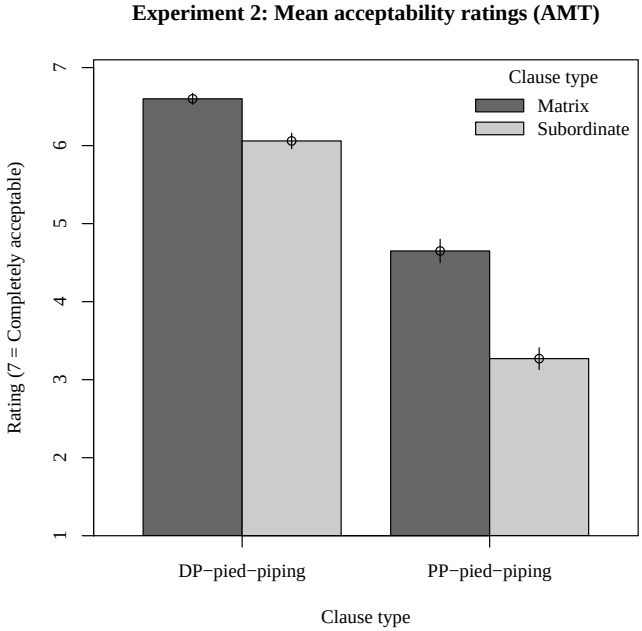
Items were presented in a Latin Square design, counterbalanced and randomly interspersed with 48 items from another experiment on Pied-piping, 61 genuine filler items, and 4 catch sentences, as discussed above, for a total of 125 items per session. Ratings and reaction times were presented and recorded over the Internet using Ibex farm. Participants from the first group were told that they had 1 hour and 20 minutes to complete the experiment, which took approximately 1/2 hour on average. Participants were invited to complete the questionnaire at their own pace, and were invited to take breaks as needed; after rating each item, participants saw a screen with the following instructions "When you're ready, press a key to see the next sentence."

3.3 Results

We first present the results from the group who were recruited from Amazon's Mechanical Turk (the AMT group), and then, as the results are virtually identical, briefly report the results obtained from student participants at the University of Massachusetts Amherst (the UMASS group). Means and standard errors from the AMT group are provided in Figure 2 below.

The procedure for data analysis was identical to that described in Experiment 1, modulo the different planned factors. We again observed a cost for subordinate clauses: pied-piping in matrix clauses (M = 5.63, SE = 0.10) was rated significantly better than pied piping in subordinate clauses (M = 4.67, SE = 0.12), t = -3.54, $p < 0.001$. In addition, there was a main effect of pied-piping type, such that participants rated pied-piping of PPs (M = 3.96, SE = 0.11) significantly worse than pied-piping

8 Removing these subjects did not affect the general patterns observed in the data.



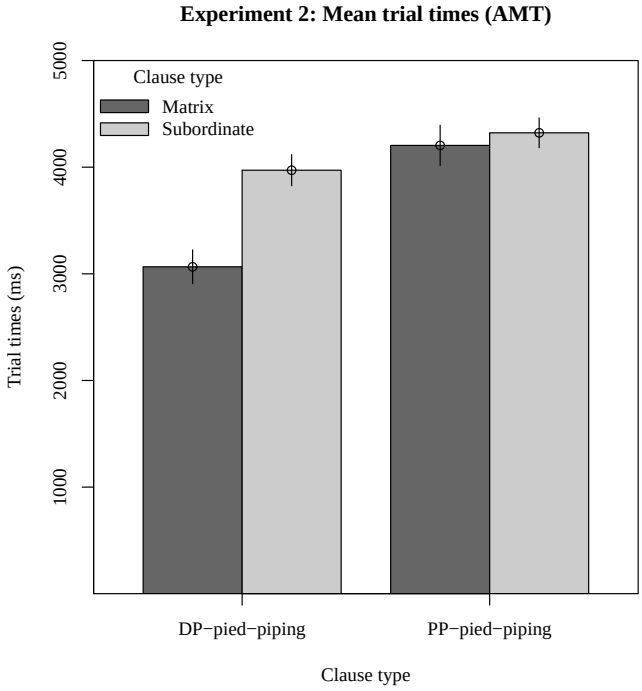
	DP-pied-piping	PP-pied-piping	Mean
Matrix	6.60 (0.07)	4.65 (0.15)	5.63 (0.10)
Subordinate	6.06 (0.10)	3.27 (0.14)	4.67 (0.12)
Mean	6.33 (0.06)	3.96 (0.11)	

Figure 2: Mean acceptability judgment ratings (standard errors in parentheses) by condition for the AMT group in Experiment 2.

of DPs (M = 6.33, SE = 0.06), t = -13.22, $p < 0.001$. Finally, we observed an interaction between the pied-piped element and embedding: the negative effect of the subordinate clause was greater when the pied-piped element was a PP (d = 1.38) than when it was a DP (d = 0.54), t = -4.03, $p < 0.001$. As the interaction is the crucial effect, we again wanted to test whether we were justified in retaining it in our model. Thus, as before, we removed the interaction and fit the data to a simpler model with just Clause type and Pied-piping element as additive factors. Although more complex, the model with the interaction term was a significantly better fit, χ^2 = 16.07, $p < 0.001$, justifying its explanatory role in the model.

We also collected the times that subjects spent on each trial. As before, this measure includes both the time to read the item and to rate it; we call this measure 'trial time' to reflect the fact that it is distinct from traditional reading time or decision time measurements. An additional caveat is in order; we hesitate to interpret the

data too closely for two reasons. First, participants were not monitored during the experiment, and thus trial times may include extraneous actions unrelated to the experiment. Second, we do not know enough about the method used to record the trial time measurement – in particular, the extent to which collection is delayed or modified during transmission across the Internet to the Ibex server. Nevertheless, we believe the results to be intriguing, and opt to present them here in Figure 3, cautions and concerns notwithstanding.



	DP-pied-piping	PP-pied-piping	Mean
Matrix	3065.98 (158.63)	4204.21 (189.47)	3631.08 (127.75)
Subordinate	3971.61 (146.00)	4322.39 (140.19)	4147.00 (101.56)
Mean	3517.20 (110.94)	4263.51 (117.57)	

Figure 3: Mean trial times in milliseconds (standard errors in parentheses) by condition for the AMT group in Experiment 2.

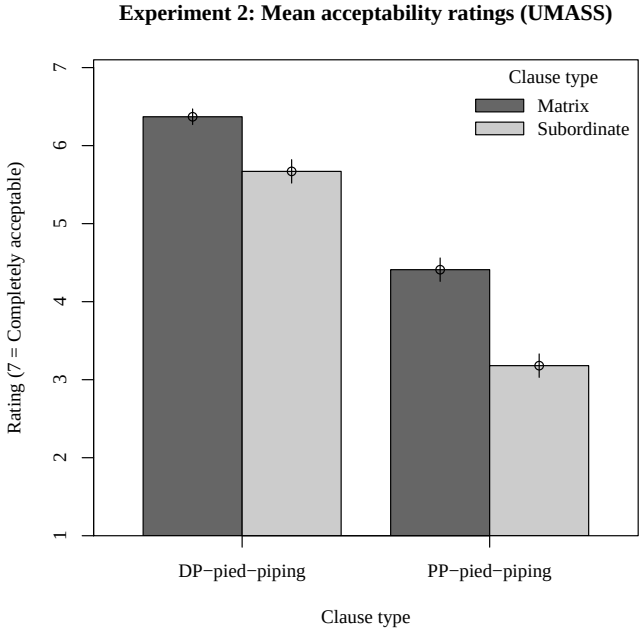
We observed a main effect of Clause type, such that items with subordinate clauses elicited significantly slower trial times (M = 4147ms, SE = 102), than Matrix counterparts (M = 3631ms, SE = 128), t = 4.435, $p < 0.001$. In addition, there was a main effect of pied-piping element: PP pied-piping items elicited significantly

slower trial times (M = 4263ms, SE = 117.57), than DP pied-piping counterparts (M = 3517ms, SE = 110.94), t = 5.15, $p < 0.001$. Finally, the clause type differentially affected the speed with which participants completed a trial: matrix clauses facilitated trial times more for DP-pied-piping structures (d = 906ms) than for PP pied-piping structures (d = 118ms), t = -2.72, $p < 0.01$. When combined with the acceptability judgments, the overall picture that emerges is that not only were DP pied-piping structures facilitated by the matrix clauses with respect to acceptability ratings, they were also facilitated with respect to trial time. Still, we hesitate to speculate about why the DP pied-piping structures were so greatly facilitated, as the pattern might be consistent with several different processing models.

While the crowd-sourcing methodology in psycholinguistics has previously shown identical results to results obtained from more traditional lab-based research (e.g., Munro et al., 2010; Sprouse, 2011; among others), it is still relatively new and little is known about what types of results generalize to other populations. Thus, despite all of our controls, we felt that we needed to replicate our core results with different subjects. Our second population was comprised of 88 undergraduate students at the University of Massachusetts Amherst, who completed the questionnaire online for course credit. We instituted the same procedure for excluding participants described above. Thirty participants were identified as non-native English speakers, duplicates, or simply inattentive, and removed from the analysis; another 15 were removed for counterbalancing purposes. The final dataset for replication consisted of 44 participants. Outliers were removed as before, with a comparable rate of data loss.

The second study indeed replicated the results from the first. Means and standard errors are provided in Figure 4 below. The statistical patterns in the acceptability rating data are in keeping with patterns obtained from the first group. In particular, participants rated pied piping in subordinate clauses (M = 4.43, SE = 0.13) significantly worse than pied-piping in matrix clauses (M = 5.39, SE = 0.11), t = -3.93, $p < 0.001$, consistent with the idea that the additional complexity of the subordinate clause introduces an independent cost. There was also a main effect of pied-piped element: participants rated structures with pied-piped PPs (M = 3.80, SE = 0.11) significantly worse than those with pied-piped DPs (M = 6.02, SE = 0.09), t = -11.12, $p < 0.001$. Lastly, an interaction between the two planned factors was observed, such that the effect of a subordinate clause on acceptability ratings was greater when the Pied-piped element was a PP (d = 1.23) than when it was a DP (d = 0.70), t = -2.10, $p < 0.05$. And, as before, this model fit the data better than a simpler model obtained from removing the interaction term, $\chi^2 = 4.41$, $p < 0.05$.

Interestingly, we did not replicate the trial time results form the AMT group: the only significant effect was that of clause type, in which subordinate clauses, as expected, elicited longer trial times than matrix counterparts. This is, as before, not



	DP-pied-piping	PP-pied-piping	Mean
Matrix	6.37 (0.10)	4.41 (0.15)	5.39 (0.11)
Subordinate	5.67 (0.15)	3.18 (0.15)	4.43 (0.13)
	6.02 (0.09)	3.80 (0.11)	

Figure 4: Mean acceptability judgment ratings (standard errors in parentheses) by condition for the UMASS group in Experiment 2.

clearly interesting, as items containing subordinate clauses were longer and more complex than items with only matrix clauses.

3.4 Discussion

As noted above, the results of Experiment 2 support the prediction presented in (7b); the effect of clause type was significantly stronger in cases of PP-pied-piping than in cases of DP-pied-piping. This supports the notion that PP-pied-piping is grammatically distinct from other pied-piping structures in English, in that it is essentially restricted to matrix environments, and therefore patterns with other cases of 'massive pied-piping'.

However, another explanation of these effects is imaginable. As noted earlier, all cases of 'non-massive' pied-piping are also not optional. That is, in English there is a

potential confound between pied-piping being 'massive' (in the technical sense) and it being optional. Thus, the special interaction between clause-type and pied-piping found for PPs might be due to its optionality rather than its 'massiveness'. On the other hand, it is currently unclear why the optionality of a pied-piping structure would lead to the effects witnessed above, while both Heck (2008) and Cable (2010) provide formal syntactic explanations for why 'massive pied-piping' should be restricted to matrix environments.

3.5 Conclusions

The experiments described above sought to provide a more rigorous test of the predictions in (7), repeated below.

(15) **Experimental Predictions to be Tested**

a. Pied-piping of PPs is less acceptable in English subordinate clauses than it is in main clauses.

b. The contrast between pied-piping in main clauses vs. subordinate clauses is seen only for PP pied-piping. A comparable contrast is not found for pied-piping of DPs by *wh*-determiners.

As reported above, Experiment 1 supported the prediction in (15a), while Experiment 2 supported that in (15b). The general picture that emerges is that the effect of subordination on acceptability ratings of structures with pied-piped PPs is greater than the effect of subordination on structures with either preposition stranding (Experiment 1) or pied-piped DPs (Experiment 2). With this experimental support, we can be more confident in the raw judgment data reported by both Heck (2008) and Cable (2010), as well as in their conclusion that pied-piping of PPs in English occupies a special status amongst pied-piping constructions. Unlike pied-piping by *wh*-determiners or possessors, pied-piping of PPs is restricted to main clauses, and thus should be categorized as 'massive pied-piping.' We interpret the results as an initial corroboration of the idea that pied-piping of PPs is thus actually 'illicit' in English, its ill-formedness simply being mitigated in matrix environments as a result of independent grammatical factors.

References

Allen, Cynthia. 1980. *Topics in Diachronic English Syntax*. New York: Garland Publishing.

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

Baayen, R. Harald, Douglas J. Davidson & Douglas M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59(4): 390–412.

Bates, Douglas & Martin Maechler. 2009. lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-31.

Cable, Seth. 2010. *The Grammar of Q: Q-Particles, Wh-Movement and Pied-Piping*. Oxford University Press.

Harris, Jesse A. & Christopher Potts. to appear. Predicting perspectival orientation for appositives. In *Proceedings from the 45th Annual Meeting of the Chicago Linguistic Society*, vol. 45. Chicago, IL: Chicago Linguistic Society.

Heck, Fabian. 2008. *On Pied-Piping: Wh-Movement and Beyond*. Studies in Generative Grammar 98, Mouton de Gruyter.

Munro, Robert, Steven Bethard, Victor Kuperman, Vicky Tzuyin Lai, Robin Melnick, Christopher Potts, Tyler Schnoebelen & Harry Tily. 2010. Crowdsourcing and language studies: The new generation of linguistic data. In *NAACL 2010 Workshop on Creating Speech and Language Data With Amazon's Mechanical Turk*. Los Angeles: ACL.

R Development Core Team. 2008. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ross, John Robert. 1967. *Constraints on variables in syntax*. Ph.D. thesis, MIT, Cambridge, MA.

Sprouse, Jon. 2011. A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods* (43).

Appendix A: Experiment 1 materials

1. a. Who did he dance with?
   b. I wonder who he danced with.
   c. With whom did he dance?
   d. I wonder with whom he danced.
2. a. Who did she talk to?
   b. I wonder who she talked to.
   c. To whom did she talk?
   d. I wonder to whom she talked.
3. a. Who did he work for?
   b. I wonder who he worked for.
   c. For whom did he work?
   d. I wonder for whom he worked.

4. a. Who did she smile at?
   b. I wonder who she smiled at.
   c. At whom did she smile?
   d. I wonder at whom she smiled.
5. a. Who did he travel with?
   b. I wonder who he traveled with.
   c. With whom did he travel?
   d. I wonder with whom he traveled.
6. a. Who did she sing to?
   b. I wonder to whom she sang.
   c. To whom did she sing?
   d. I wonder to whom she sang.
7. a. Who did he yell at?
   b. I wonder who he yelled at.
   c. At whom did he yell?
   d. I wonder at whom he yelled.
8. a. Who did she cook for?
   b. I wonder who she cooked for.
   c. For whom did she cook?
   d. I wonder for whom she cooked.

Appendix B: Experiment 2 materials

1. a. About what did she write?
   b. It's obvious about what she wrote.
   c. Which book did she write?
   d. It's obvious which book she wrote.
2. a. With what does Bill work?
   b. I asked with what Bill works.
   c. Which register does Bill work?
   d. I asked which register Bill works.
3. a. On what did Sarah cook?
   b. They wonder on what Sarah cooked.
   c. Which soup did Sarah cook?
   d. They wonder which soup Sarah cooked.

4. a. Near what did Lou perform?
   b. Kate knows near what Lou performed.
   c. Which act did Lou perform?
   d. Kate knows which act Lou performed.
5. a. From what did Bill run?
   b. I know from what Bill ran.
   c. Which race did Bill run?
   d. I know which race Bill ran.
6. a. With what did he draw?
   b. Fran asked with what he drew.
   c. Which picture did he draw?
   d. Fran asked which picture he drew.
7. a. With what did John doodle?
   b. I wonder with what John doodled.
   c. Which picture did John doodle?
   d. I wonder which picture John doodled.
8. a. About what did Joan mumble?
   b. It's obvious about what Joan mumbled.
   c. Which excuse did Joan mumble?
   d. It's obvious which excuse Joan mumbled.
9. a. From what did Peter drink?
   b. Barb wonders from what Peter drank.
   c. Which wine did Peter drink?
   d. Barb wonders which wine Peter drank.
10. a. From where did she escape?
    b. I know from where she escaped.
    c. Which prison did she escape?
    d. I know which prison she escaped.
11. a. With what did she eat?
    b. He wonders with what she ate.
    c. Which dish did she eat?
    d. He wonders which dish she ate.
12. a. Towards what did Bill drive?
    b. I asked towards what Bill drove.
    c. Which car did Bill drive?
    d. I asked which car Bill drove.

Catch items. Participants saw each of the following four items, randomly dispersed within the experiment. Participants who rated any one of these items a 3 or above on the 7-point scale were excluded from analysis.

1. Most book did Lance enjoy in.
2. More scooters than bicycles students has riden.
3. Doris talked a book a page.
4. I think he remembers he.