Ceiling Effects on Weight in Heavy NP Shift

David J. Medeiros
Paul Mains
Kevin B. McGowan

Abstract: This squib tests theories of heavy NP shift that link constituent order to parsing. Our results indicate that increasing the weight of an object NP cannot make a heavy NP shift construction more acceptable than a comparison sentence in the canonical order, contra parsing-based theories. In addition, we examine acceptability as it relates to verb disposition. We find no significant differences between different subcategorizations, contrary to the findings of corpus studies. These results reveal a disconnect between production and comprehension; we further conclude that parser sensitivity to constituent structure is unlikely to affect speaker production of heavy NP shift.

Keywords: heavy NP shift, constituent order, English, experimental syntax

1 Introduction

The idea that constituent length (often characterized as end-weight) affects word order was originally discussed by Behaghel (1910), who observed that larger constituents generally follow shorter constituents. For Behaghel, a larger constituent is one with more words, such that weight is equated to length, a notion echoed in Hawkins 1990 and later work. For example, Hawkins (1994) suggests that constituent ordering in language can, very generally, be reduced to a parsing principle favoring linear orders that allow identification of constituent structure (e.g., within a VP) to be recognized as quickly as possible, thereby explaining the length effect.

To see how this principle works, consider (1), from Hawkins 1994:57), which illustrates canonical English word order in (1a) and the heavy NP shift (HNPS) word order in (1b). According to Hawkins, (1b) is preferred to (1a) because all of the constituents within the VP are encountered (not necessarily completed) more quickly in the HNPS word order. Specifically, in (1b) the V, PP, and NP constituents are encountered within four words from the beginning of the VP (i.e., when the initial *the* in the object NP is reached), whereas in (1a) the parser must wait until the initial *to* in the PP for all of the subconstituents in VP to be encountered. Hawkins (1994:77) formalizes this as the *Early Immediate Constituents* principle, which states, informally, that "the size of a CRD [Constituent Recognition Domain] will be as small and efficient as it can possibly be for the recognition of a mother node and its immediate constituents."[1] (See Hawkins 1994:77 for full formalization.)

The authors would like to acknowledge the audience at the 89th Annual Meeting of the Linguistic Society of America, as well as two anonymous reviewers, whose feedback greatly improved the manuscript. We dedicate this squib to the memory of Samuel David Epstein, who will always remain an inspiration to those he mentored.

[1] Hawkins subsumes the Early Immediate Constituents principle under a new principle, Minimize Domains, in later work (Hawkins 2004, 2014).

Linguistic Inquiry, Volume 52, Number 2, Spring 2021
426–440

https://doi.org/10.1162/ling_a_00382

(1) a. I $_{\text{VP}}$[gave $_{\text{NP}}$[the valuable book that was extremely difficult to find] $_{\text{PP}}$[to Mary]].
b. I $_{\text{VP}}$[gave $_{\text{PP}}$[to Mary] $_{\text{NP}}$[the valuable book that was extremely difficult to find]].

While a number of factors are known to affect likelihood of the HNPS construction in corpora, the theory of constituent ordering expressed in Hawkins 1994 and subsequent research crucially links constituent length, or weight, to parser identification of constituent structure.

Alternatively, Wasow (1997, 2002) articulates the view that HNPS may be related to speech planning, not parsing. Wasow argues that the interests of speech production conflict with parsing. Under this view, parsers (echoing Hawkins's analysis) are taken to prefer knowledge of VP-internal constituent structure as early as possible, what Wasow terms "early commitment." By contrast, speech producers are expected to keep options open for decision making, preferring "late commitment." Wasow (1997) offers evidence from written and spoken corpora which show that HNPS is more common among optionally transitive verbs as compared to obligatorily transitive verbs, insofar as speech producers are more likely to use HNPS with optionally transitive verbs as they delay their commitment to a transitive subcategorization. At the same time, Wasow suggests that the opposite pattern (greater preference for HNPS with obligatorily transitive verbs, for which the subcategorization is known) would be expected, if indeed constituent identification in comprehension is relevant for HNPS. However, prior research has not directly examined the relevance of constituent recognition for parsing in a comprehension study.

In this squib, we test the theory that the parser's early identification of VP-internal constituent structure is related to HNPS. To this end, we present two analyses of an acceptability judgment task. The first analysis relates the length of direct object NPs to acceptability, while controlling other known factors to the extent possible. The second analysis compares acceptability of HNPS to verb disposition, comparing obligatorily to optionally transitive verbs. In the first analysis, our results show a ceiling effect in the relationship between weight and acceptability. In other words, increasing the weight of an object NP in a HNPS construction increases acceptability, but only to a point, such that further increase in weight does not result in higher acceptability. Likewise, increasing the length of an object NP in the canonical order lowers acceptability only to a point, *pace* Hawkins (1994), and never results in significant preference for the shifted order. In the second analysis, we do not find greater acceptability of HNPS for obligatorily transitive verbs, as an "early commitment" theory would predict. Given these results, we suggest that parser identification of VP-internal constituent structure does not explain patterns of HNPS. To the extent that the weight effect in HNPS is similar to weight effects in other constructions, such as dative alternation, particle movement, and the ordering of multiple PPs (see, e.g., Wasow 1997, Hawkins 1999, Wasow and Arnold 2003, Melnick 2017), the results presented here may also have ramifications for predictive models of constituent ordering beyond HNPS.

2 Background

HNPS has been an important topic of discussion in both formal and psycholinguistic research for some time. For example, Chomsky (1975) argues that grammatical weight should be equated with the complexity of the relevant NP, where complexity is defined as the inclusion, within the NP, of other phrasal nodes (see Ross 1967, Quirk 1972, Emonds 1976, Erdmann 1988, Dik 1989, Hawkins 1994, and Rickford et al. 1995 for related proposals). One complicating factor is that the length of an NP and its complexity (under any of the phrasal-node-based accounts) are highly correlated, making any judgment between these two metrics difficult. Additional views on grammatical weight identify weight with phonological phrases (two or more; Inkelas and Zec 1990) and information structure (givenness versus newness; Niv 1992).[2]

Wasow 1997, based primarily on corpus evidence, initiated a new interest in weight-based phenomena, including HNPS as well as the dative alternation and particle movement.[3] Three key findings in Wasow 1997 are the following. First, weight is a gradient principle, such that increasing weight predicts greater likelihood of HNPS in a given corpus. Second, measurements of weight by length, total number of nodes (contained in the NP), and total number of phrasal nodes are all highly correlated, though Wasow (1997) suggests that length in terms of word count is the clearest predictor. Finally, the weight of a shifted constituent relative to an intervening constituent is a stronger predictor of HNPS than the weight of the shifted constituent alone (see also Erdmann 1988, Stallings and MacDonald 2011).

In Wasow 1997 and subsequent literature, the predictors of HNPS and other weight-related phenomena have been made more precise, and we summarize several of these in the remainder of this section.

[2] We focus only on English here, but weight-sensitive phenomena occur in other languages as well. For example, Hawkins (1994) shows that longer constituents tend to precede shorter constituents in head-final languages such as Japanese and Korean. Hawkins (1994:66–68) explains the long-before-short preference in head-final languages by computing the "window" for a CRD as soon as the first *head* (not word) of the first subconstituent is reached. Schematically, in a head-final VP structure in which x_n is a word inside phrases AP and BP with heads A & B such as $[[x_1\ x_2\ x_3\ A]_{AP}\ [x_1\ x_2\ B]_{BP}\ V]_{VP}$, the CRD is computed starting at head A, not x_1 inside A, although it does include x_1, x_2, and B inside BP, parallel to how the CRD in (1b) extends from *gave* to *the*, but does not include any further words inside NP.

[3] Note that Wasow uses terms such as *HNPS* and *particle "movement"* due to convention, and not as an endorsement of transformational approaches.

As discussed above, Wasow (1997) shows that optionally transitive verbs are more likely to occur in HNPS constructions than those that are obligatorily transitive. In a related finding, Stallings, MacDonald, and O'Seaghdha (1998) show that verbs that can take either NP or S complements are more acceptable in HNPS constructions than verbs that only take NP complements, arguing that the frequent occurrence of V-PP/Adv-S word order for these verbs facilitates the processing of HNPS (see also Staub, Clifton, and Frazier 2006).

The characterization of heaviness is made further precise by Arnold et al. (2000), who find that NP complexity is a predictor of acceptability of HNPS, with corpus data showing that length and complexity both predict HNPS, although these two measures are highly correlated. In one of the few acceptability studies on HNPS, Wasow and Arnold (2005) confirm the role of NP complexity when length is held constant. In addition to length and complexity, Arnold et al. (2000) show that information structure is a predictor, with new information following given information (they also note that newness is correlated with length as well).

Wasow (1997) and Arnold et al. (2000) also discuss semantic connectedness of constituents as a predictor of HNPS. For example, Wasow (1997) shows that V + PP idioms such as *take into account* are much more likely to occur in HNPS constructions than nonidioms, such that (e.g.) *take into account our concerns* is more likely to occur than *share with others that cost*, because only the former involves an idiomatic interpretation. Hawkins (1999) and Lohse, Hawkins, and Wasow (2004) show similar semantic effects on constituent placement in related constructions.

More recently, and on the basis of detailed corpus evidence, Melnick (2017) shows that individuals vary with respect to their use of weight-related constructions, arguing that this favors an approach to weight effects based on cognitive resources (since these resources are taken to vary across individuals). Melnick also shows that this individual variation holds across weight-sensitive constructions, and also extends to dialect variation (e.g., American vs. Australian English).

In sum, the prior literature has identified several factors relevant to HNPS and other weight-related constructions; no single factor can account for observed constituent order alternations (see also McDonald, Bock, and Kelly 1993). Nevertheless, end-weight, now typically formalized as relative constituent length, remains a central factor for any understanding of HNPS (Wasow and Arnold 2003). Note, however, that the evidence given in support of the research cited above is largely drawn from corpora or speech production (possibly after a planning period, as in Stallings et al. 1998), prompting the comprehension-oriented study described in the following section.

3 Experiment

Given that constituent length relates to constituent identification within the parsing-based model discussed above, the present experiment compares object NP weight in HNPS constructions to acceptability using NP lengths from 2 to 11 words, while holding the intervening constituent to 2 words in all items. Word and syllable length are modeled as continuous, rather than categorical, variables, allowing us to fit a regression model to these variables rather than comparing variance between length means. Finally, the use of linear mixed-effects regression allows us to control for differences in acceptability rating due to a particular verb or due to individual participant differences (e.g., some participants may just give lower ratings overall or alter their ratings differently as a function of NP length).

We recruited participants via Amazon Mechanical Turk, which has been independently validated for linguistics research by Sprouse (2011) and Munro et al. (2010). We restricted our survey to Mechanical Turk workers with IP addresses in the United States. Further, workers needed to have completed at least 50 tasks within Mechanical Turk with a 95% approval rating for all prior tasks in order to take part in this experiment. Having satisfied these requirements, 209 participants started the survey, with 196 completing it. Participants gave informed consent via the survey prior to seeing instructions for the experiment. All of them were paid for their participation regardless of their response or completion.

3.1 Materials and Method

In order to isolate the effect of NP weight on HNPS, we held constant other known factors that predict HNPS acceptability. We created 10 lists, each based upon one verb that selects both NP and S. Each of the 10 lists consisted of 20 associated sentences (based upon a single verb), including 10 nonshifted/shifted pairs—200 test sentences total. All of the test sentences appear in the online appendix (https://www.mitpressjournals.org/doi/suppl/10.1162/ling_a_00382).

From the 10 lists (= 200 test items) described above, we created 20 sublists, such that each sublist contained only one test item per verb (controlling, to the extent possible, effects of givenness/newness), drawing either one shifted or nonshifted sentence from each verb. Each sublist therefore contained 10 test items, 5 with shifted and 5 with nonshifted word order, with shifted and nonshifted items balanced through the list (i.e., 2-word NP shifted, 3-word NP nonshifted, 4-word NP shifted, etc.). An example of one such sublist appears in (2). Each of the 20 sublists also contained 30 fillers, which represented a range of acceptability. Qualtrics survey software controlled random presentation of each sublist (10 test items plus 30 fillers) to participants. All survey items (test items and fillers) began with a proper name followed by a verb in the past tense.

(2) a. Alex indicated to Curtis the problems.
b. Kim proposed some policy changes to George.
c. Todd suggested to Frank a large gift basket.
d. Jacob recommended lots of rest and water to Jerry.

e. Matthew confessed to Kenny the story about his secret relationship.
f. Tim stated the guidelines for writing successful science papers to Kelly.
g. Bob announced to Ian the company's earnings from the last fiscal year.
h. Angie muttered some sarcastic and critical comments about the romantic comedy to Jen.
i. Erica explained to Joan some surprisingly useful tips and tricks about how computers work.
j. Mary mentioned the real estate listing for an expensive house on lakefront property to Frank.

All verbs in the test items subcategorize for both NP and S, which is known to increase HNPS acceptability (Stallings, MacDonald, and O'Seaghdha 1998). We also limit all potential interveners to 2-word PPs, which should again increase overall acceptability (as compared to longer interveners) (Wasow 1997). While the longest object NPs in our study include NP conjunction and embedded PPs, none of them contain embedded clauses, which should have the overall effect of decreasing acceptability for the HNPS items (Arnold et al. 2000).

Each participant also saw introductory and concluding sections; the introductory section contained 2 additional fillers in fixed order (i.e., all participants saw the same 2 introductory fillers), and the concluding section contained 1 additional fixed filler (participants were not aware that these introductory and concluding fillers were not part of the randomized survey). Participants therefore rated 43 sentences in total: the 10 test items and 30 fillers in random presentation, as well as the 2 fillers presented in fixed order at the beginning of the survey and the 1 fixed filler at the end of the survey. The introductory section also queried participants about native language status.

Finally, we identified 4 nontest items, 1 unquestionably acceptable and 3 unquestionably unacceptable, which we used to monitor participants' responses. We did this because it is known that some participants in this type of survey experiment will answer randomly. Any participant who did not assign comparatively low or high values to these preselected fillers was discarded from the analysis.

Participants were asked to read and then rate each sentence along a 7-point Likert scale, with 1 designated as "not at all natural," 4 designated as "somewhat natural," and 7 designated as "very natural." Participants saw one item per screen, and the designations for the scale appeared with every survey item. Participants had to provide a ranking on the 1–7 scale before moving on to the next item.

In all, 209 participants took the survey. Of these, 16 were excluded from the analysis. Of the 16 participants whose data were excluded, 13 did not finish. While all participants reported affirmatively that they spoke "English at home while growing up," 1 participant reported "other" for region of origin and was therefore excluded. The final 2 participants who were excluded appeared to answer the survey questions randomly, given their performance on the preselected filler items. After we excluded these 16 participants, data from 193 participants remained, providing a total of 1,930 ratings on test sentences.[4]

3.2 Analysis 1: Length and Acceptability

Data were analyzed using linear mixed-effects modeling with the lme4 package (Bates et al. 2015) in the open-source R statistical computing environment (R Core Team 2016). All categorical variables were sum-coded to allow for the interpretation of possible lower-order effects in the models as main, rather than simple, effects. Models were fitted with the maximal random effects structure justified by the data and by model comparison to avoid the inflated risk of type 1 errors in models that include random intercepts without random slopes (Barr et al. 2013). This structure includes random intercepts for participant (modeling the case where some participants habitually provide higher or lower acceptability ratings than others), random slopes for participant (modeling the case where some participants change their acceptability ratings quickly in response to changes in NP length while other participants may react more slowly), and random intercepts and slopes for verb (assuming that verbs may not be uniform in licensing shift or the extent to which length increases acceptability). Model comparison was also used to justify the inclusion of fixed effects and interaction terms in the linear models, while statistical significance within the resulting models is reported using Satterthwaite's approximations as implemented in the R package lmerTest (Kuznetsova, Brockhoff, and Christensen 2013).

A linear mixed-effects model in which acceptability score was the dependent variable and in which word length and whether the item was shifted or nonshifted are included as predictors (fixed effects). There are significant main effects both of word length ($\beta = 0.034$, $t = 3.365$, $p < .001$) and of shiftedness ($\beta = 1.118$, $t = 12.541$, $p < .001$). Model comparison motivates inclusion of an interaction term that is significant ($\beta = -0.118$, $t = -11.515$, $p < .001$). This interaction is visible in figure 1, which plots mean scores across all participants for shifted and nonshifted sentences by word length of the object NP. Error bars on data points represent the confidence intervals computed by the statistical analysis, in which significant differences in acceptability obtain between shifted and nonshifted sentences when NP length varies from 2 words to 7 words, the nonshifted (canonical) word order being more acceptable. Once NP length exceeds 7 words, no significant differences are found between the shifted and nonshifted word orders, both orders having an average acceptability rating above 5 for all but one item type (shifted sentences with 8-word-long NPs).

[4] Data files including the original data collected (which includes data from the 16 excluded subjects), as well as a file including only the data used in the analysis, are available at https://www.mitpressjournals.org/doi/suppl/10.1162/ling_a_00382.

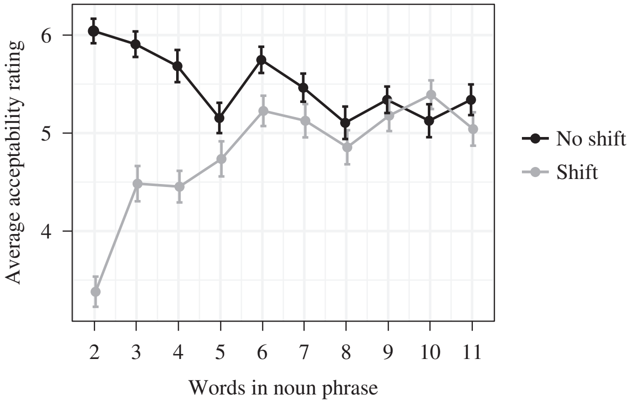

Figure 1
Acceptability by length (words). Each vertical pair of dots represents NP word length, from 2 to 11 words.

At the same time, the shifted word order is least acceptable when NP length is 2 words, and acceptability for the shifted word order increases gradiently until NP length reaches 7 words, at which point acceptability reaches a ceiling.

We also examined an alternative potential measure of weight, comparing the word length measure to syllable count. In our stimuli, syllable increase does not increase at the same rate as word increase. We found essentially the same pattern of data when defining weight as syllable length as we observed when defining weight by word count. While increasing syllable length also increased the acceptability of shifted items ($\beta = 1.001$, $t = 12.294$, $p < .001$), a ceiling effect was again observed—visible as an interaction between syllable length and shiftedness ($\beta = -.062$, $t = -10.795$, $p < .001$). At no point were shifted items significantly more acceptable than nonshifted items, as indicated in figure 2. Significant differences between shifted and nonshifted items obtain until 12 syllables, beyond which point both shifted and nonshifted examples are quite acceptable. Model comparison of the full model for word count and the full model for syllable count reveals that word count provides a better model of the change in acceptability ratings; this difference is significant ($p < .001$).

3.3 Analysis 2: Transitivity and Acceptability

The results of the first analysis suggest that a direct relationship between NP length and acceptability does not hold past the ceiling, counter to the predictions of the theory presented in Hawkins 1994 and subsequent work. We further examined the role of constituent identification for our subjects in a post hoc analysis of verb disposition and acceptability.

According to Wasow (1997), corpus data show a higher rate of HNPS for verbs that are optionally transitive than for those that are

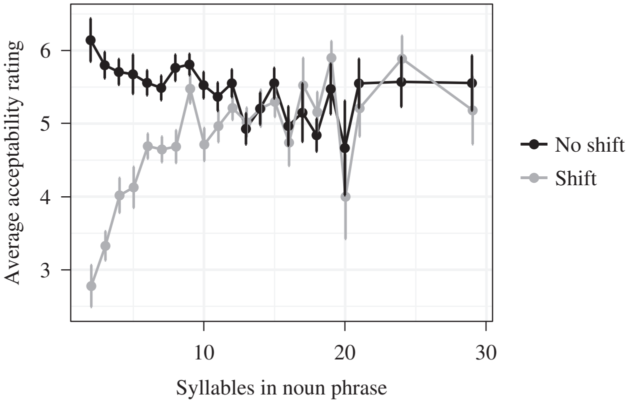

Figure 2
Acceptability by length (syllables). Each vertical pair of dots represents NP syllable length, from 2 to 29 syllables.

obligatorily transitive. Limiting his sample to verbs that occurred in HNPS and/or with both NP and PP sisters a specified number of times, Wasow (1997:97–100) reports significantly more occurrences of HNPS for optionally transitive verbs in the Brown ($p < .02$) and Switchboard ($p < .01$) corpora. Wasow suggests that this pattern of data can be explained by "late commitment" in speech production, such that speakers and writers will delay committing to a transitive subcategorization if the relevant verb allows both intransitive and transitive subcategorizations. Wasow speculates that speech comprehenders would behave differently, preferring "early commitment," such that the subcategorization of an optionally transitive verb should be identified as soon as possible, by avoiding HNPS (and likewise favoring HNPS for obligatorily transitive verbs). Given his corpus results, Wasow suggests that speech production, and not parsing, is the major factor that drives HNPS rates in corpora.

To determine if our participants found HNPS more acceptable with obligatorily transitive verbs, we coded our verbs as V_t if obligatorily transitive and as V_p if optionally transitive. The transitivity status was determined in consultation with several dictionaries; because the IP addresses of our subjects were restricted to those originating in the United States, we used American English when dialect differences obtained. We judged the verbs as follows:[5]

[5] Classifying these verbs is nontrivial. *Indicate* has an intransitive use in the United Kingdom (= 'to turn-signal') and *mention* has an intransitive entry in the *Oxford English Dictionary* ('to mention about something', with the most recent example in 1925), but no other dictionary that we consulted included this. The intransitive use of *announce* may be limited to American English, as in announcing for candidacy or working as an announcer (e.g., in broadcast media).

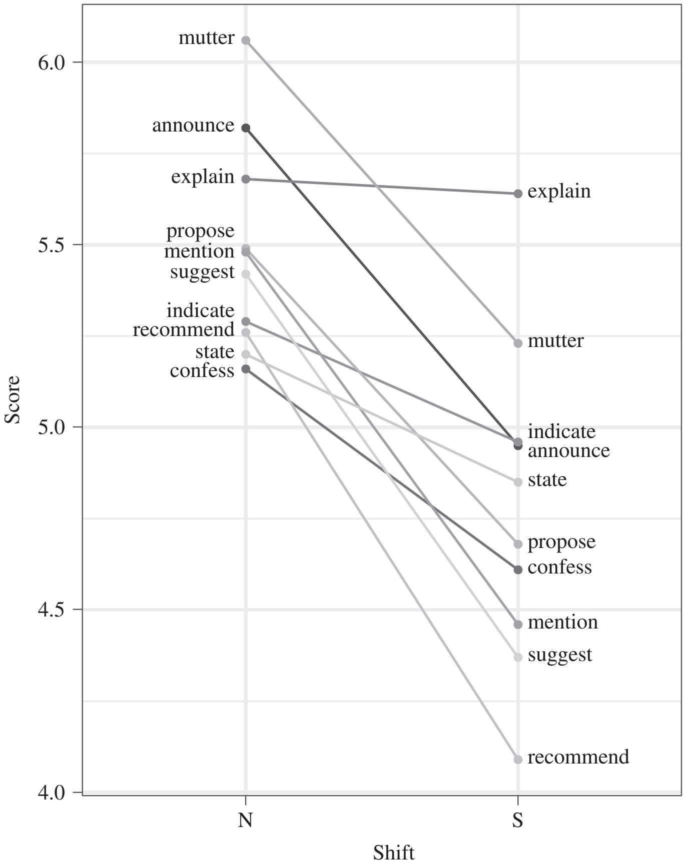

Figure 3
Acceptability by verb in nonshifted (N) and shifted (S) conditions

(3) a. V_t: indicate, mention, recommend, state, suggest
b. V_p: announce, confess, explain, mutter, propose

First, we examined overall means for each verb in the shifted and nonshifted conditions. These descriptive statistics are visualized in figure 3, in which the overall mean for the verb in the nonshifted condition appears on the left and the overall mean in the shifted condition appears on the right. Figure 3 shows that moving from the canonical to the HNPS condition affects certain verbs more than others (also, the verbs are more closely clustered in the nonshifted condition than the shifted condition). For example, *explain* is hardly affected at all, though it is readily used intransitively. By contrast, *recommend* and *suggest* have a much lower score in the shifted condition than the nonshifted condition. These results already suggest that comprehenders do not have a preference for HNPS with obligatorily transitive verbs.

Using the same software packages described above, we additionally analyzed the effect of verb disposition with a linear effects model in which acceptability score was the dependent variable and transitivity status was the predictor. Examining all sentences with HNPS ($n =$ 960 sentences), the model includes random intercepts for participants and random intercepts for verbs, assuming that these differ in their transitivity bias (separate from subcategorizations). Our participants did not have a statistically significant preference for HNPS with either verb disposition, with a nonsignificant trend toward preferring optionally transitive verbs ($\beta = 0.4683$, $t = 1.876$, $p = .0972$), as illustrated in figure 4.

While Wasow (1997) suggests that speech comprehenders would strongly prefer HNPS in the V_p condition if constituent identification is relevant for the parsing of HNPS, we found a nonsignificant trend in the opposite direction, aligning our participants (who represent speech comprehension) with the corpus data (speech production) in this respect.[6]

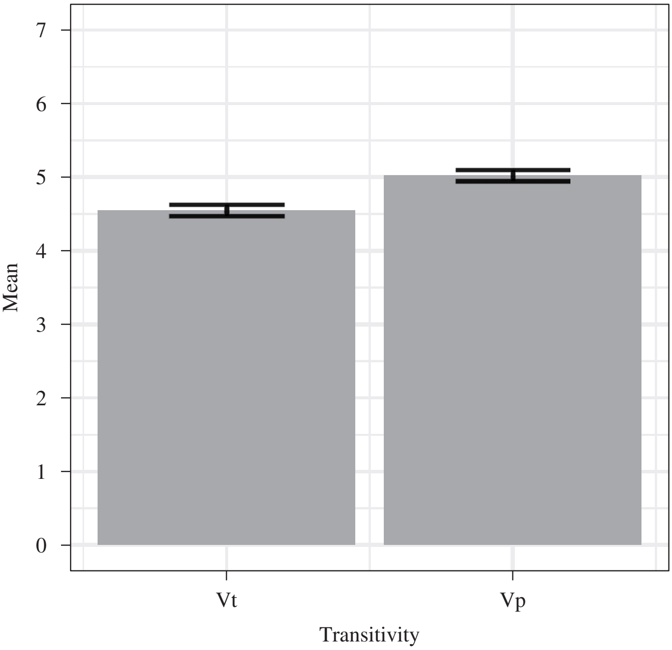

Figure 4
Acceptability by verb disposition. Error bars represent standard error.

[6] Given the verb means visualized in figure 3, excluding verbs that are controversially V_t or V_p only increases the effect already observed here. This is because the highest-rated verbs, *explain* and *mutter*, are noncontroversially V_p, while the lowest-rated, *recommend* and *suggest*, are noncontroversially V_t.

4 Discussion

The results described above fail to support theories of HNPS that relate use of HNPS to consideration of the parser's ability to identify subconstituents within VP. Within Hawkins's (1994, 2004) theory, HNPS is increasingly favored as the argument NP's length increases relative to that of an intervener between the argument and selecting verb. However, our participants assigned higher ratings to longer NPs (relative to the PP interveners) only up to the observed ceiling. Similarly, the presence of a very long object NP does not drive the acceptability of the canonical word order down indefinitely; here a floor effect is observed. In fact, the canonical word order items never average a score below 5 (on the 7-point scale) in our results, regardless of NP length.

In addition, our participants did not prefer HNPS for obligatorily transitive verbs, contrary to what Wasow (1997) suggests would be the case if parser sensitivity to constituent identification obtains. We therefore conclude that recognition of constituent structure on behalf of the parser is unlikely to be a factor that influences the occurrence of HNPS.

We also note a disconnect between corpus and experimental studies given our results. While our participants are sensitive to end-weight up to the ceiling effect, increasing NP length appears to bring HNPS only to a par with the canonical word order, such that HNPS is never significantly preferred over the canonical word order. However, Wasow (2002) reports that rates of HNPS can exceed 50% when relative weight differences are large.[7] The source of this disconnect between the acceptability judgment task reported here and corpus data warrants further research and highlights the importance of testing theories against multiple types of data (see, e.g., Arppe et al. 2010). We tentatively suggest two possibilities for this disconnect. First, Wasow (1997, 2002) may be correct that HNPS largely serves the needs of speech production, not comprehension, though our study does show an effect of end-weight, if limited. Or it could be that the alternative, largely discourse or syntactic (via NP complexity) factors that are known to correlate with weight are underestimated in the prior literature (several of these are discussed in section 2). Under this view, weight itself may be less important than previously thought, and therefore does not influence the results of isolated sentences in an acceptability task in the way corpus research would lead one to expect.

References

Arnold, Jennifer E., Anthony Losongco, Thomas Wasow, and Ryan Ginstrom. 2000. Heaviness vs. newness: The effects of struc-

[7] A reviewer notes that a similar discrepancy between corpus and acceptability judgments occurs with English resumptive pronouns, which are common in usage but typically judged unacceptable; see Cann, Kaplan, and Kempson 2005 and Polinsky et al. 2013 for examples and review.

tural complexity and discourse status on constituent ordering. *Language* 76:28–55.

Arppe, Antti, Gaëtanelle Gilquin, Dylan Glynn, Martin Hilpert, and Arne Zeschel. 2010. Cognitive corpus linguistics: Five points of debate on current theory and methodology. *Corpora* 5:1–27.

Barr, Dale J., Roger Levy, Christoph Scheepers, and Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68:255–278.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67:1–48.

Behaghel, Otto. 1910. Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern. *Indogermanische Forschungen* 25: 110–142.

Cann, Ronnie, Tami Kaplan, and Ruth Kempson. 2005. Data at the grammar-pragmatics interface: The case of resumptive pronouns in English. *Lingua* 115:1551–1577.

Chomsky, Noam. 1975. *The logical structure of linguistic theory.* New York: Plenum Press.

Dik, Simon C. 1989. *The theory of functional grammar.* Part 1, *The structure of the clause.* Dordrecht: Foris.

Emonds, Joseph E. 1976. *A transformational approach to English syntax: Root, structure-preserving, and local transformations.* New York: Academic Press.

Erdmann, Peter. 1988. On the principle of 'weight' in English. In *On language, rhetorica phonologica syntactica: A festschrift for Robert P. Stockwell from his friends and colleagues*, ed. by Caroline K. Duncan-Rose and Theo Vennemann, 325–339. London: Routledge.

Hawkins, John A. 1990. A parsing theory of word order universals. *Linguistic Inquiry* 21:223–261.

Hawkins, John A. 1994. *A performance theory of order and constituency.* Cambridge: Cambridge University Press.

Hawkins, John A. 1999. The relative order of prepositional phrases in English: Going beyond manner–place–time. *Language Variation and Change* 11:231–266.

Hawkins, John A. 2004. *Efficiency and complexity in grammars.* Oxford: Oxford University Press.

Hawkins, John A. 2014. *Cross-linguistic variation and efficiency.* Oxford: Oxford University Press.

Inkelas, Sharon, and Draga Zec. 1990. *The phonology-syntax connection.* Chicago: University of Chicago Press.

Kuznetsova, Alexandra, Per B. Brockhoff, and Rune H. B. Christensen. 2013. *lmerTest: Tests for random and fixed effects for linear mixed effect models.* https://CRAN.R-project.org/package=lmerTest, r package version 2.0-11.

Lohse, Barbara, John A. Hawkins, and Thomas Wasow. 2004. Domain minimization in English verb-particle constructions. *Language* 80:238–261.

McDonald, Janet L., Kathryn Bock, and Michael H. Kelly. 1993. Word and world order: Semantic, phonological, and metrical determinants of serial position. *Cognitive Psychology* 25:188–230.

Melnick, Robin. 2017. Consistency in variation: On the provenance of end-weight. Doctoral dissertation, Stanford University.

Munro, Robert, Steven Bethard, Victor Kuperman, Vicky Tzuyin Lai, Robin Melnick, Christopher Potts, Tyler Schnoebelen, and Harry Tily. 2010. Crowdsourcing and language studies: The new generation of linguistic data. In *CSLDAMT '10: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 122–130. Stroudsburg, PA: Association for Computational Linguistics.

Niv, Michael. 1992. Right association revisited. In *ACL '92: Proceedings of the 30th annual meeting of the Association for Computational Linguistics*, 285–287. Stroudsburg, PA: Association for Computational Linguistics.

Polinsky, Maria, Lauren Eby Clemens, Adam Milton, Ming Xiang Morgan, and Dustin Heestand. 2013. Resumption in English. In *Experimental syntax and island effects*, ed. by Jon Sprouse and Norbert Hornstein, 341–359. Cambridge: Cambridge University Press.

Quirk, Randolph. 1972. *A grammar of contemporary English.* London: Longman.

R Core Team. 2016. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.

Rickford, John R., Thomas Wasow, Norma Mendoza-Denton, and Julie Espinoza. 1995. Syntactic variation and change in progress: Loss of the verbal coda in topic-restricting *as far as* constructions. *Language* 71:102–131.

Ross, John R. 1967. Constraints on variables in syntax. Doctoral dissertation, MIT.

Sprouse, Jon. 2011. A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods* 43:155–167.

Stallings, Lynne M., and Maryellen C. MacDonald. 2011. It's not just the "heavy NP": Relative phrase length modulates the production of heavy-NP shift. *Journal of Psycholinguistic Research* 40:177–187.

Stallings, Lynne M., Maryellen C. MacDonald, and Padraig G. O'Seaghdha. 1998. Phrasal ordering constraints in sentence production: Phrase length and verb disposition in heavy-NP shift. *Journal of Memory and Language* 39:392–417.

Staub, Adrian, Charles Clifton Jr., and Lyn Frazier. 2006. Heavy NP shift is the parser's last resort: Evidence from eye movements. *Journal of Memory and Language* 54:389–406.

Wasow, Thomas. 1997. Remarks on grammatical weight. *Language Variation and Change* 9:81–105.

Wasow, Thomas. 2002. *Postverbal behavior.* Stanford, CA: CSLI Publications.

Wasow, Thomas, and Jennifer Arnold. 2003. Post-verbal constituent ordering in English. *Topics in English Linguistics* 43:119–154.

Wasow, Thomas, and Jennifer Arnold. 2005. Intuitions in linguistic argumentation. *Lingua* 115:1481–1496.

David J. Medeiros
Department of Linguistics/TESL
California State University, Northridge

david.medeiros@csun.edu

Paul Mains
Google

paul.m.mains@gmail.com

Kevin B. McGowan
Department of Linguistics
University of Kentucky

kbmcgowan@uky.edu