Ceiling Effects on Weight in Heavy NP Shift

To appear in Linguistic Inquiry

David J. Medeiros

California State University, Northridge

Paul Mains

Google

Kevin B. McGowan

University of Kentucky

Abstract

The tendency for longer, complex phrases to come at the end of clauses has been formalized in a number of ways, for instance by theories which relate relative end-weight to the acceptability of different constituent orders. One class of theories suggests that Heavy NP Shift is driven by consideration of the parser, which has been argued to favor Heavy NP Shift when the shifted order reveals VP-internal constituent structure more quickly than the canonical order. This paper tests theories of Heavy NP Shift which link constituent order to constituent identification. While controlling for other factors known to affect Heavy NP Shift, our results indicate that increasing the weight of an object NP can not make a Heavy NP Shift construction more acceptable than a comparison sentence in the canonical order, regardless of constituent identification. In addition, we examine acceptability as it relates to verb disposition, which has been linked to constituent identification in prior literature. We find no significant differences between verbs with different subcategorizations, contrary to the findings of corpus studies. We therefore reveal a disconnect between production and comprehension, and we conclude that parser sensitivity to constituent structure is unlikely to affect speaker production of Heavy NP Shift.

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 that favors linear orders which allow identification of constituent structure (e.g. within a VP) to be recognized as quickly as possible, thereby explaining the length effect.

To illustrate this principle, consider (1), from Hawkins (1994, p. 57), which illustrates canonical English word order in (1a) and the Heavy NP Shift (HNPS, henceforth) word order in (1b). According to Hawkins, (1b) is preferred to (1a) because all of the 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 sub-constituents in VP to be encountered. Hawkins (1994) formalizes this as the *Early Immediate Constituents* principle (p. 77), 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, p. 77) for full formalization.)

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

b. I $_{VP}$[gave $_{PP}$[to Mary] $_{NP}$[the valuable book that was extremely difficult to find]]

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 paper to the memory of Samuel D. 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).

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 Heavy NP Shift may be related to speech planning, not parsing. Wasow argues that the interests of speech production are in conflict with parsing. Under this view, parsers (echoing Hawkins' analysis) are taken to prefer knowledge of VP-internal constituent structure as early as possible, what Wasow terms 'early commitment.' Alternatively, 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 paper, 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 versus 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 only lowers acceptability 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]

[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, pp.66-68) explains the long-before-short preference for head final languages by computing the 'window' for a CRD as soon as the first *head* (not word) of the first sub-constituent 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.

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 and 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. As discussed above, Wasow (1997) shows that optionally transitive verbs are more likely to occur in HNPS constructions than those which are obligatorily transitive. In a related finding, Stallings et al. (1998) show that verbs which can take either NP or S complements are more acceptable in HNPS constructions as compared to verbs which 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 et al. 2006).

The characterization of heaviness is made further precise in 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 also 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

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

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 non-idioms, 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 et al. (2004) show similar semantic effects on constituent placement in related constructions.

More recently, and based on 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. US versus 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 et al. 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, such as 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 subject differences (e.g. some subjects may just give lower ratings overall or alter their ratings differently as a function of NP length).

We recruited subjects via Amazon MTurk, which has been independently validated for linguistics research by Sprouse (2011) and Munro et al. (2010). We restricted our survey to only Amazon MTurk workers with IP addresses in the United States. Further, participants needed to have completed at least 50 tasks within Mechanical Turk with a 95 percent approval rating for all prior tasks in order to take this experiment. 209 subjects started the survey with 196 completing it. Participants gave informed consent via the survey prior to seeing instructions for the experiment. All subjects 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 which predict HNPS acceptability. We created 10 lists, each based upon one NP-and-S selecting verb. Each of the 10 lists had 20 associated sentences (based upon a single verb), with 10 non-shifted/shifted pairs, with 200 test sentences total. All of the test sentences appear in the appendix of this paper.

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

test items plus 30 filters) to subjects. 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 et al. 1998). We also limit all potential interveners to two-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 subject also saw introductory and concluding sections; the introductory section contained two additional fillers in fixed order (i.e. all subjects saw the same two introductory fillers) and the concluding section contained one additional fixed filler (subjects were not aware that these introductory and concluding fillers were not part of the randomized survey). Subjects therefore rated 43 sentences in total: the 10 test items and 30 fillers in random

presentation, as well as the two fillers presented in fixed order at the beginning of the survey and one fixed filler at end of the survey. The introductory section also queried subjects about native language status.

Finally, we identified four non-test items, one unquestionably acceptable and three unquestionably unacceptable, which we then used to monitor subjects' responses. We did this because it is known that some subjects in this type of survey experiment will answer randomly. Any subject who did not assign comparatively low or high values to these pre-selected fillers were discarded from the analysis.

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

209 subjects took the survey. Of these, 16 were excluded from the analysis. Of the 16 subjects whose data were excluded, 13 did not finish. While all subjects reported affirmatively that they spoke "English at home while growing up," one subject reported 'other' for region of origin and was therefore excluded. The final two subjects who were excluded appeared to answer the survey questions randomly, based upon their performance on the pre-selected filler items. After excluding these 16 subjects, data from 193 subjects 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 max-

[4]Data files including the original data collected (which includes the 16 excluded subjects) as well as a file including only the data used in the analysis are available on the first author's website.

imal 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 also 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 will be reported using Satterthwaite's approximations as implemented in the R package `lmerTest` (Kuznetsova et al. 2013).

A linear mixed effects model in which acceptability score was the dependent variable and word length and whether the item was shifted or non-shifted are included as predictors (fixed effects). There are significant main effects both of word length ($\beta = 0.034; t = 3.365; p < 0.001$) and of shiftedness ($\beta = 1.118; t = 12.541; p < 0.001$). Model comparison motivates inclusion of an interaction term which is significant ($\beta = -0.118; t = -11.515; p < 0.001$). This interaction is visible in Figure 1, which plots mean scores across all participants for shifted and non-shifted 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 non-shifted sentences when NP length varies from 2 words to seven words, with the non-shifted (canonical) word order more acceptable. Once NP length exceeds 7 words, no significant differences are found between the shifted and non-shifted word orders, with both orders having an average acceptability rating above 5 for all but one item type (shifted sentences with 8-word-long NPs). 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

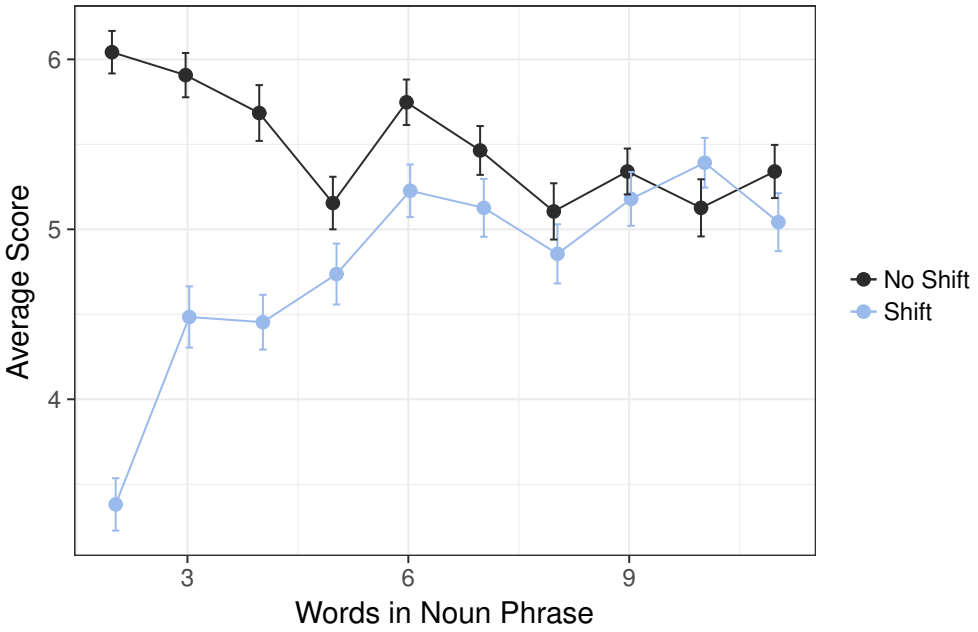

Figure 1: *Acceptability by Length (Words): Each vertical pair of dots represents NP word length, from 2 to 11 words*

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 ($\beta = 1.001; t = 12.294; p < 0.001$), a ceiling effect was again observed – visible as an interaction between syllable length and shiftedness ($\beta = -.062; t = -10.795; p < 0.001$). At no point were shifted items significantly more acceptable than non-shifted items, as indicated in Figure 2. Significant differences between shifted and non-shifted items obtain until 12 syllables, beyond which point both shifted and non-shifted 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 ($p < .001$).

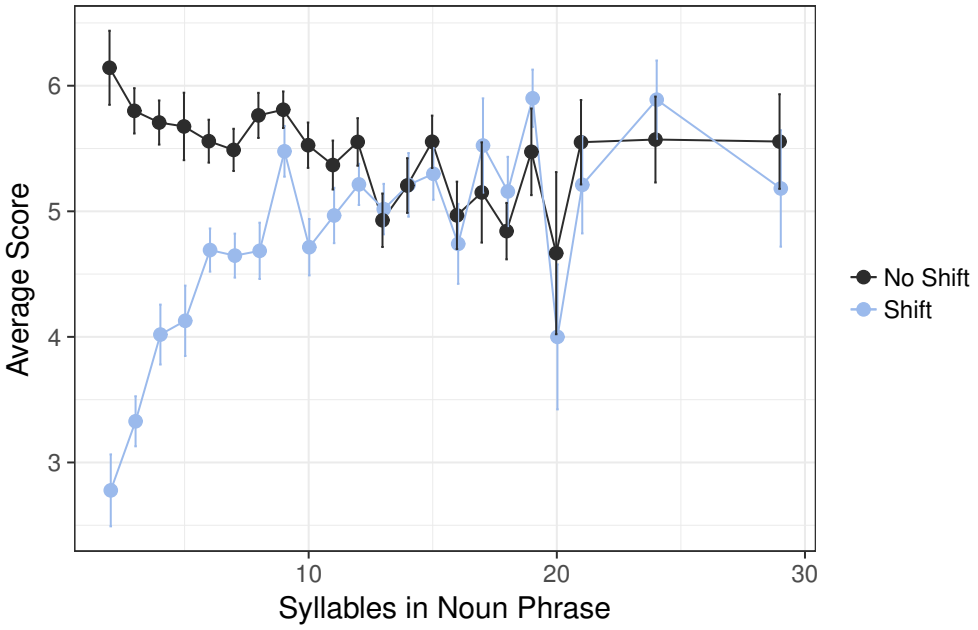

Figure 2: *Acceptability by Length (Syllables)*

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 which are optionally transitive as compared to those which are obligatority transitive. Limiting his sample to verbs which occurred in HNPS and/or with both NP and PP sisters a specified number of times, Wasow (1997, pp.97-100) reports significantly more occurrences of HNPS for optionally transitive verbs in the Brown ($p < .02$) and Switchboard ($p < .01$) corpus. 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 which drives HNPS rates in corpora.

In order to determine if our subjects found HNPS more acceptable with obligatorily transitive verbs, we coded our verbs as V_t if obligatorily transitive and 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 each verb as follows.[5]

(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 non-shifted condition. These descriptive statistics are visualized in Figure 3, in which the overall mean for the verb in the non-shifted 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 non-shifted condition versus the shifted condition). For example, 'explain' is hardly affected at all, though it is readily used intransitively. Alternatively, 'recommend' and 'suggest' have a much lower score in the shifted condition than the non-shifted 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

[5]Classifying these verbs is non-trivial. 'Indicate' has an intransitive use in the UK (= 'to turn signal') and 'mention' has an intransitive entry in the OED ('to mention about something', with the most recent example in 1925), but no other dictionary 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).

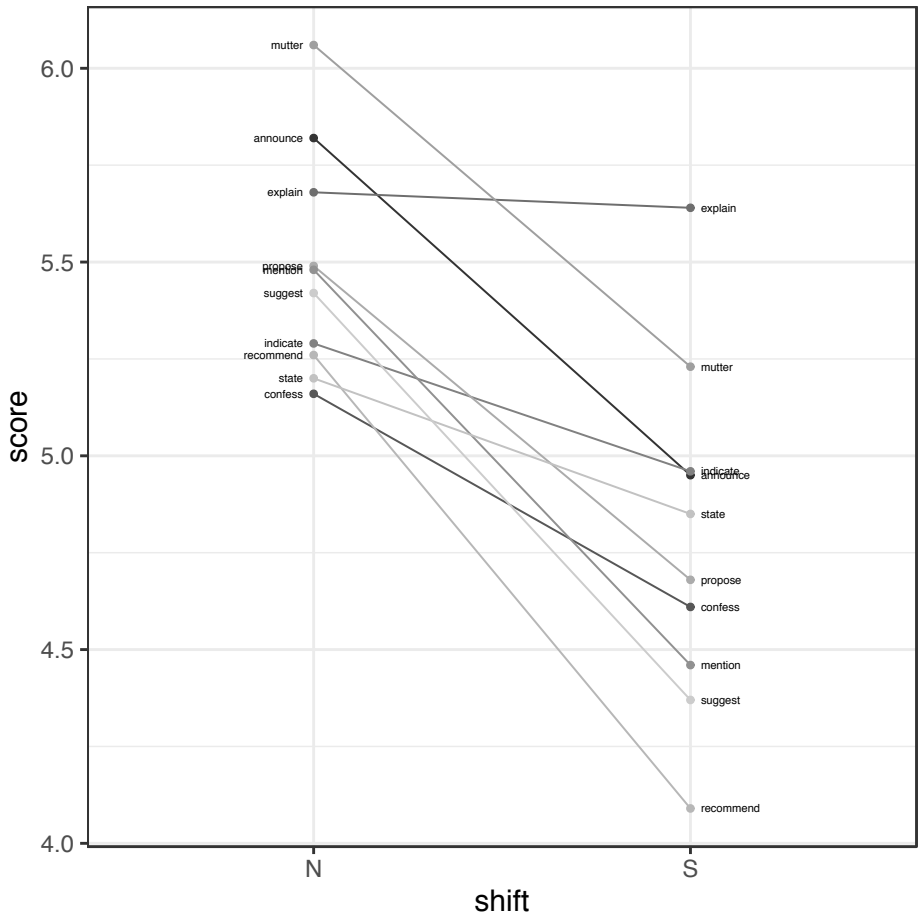

Figure 3: *Acceptability by Verb in Non-Shifted and Shifted Conditions*

variable and transitivity status is 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 subjects did not have a statistically significant preference for HNPS with either verb disposition, with a non-significant trend towards preferring optionally transitive verbs ($\beta = 0.4683$, $t = 1.876$, p = .0972), as illustrated in Figure 4 (error bars show Standard Error).

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 non-significant trend in the opposite direction, aligning our subjects (who represent speech

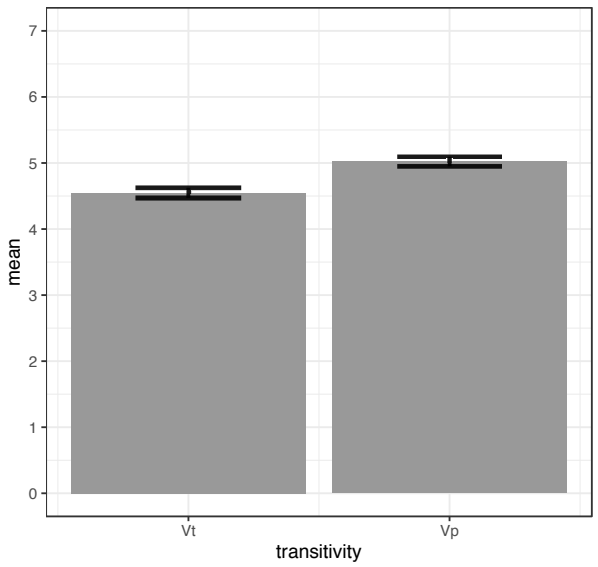

Figure 4: *Acceptability by Verb Disposition, with Error Bars Representing Standard Error*

comprehension) with the corpus data (speech production) in this respect.[6]

4 Discussion

The results described above fail to support theories of HNPS which relate use of HNPS to consideration of the parser's ability to identify subconstituents within VP. Within Hawkins's theory (Hawkins 1994, 2004), HNPS is increasingly favored as the argument NP's length increases relative to an intervener between the argument and selecting verb. However, our subjects did not rate HNPS sentences higher as the NP length increased relative to the PP interveners past 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 seven point scale) in our results, regardless of NP length.

In addition, our subjects did not prefer HNPS for obligatorily transitive verbs, contrary

[6]Given the verbs means visualized in Figure 3, excluding verbs which 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 non-controversially V_p, while the lowest rated, 'recommend' and 'suggest,' are non-controversially V_t.

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 which influences the occurrence of HNPS.

We also note a disconnect between corpus and experimental studies given our results. While our subjects are sensitive to end-weight up to the ceiling effect, increasing NP length appears to only bring HNPS 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. Alternatively, it could be that the alternative, largely discourse or syntactic (via NP complexity) factors which 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 expected based on corpus research.

[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 et al. (2005) and Polinsky et al. (2013) for examples and review.

References

Arnold, Jennifer E, Anthony Losongco, Thomas Wasow, and Ryan Ginstrom. 2000. Heaviness vs. newness: The effects of structural 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* 68:255278.

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: Springer.

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*. Cambridge, MA: 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. Caroline K. Duncan-Rose and Theo Vennemann, 325–39. 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*. URL `http://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 Jeffrey. 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 *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 *Proceedings of the 30th annual meeting on 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. Jon Sprouse and Norbert Hornstein, 341–359. Cambridge: Cambridge University Press.

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

R Core Team. 2016. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `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, 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: Center for the Study of Language and Information.

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.

Appendix: Experimental Items

All experimental items in the Heavy NP Shift construction are given below, representing half of all of the experimental items. These items represent the HNPS construction because, in each case, a prepositional phrase (which we have placed in brackets for the purpose of this appendix) intervenes between the verb and its object. The other half of the items involve the same sentences which appear below, but with the prepositional phrase following the objects. For example, the first item below is *Alex indicated to Curtis the problems*, which is in the HNPS construction, and the non-HNPS item (not given below) is *Alex indicated the problems to Curtis.* Therefore, while 100 items appear in this appendix, the experiment had 200 items in total.

(A1) a. Alex indicated [to Curtis] the problems.

b. Alex indicated [to Curtis] the plan's problems.

c. Alex indicated [to Curtis] the most glaring problems.

d. Alex indicated [to Curtis] some problems with the plan.

e. Alex indicated [to Curtis] some problems in the hospital's plan.

f. Alex indicated [to Curtis] some weak points in the hospital's plan.

g. Alex indicated [to Curtis] some weak points in the hospital's strategic plan.

h. Alex indicated [to Curtis] each of the weak points in the hospital's plan.

i. Alex indicated [to Curtis] each of the weak points in the hospital's strategic plan.

j. Alex indicated [to Curtis] each of the weak points in the hospital's new strategic plan.

(A2) a. Kim proposed [to George] some policies.

b. Kim proposed [to George] some policy changes.

c. Kim proposed [to George] some major policy changes.

d. Kim proposed [to George] some changes to company policy.

e. Kim proposed [to George] some changes to the company policy.

f. Kim proposed [to George] some major changes to the company's policies.

g. Kim proposed [to George] some changes to the company's customer service policies.

h. Kim proposed [to George] some major changes to the company's customer service policies.

i. Kim proposed [to George] some significant changes to the company's inadequate customer service policies.

j. Kim proposed [to George] some significant changes to the company's notoriously inadequate customer service policies.

(A3) a. Todd suggested [to Joel] a gift.

b. Todd suggested [to Joel] a gift basket.

c. Todd suggested [to Joel] a large gift basket.

d. Todd suggested [to Joel] a gift basket of chocolates.

e. Todd suggested [to Joel] a gift of chocolates and flowers.

f. Todd suggested [to Joel] a gift of chocolates, flowers, and candies.

g. Todd suggested [to Joel] a gift basket filled with chocolates and flowers.

h. Todd suggested [to Joel] a gift basket filled with chocolates, flowers, and candies.

i. Todd suggested [to Joel] a large gift basket filled with chocolates, flowers, and candies.

j. Todd suggested [to Joel] a large gift basket filled with assorted chocolates, flowers, and candies.

(A4) a. Jacob recommended [to Jerry] some medicine.

b. Jacob recommended [to Jerry] lots of rest.

c. Jacob recommended [to Jerry] rest and clear fluids.

d. Jacob recommended [to Jerry] lots of rest and water.

e. Jacob recommended [to Jerry] lots of rest and clear fluids.

f. Jacob recommended [to Jerry] rest and clear fluids like apple juice.

g. Jacob recommended [to Jerry] rest and clear fluids like water and juice.

h. Jacob recommended [to Jerry] lots of rest and clear fluids like apple juice.

i. Jacob recommended [to Jerry] lots of rest and clear fluids like water and juice.

j. Jacob recommended [to Jerry] lots of rest and clear fluids like water and apple juice.

(A5) a. Matthew confessed [to Kenny] the story.

b. Matthew confessed [to Kenny] the scandalous story.

c. Matthew confessed [to Kenny] the story about himself.

d. Matthew confessed [to Kenny] the story about his boss.

e. Matthew confessed [to Kenny] the story about his secret relationship.

f. Matthew confessed [to Kenny] the story about his secret romantic relationship.

g. Matthew confessed [to Kenny] the story about his relationship with his boss.

h. Matthew confessed [to Kenny] the scandalous story about his relationship with his boss.

i. Matthew confessed [to Kenny] the scandalous story about his secret relationship with his boss.

j. Matthew confessed [to Kenny] the scandalous story about his secret romantic relationship with his boss.

(A6) a. Tim stated [to Kelly] the guidelines.

b. Tim stated [to Kelly] the strict guidelines.

c. Tim stated [to Kelly] the overly strict guidelines.

d. Tim stated [to Kelly] the guidelines for successful writing.

e. Tim stated [to Kelly] the guidelines for writing science papers.

f. Tim stated [to Kelly] the guidelines for writing successful science papers.

g. Tim stated [to Kelly] the guidelines for writing papers in the sciences.

h. Tim stated [to Kelly] the guidelines for writing successful papers in the sciences.

i. Tim stated [to Kelly] the guidelines for writing successful papers in the social sciences.

j. Tim stated [to Kelly] the guidelines for writing successful research papers in the social sciences.

(A7) a. Bob announced [to Ian] the earnings.

b. Bob announced [to Ian] the yearly earnings.

c. Bob announced [to Ian] the company's yearly earnings.

d. Bob announced [to Ian] the earnings from last year.

e. Bob announced [to Ian] the company's earnings from last year.

f. Bob announced [to Ian] the earnings from the last fiscal year.

g. Bob announced [to Ian] the company's earnings from the last fiscal year.

h. Bob announced [to Ian] the earnings from the first half of the year.

i. Bob announced [to Ian] the earnings from the first half of the fiscal year.

j. Bob announced [to Ian] the company's earnings from the first half of the fiscal year.

(A8) a. Angie muttered [to Jen] some comments.

b. Angie muttered [to Jen] some sarcastic comments.

c. Angie muttered [to Jen] a few sarcastic comments.

d. Angie muttered [to Jen] some comments about the movie.

e. Angie muttered [to Jen] some sarcastic comments about the movie.

f. Angie muttered [to Jen] some sarcastic comments about the romantic comedy.

g. Angie muttered [to Jen] some sarcastic and critical comments about the movie.

h. Angie muttered [to Jen] some sarcastic and critical comments about the romantic comedy.

i. Angie muttered [to Jen] some sarcastic and critical comments about the cheesy romantic comedy.

j. Angie muttered [to Jen] some bitingly sarcastic and critical comments about the cheesy romantic comedy.

(A9) a. Erica explained [to Joan] some information.

b. Erica explained [to Joan] some useful information.

c. Erica explained [to Joan] some information about computers.

d. Erica explained [to Joan] some useful information about computers.

e. Erica explained [to Joan] some tips and tricks about computers.

f. Erica explained [to Joan] some useful information about how computers work.

g. Erica explained [to Joan] some tips and tricks about how computers work.

h. Erica explained [to Joan] some useful tips and tricks about how computers work.

i. Erica explained [to Joan] some surprisingly useful tips and tricks about how computers work.

j. Erica explained [to Joan] some surprisingly useful tips and tricks about how laptop computers work.

(A10) a. Mary mentioned [to Frank] the ad.

b. Mary mentioned [to Frank] the real estate.

c. Mary mentioned [to Frank] the real estate listings.

d. Mary mentioned [to Frank] the local real estate listings.

e. Mary mentioned [to Frank] the listing for a new house.

f. Mary mentioned [to Frank] the listing for a large brick house.

g. Mary mentioned [to Frank] the listing for a house on lakefront property.

h. Mary mentioned [to Frank] the real estate listing for a large brick house.

i. Mary mentioned [to Frank] the real estate listing for a house on lakefront property.

j. Mary mentioned [to Frank] the real estate listing for an expensive house on lakefront property.