The reliability of acceptability judgments across languages*

Tal Linzen and Yohei Oseki

November 17, 2015

1 Introduction

Acceptability judgments are a major source of data in theoretical linguistics. Most of the contrasts reported in the literature reflect the judgment of one individual, typically the author of the article. The reliability of such judgments has repeatedly come under criticism (Edelman & Christiansen, 2003; Gibson & Fedorenko, 2010; Gibson, Piantadosi, & Fedorenko, 2013; Langendoen, Kalish-Landon, & Dore, 1973; Schütze, 1996). Wasow and Arnold (2005), for example, argue that "the journals are full of papers containing highly questionable data, as readers can verify simply by perusing the examples in nearly any syntax article about a familiar language." (p. 1484). If this criticism turned out to be correct, decades of syntactic theory would appear to be standing on shaky empirical ground.[1]

Other authors have defended the field's reliance on individual judgments (Phillips, 2010; Phillips & Lasnik, 2003; Featherston, 2009). Proponents of this methodology have pointed out that while acceptability judgments reported in linguistics articles are initially made by one or two authors, they are subjected to several stages of formal and informal

*We thank Alec Marantz for feedback and for teaching the seminar *Linguistics as Cognitive Science*, which led to this project. We also thank Jeremy Kuhn for inspiring the idea of a crowdsourced acceptability judgment database.

[1] Although some of the critics have targeted generative syntax in particular for opprobrium, this issue applies to other traditions in syntax; see, for example, the debate around the use of introspective judgments in The Cambridge Grammar of the English Language (Huddleston & Pullum, 2002a, 2002b).

peer review before being published, and as such are likely to be robust. This defence of individual judgments is bolstered by the results of two recent judgment collection experiments, which replicated the overwhelming majority of English judgments in a Minimalist syntax textbook and in *Linguistic Inquiry* articles (Sprouse & Almeida, 2012; Sprouse, Schütze, & Almeida, 2013).

The debate around the reliability of acceptability judgments has so far been limited to English. While English is the source of a sizable proportion of the judgments in the literature — it accounts for about half of the syntactic acceptability judgments in *Linguistic Inquiry* between 2001 and 2010 (Sprouse et al., 2013) — theoretical developments in generative linguistics are often driven by data from other languages. The formal peer review mechanisms that are currently in place may not be as effective outside of English. Journal articles that include data in a less widely spoken language are often reviewed by at most one or two native speakers of that language. Much of the work in linguistics, including book chapters, dissertations and conference proceedings papers, does not undergo formal peer review at all. Most worryingly, the extent of informal peer review that judgments in languages other than English are subjected to is often fairly limited: a native speaker of Estonian working on her own language in the United States may never have to present her work to any other Estonian speakers.

Is the current level of peer review sufficient to guarantee the validity of judgments in less widely spoken languages? We address this question by investigating the reliability of acceptability judgments in Hebrew and Japanese. This paper is organized as follows. In Section 2, we briefly describe the methods of our experiments in which a large sample of naive participants were asked to rate questionable acceptability judgments selected from the literature. In Section 3, we show that half of the Hebrew contrasts and a third of the Japanese contrasts that we deemed to be questionable failed to replicate in formal experiments. In Section 4, we discuss the interpretation of these results and suggests ways in

which the benefits of informal peer review can be extended beyond English. Section 5 concludes the paper.

2 Methods

Previous judgment replication experiments have aimed to estimate the proportion of English acceptability contrasts that can be replicated in a formal experiment (Sprouse & Almeida, 2012; Sprouse et al., 2013). The proportion of replicable contrasts is always defined with respect to a particular set of constrasts (the "population", in statistical terms). Should the set of contrasts include every single judgment in a certain body of work or only a subset of the judgments (see also Gibson et al., 2013, p. 231)? We illustrate this dilemma using the classification of acceptability judgments proposed by Marantz (2005).

The first category of judgments discussed by Marantz, which we refer to as Type I judgments, consists of "word salads" — sequences of words that are so far from the grammar of the language that they cannot even be assigned a phonological representation. The following "word salad", for example, illustrates what English sentences would look like if English were a head-final language like Japanese (Marantz, 2005, p. 433):

(1) *Man the book a woman those to given has.

The second category (Type II) includes judgments that illustrate uncontroversial facts about the grammar of the language, facts of the sort that might be presupposed in theoretical analyses. The following contrast, for example, shows that English verbs agree in number with their subject (Marantz, 2005, p. 434):

(2) a. The men are leaving.

b. *The men is leaving.

The third category (Type III) includes more subtle contrasts, such as constraints on *wh*-movement or on possible coreference relations across noun phrases (what Gibson et al. (2013) refer to as "theoretically meaningful contrasts"). The judgments that critics take issue with typically fall into this category.

Sprouse and Almeida's work on English took the set of all published judgments as their relevant population, regardless of their category. One of the contrasts from Adger (2003) replicated by Sprouse and Almeida (2012) is shown in (3):

(3) a. The bear snuffled.

b. *The bear snuffleds.

This contrast was unsurprisingly replicated by a large margin, as were other Type II judgments.

Since we believe that Type I/II judgments in any language are extremely likely to replicate, our study focuses on Type III judgments. Unfortunately, there is no clear cut boundary between Type II and Type III judgments. For example, the contrast in (4), which illustrates the Coordinate Structure Constraint on *wh*-movement, seems to be as self-evident as many Type II judgments — few would be willing to bet on a non-replication of this contrast (Hickok, 2010) — but it patterns with many Type III judgments in that it illustrates a constraint on *wh*-movement:

(4) a. Who did you see Mary with?

b. *Who did you see Mary and?

For the experiments reported below, the authors — linguists who are native speakers of Hebrew or Japanese — selected contrasts in each language that they believed were potentially questionable Type III judgments. A total of 18 contrasts were selected in each language, 14 of which were controversial Type III contrasts from the literature (henceforth "critical

items") and four were uncontroversial Type II contrasts (henceforth "control items"). The full list of materials is given in Appendix A (for Hebrew) and Appendix B (for Japanese).

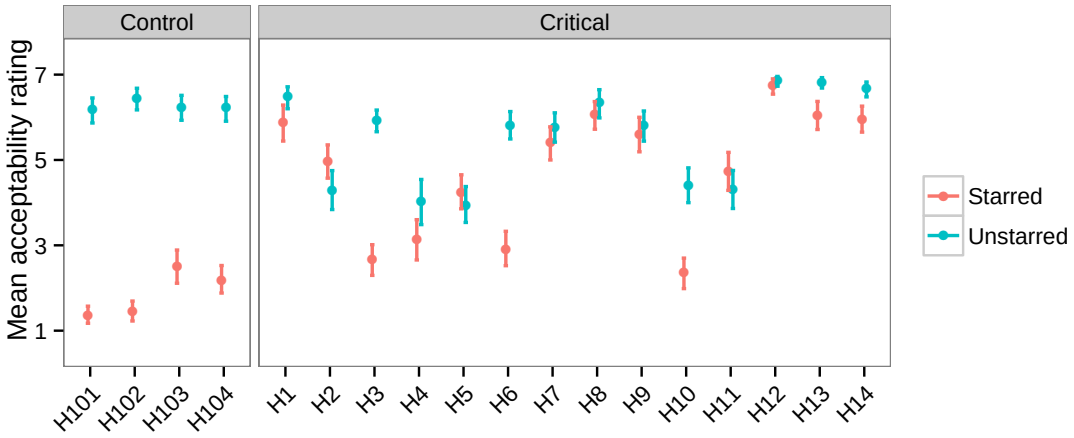
Two acceptability rating experiments were then conducted, one in Hebrew and one in Japanese. The Hebrew experiment was completed by 76 participants, and the Japanese experiment by 98 participants. All participants were recruited through Facebook. The experiments were administered using a website created for this purpose. The instructions were based on those used by Sprouse and Almeida (2012). The participants were requested to rate each sentence on a scale from 1 (very bad) to 7 (very good). All participants rated all sentences, but the order of sentences was randomized across participants. See Appendix C for additional details on the experimental methods.

3 Results

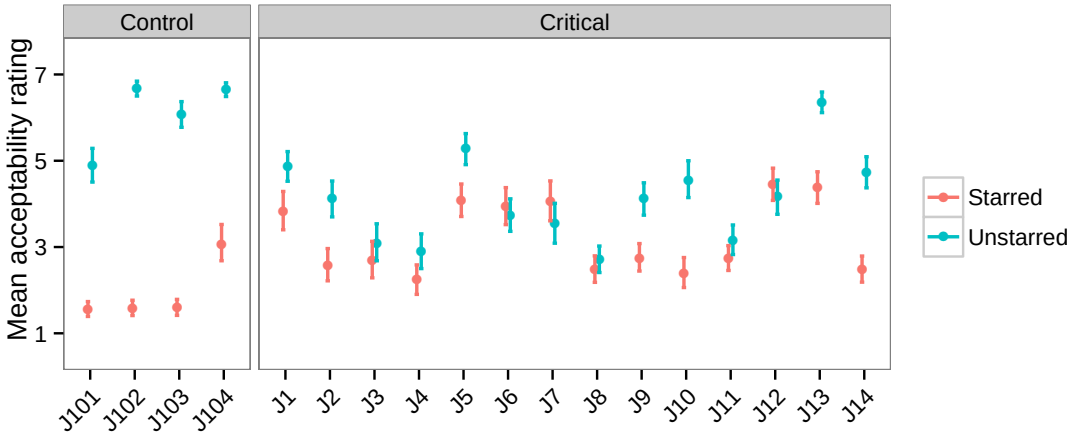
Figures 1a and 1b show the mean acceptability ratings for each of the judgments in Hebrew and Japanese, respectively. To assess the statistical significance of the results, we conducted a two-tailed paired t-test for each contrast separately (see Sprouse et al. (2013) for a discussion of analysis methods for this paradigm). The full numerical results are reported in Table 1 for Hebrew and Table 2 for Japanese.

Contrast	Starred	Unstarred	Diff	sd(Diff)	t	p	Sig.
H1	6.49	5.89	0.60	1.57	3.19	0.002	**
H2	4.29	4.97	-0.68	1.82	-3.06	0.003	**!
H3	5.93	2.66	3.27	1.74	15.82	< 0.001	**
H4	4.03	3.13	0.90	2.37	3.17	0.002	**
H5	3.94	4.25	-0.30	1.84	-1.37	0.174	!
H6	5.81	2.91	2.90	1.93	12.27	< 0.001	**
H7	5.76	5.40	0.36	1.81	1.62	0.109	
H8	6.35	6.06	0.29	1.57	1.55	0.126	
H9	5.81	5.60	0.21	1.50	1.13	0.262	
H10	4.41	2.36	2.06	1.90	9.05	< 0.001	**
H11	4.32	4.74	-0.42	1.76	-2.07	0.042	*!
H12	6.86	6.74	0.11	0.67	1.42	0.159	
H13	6.81	6.06	0.76	1.26	5.04	< 0.001	**
H14	6.68	5.96	0.72	1.50	4.00	< 0.001	**
H101	6.19	1.36	4.83	1.58	26.45	< 0.001	**
H102	6.44	1.45	4.99	1.54	28.08	< 0.001	**
H103	6.24	2.50	3.74	2.09	15.17	< 0.001	**
H104	6.22	2.18	4.04	2.07	16.97	< 0.001	**

Table 1: Hebrew results: The columns Unstarred and Starred indicate the mean rating for the unstarred and starred sentence in each contrast, respectively; Diff and sd(Diff) are the mean and standard deviation of the difference in rating between the unstarred and starred sentences. The rightmost column has one star if $0.01 \leq p < 0.05$ and two stars if $p < 0.01$. An exclamation mark indicates that the sign of the difference between the means is negative — i.e., the mean rating for the unstarred version was lower than the mean rating for the starred one.



(a)



(b)

Figure 1: Results of the experiments: (a) Hebrew and (b) Japanese. Error bars represent bootstrapped 95% confidence intervals.

Contrast	Starred	Unstarred	Diff	sd(Diff)	t	p	Sig.
J1	4.86	3.84	1.02	1.98	4.63	< 0.001	**
J2	4.12	2.58	1.54	2.26	6.22	< 0.001	**
J3	3.09	2.70	0.38	1.54	2.38	0.019	*
J4	2.89	2.24	0.65	1.65	3.55	< 0.001	**
J5	5.28	4.09	1.19	1.92	5.86	< 0.001	**
J6	3.74	3.95	-0.21	1.62	-1.13	0.263	!
J7	3.54	4.06	-0.52	2.85	-1.62	0.109	!
J8	2.70	2.49	0.22	1.73	1.17	0.245	
J9	4.12	2.74	1.39	1.88	6.90	< 0.001	**
J10	4.55	2.39	2.16	2.17	8.99	< 0.001	**
J11	3.16	2.74	0.42	1.34	3.02	0.003	**
J12	4.18	4.45	-0.26	1.86	-1.32	0.189	!
J13	6.36	4.38	1.98	2.01	9.14	< 0.001	**
J14	4.73	2.48	2.26	2.12	9.86	< 0.001	**
J101	4.90	1.55	3.34	2.02	15.47	< 0.001	**
J102	6.69	1.59	5.10	1.20	40.33	< 0.001	**
J103	6.08	1.60	4.48	1.62	26.15	< 0.001	**
J104	6.66	3.07	3.59	2.07	16.31	< 0.001	**

Table 2: Japanese results. See the caption of Table 1 for the interpretation of the columns.

Control: The control contrasts in both languages were robustly replicated. The unstarred sentences within each contrast were rated 5 or higher on average, whereas the starred sentences were rated 3 or lower on average (for all differences, $t > 15$, $p < 0.001$).

Critical—Hebrew: Seven of the 14 contrasts were replicated at the conventional statistical threshold of $p < 0.05$. Two contrasts showed a significant difference in the opposite direction than expected: the starred sentence was rated higher than the unstarred one (H2: $p = 0.003$; H11: $p = 0.04$). The difference in ratings within the remaining five contrasts failed to reach significance. The sign of the difference in four of those contrasts was consistent with the originally reported judgments. This suggests that a larger sample size may result in a higher replication rate, though it should be kept in mind that the sample was already quite large ($n = 76$). Based on the variability of the responses, we estimate that the experiment was sensitive enough on average to detect a difference of 0.55 in ratings (see Appendix D for details).

Critical—Japanese: Ten of the 14 contrasts were replicated at the $p < 0.05$ level. The remaining four contrasts did not reach significance in either direction; the numerical difference in three of these contrasts went in the opposite direction than predicted. The higher proportion of replicated Japanese contrasts was not due to the larger sample of participants — the sensitivity of the Japanese experiment was almost identical to the Hebrew experiment, with an average detectable difference of 0.54 — but rather to somewhat larger effect sizes: the average difference in ratings between unstarred and starred sentence in Japanese was 0.87 compared to 0.77 in Hebrew (see Appendix D for details).

Variability across contrasts: Strictly speaking, only differences in rating within each contrast are relevant to the replicability debate. Still, there was striking variability in ratings across contrasts. For instance, while contrast H8 was replicated ($p < 0.001$), its starred version received an average rating of 6.06 — higher than the rating of nine of the 14 unstarred critical sentences in the Hebrew experiment. At the same time, while contrast J4 was replicated ($p < 0.001$), its unstarred version was rated 2.89 on average — lower than six of the 14 starred critical sentences in the Japanese experiment. This pattern of results illustrates the care that should be taken in relating acceptability to grammaticality; clearly, it makes little sense to interpret a mean acceptability rating of 6.06 as showing that a sentence cannot be grammatical if a mean acceptability rating of 2.89 is taken to show that a sentence is grammatical.

4 Discussion

Half of the Hebrew contrasts and a third of the Japanese contrasts that were identified by the authors as potentially controversial did not replicate in a formal experiment. Three of the controversial contrasts in each experiment were rated in the opposite direction to the originally reported judgments: the starred sentences received higher ratings than their unstarred counterparts. At the same time, contrasts that the authors judged to be clear Type II contrasts (control items) were consistently replicated by a comfortable margin.

Our results reinforce the concern that some Type III contrasts in the literature may not be replicable, and suggest that replicability issues may be more common in languages other than English. Gibson et al. (2013) propose a fairly extreme approach for addressing this concern: they argue that every acceptability judgment must be validated in a formal experiment. However, given that the robustness of Type II contrasts is obvious to any native speaker of the language, it would be a waste of resources to test each and every judg-

ment in a formal experiment (Culicover & Jackendoff, 2010; Poeppel, 2010), especially in smaller language communities where a large sample of participants is difficult to recruit. We suggest that linguists concerned with data quality should focus on the small minority of potentially questionable Type III contrasts. Formal acceptability rating experiments are necessary only in few cases in which there is considerable disagreement among linguists about a particular Type III judgment. The results of the experiments presented in this paper indicate that individual linguists can identify such contrasts with considerable accuracy. This validates the intuition that informal peer review can effectively weed out such judgments from the literature (Phillips, 2010).

While thorough peer review is essential for guaranteeing the quality of acceptability judgments without formal experiments, our results also imply that peer review is less extensive in languages like Hebrew and Japanese. In order to clarify the reason, it is instructive to divide the review mechanisms that Phillips (2010) discusses into three stages. The first stage is pre-publication peer review, which takes place mainly in conferences. As we have pointed out, pre-publication peer review is likely to be less rigorous in languages other than English: most conference are likely to have few, if any, native speakers of the language in question, with the exception of conferences that focus on particular language families.

The second stage is the formal review that takes place as part of the journal publication process. Many papers do not undergo this process at all (e.g., book chapters, dissertations, and conference proceeding papers). Questionable judgments can slip even into those papers that are formally peer-reviewed; indeed, some of the judgments that failed to replicate in our experiments were drawn from journal articles. This issue is likely to be more acute in less widely spoken languages, where journal editors often struggle to find reviewers who are both native speakers of the language and experts on the topic of the submitted article. This situation is further exacerbated when the paper isn't predominately about a particular language but includes one or two judgments from several languages.

The third stage in the process outlined by Phillips can be referred to as "historical peer review". He argues that questionable judgments do not make it into the "lore" of the discipline: they are ignored by subsequent researchers. Yet it is unclear how well this process works in practice for those who do not speak the language in question and are incapable of evaluating the judgments. For example, contrast H8 in the Hebrew experiment has been challenged by Siloni (2001, fn. 15), but this fact does not seem to have undermined the influence of the analysis motivated in part by that contrast (Shlonsky, 2004). It is unclear whether the field is aware of Siloni's challenge to the validity of the contrast: a later paper on Welsh cites the Hebrew contrast without noting the disagreement about its status (Willis, 2006).

The weaknesses of the peer review process for less widely spoken languages can be straightforwardly mitigated. We propose an online crowdsourced database of published acceptability judgments, modeled after existing community resources such as Stack Overflow and Urban Dictionary. To help linguists who are not experts on a particular language to discover existing post-publication criticisms of published judgments in that language, links between different papers that discuss a given judgment will be automatically generated. Users will be given the option to comment on judgments online. Such comments might specify the set of contexts in which the judgment is valid, or provide attested examples that challenge it. Furthermore, a voting mechanism would allow users to quickly evaluate a judgment without commenting on it. Such "upvotes" and "downvotes" have been successful in weeding out uninformed answers to questions on Stack Overflow or fraudulent definitions in Urban Dictionary. The website could also provide facilities for collecting judgments from a large sample of naive participants, in a handful of cases in which this will be found to be necessary.

Some of the issues with peer review processes as currently implemented apply to widely studied languages as well: a questionable English judgment that has made it into a pub-

lished paper may mislead linguists who are not native English speakers and are not aware of the controversy surrounding the judgment. We therefore believe that work on English will also benefit from the online crowdsourced database sketched above.

5 Conclusion

The vast majority of published English judgments can be replicated with naive participants (Sprouse & Almeida, 2012; Sprouse et al., 2013). We argued that this is due to two reasons. First, a large proportion of the acceptability judgments illustrate obvious and uncontroversial contrasts (Type I/II judgments). Second, more subtle contrasts (Type III judgments) are informally vetted by a large community of linguists who are native English speakers. While not foolproof, this informal peer review process weeds out most questionable judgments (Phillips, 2010).

To examine the efficacy of the peer review process in languages other than English, we selected acceptability judgments in Hebrew and Japanese that we deemed to be questionable. A half (in Hebrew) or a third (in Japanese) of the Type III judgments failed to replicate, while all Type II judgments were robustly replicated. These results suggest that (1) formal acceptability rating experiments are not necessary for each and every judgment, (2) linguists can effectively identify questionable contrasts, and (3) informal peer review mechanisms are less effective for languages spoken by a smaller number of linguists. We proposed an online community resource that can extend the benefits of informal peer review to less widely spoken languages.

A List of Hebrew grammaticality contrasts

The judgments were primarily drawn from peer-reviewed articles, in particular from the Special Hebrew Issue of *Natural Language and Linguistic Theory* (August 1995). In addition, two books were included: a recent collection of articles (Armon-Lotem, Danon, & Rothstein, 2008) and a frequently cited dissertation published as a book (Shlonsky, 1997). Other judgments were taken from papers published in various issues of *Natural Language and Linguistic Theory* and *Linguistic Inquiry*.

Some of the Hebrew judgments were for DPs rather than for entire sentences, such as the following judgment (Belletti & Shlonsky, 1995, p. 517):

(5) a. ha-haxzara shel ha-shtaxim la-falastinim
the-handing.over of the-territories to.the-Palestinians

b. *ha-haxzara la-falastinim shel ha-shtaxim
the-handing.over to.the-Palestinians of the-territories

These DPs were presented to participants embedded in simple sentence (see Supplementary Materials). The added material is represented below in brackets.

The original articles have used a variety of romanization schemes; here we use a unified scheme that reflects modern pronounciation (*x* represents the voiceless velar fricative [x]). We kept the glosses used in the original articles even when our own judgments about the meaning of some words diverged from the original ones. The gloss ACC refers to the accusative marker *et*.

A.1 Critical items

(6) H1 (Arad, 2003):

a. Dani tsava et ha-kirot be-laka.
Dani painted ACC the-walls in-varnish
Dani painted the walls with varnish.

b. *Dani siyed et ha-kirot be-laka.
Davi whitewashed ACC the-walls in-varnish
Dani whitewashed the walls with varnish.

(7) H2 (Shlonsky, 1992):

a. elu ha-sfarim she-Dan tiyek otam bli likro otam.
these the-books that-Dan filed them without to-read them
'These are the books that Dan filed without reading.'

b. *elu ha-sfarim se-Dan tiyek otam bli likro.
these the-books that-Dan filed them without to-read
'These are the books that Dan filed without reading.'

(8) H3 (Belletti & Shlonsky, 1995):

a. [hitvakaxnu al] ha-haxzara shel ha-shtaxim la-falastinim.
[we.argued about] the-return of the-territories to.the-Palestinians
'We argued about the return of the territories to the Palestinians.'

b. [hitvakaxnu al] ha-haxzara la-falastinim shel ha-shtaxim.
[we.argued about] the-return to.the-Palestinians of the-territores
'We argued about the return to the Palestinians of the territories.'

(9) H4 (Shlonsky, 1990):

a. et mi lo yadata im ha-xayalim atsru?
ACC who no you.knew if the-soldiers detained
'Who didn't you know whether the soldiers detained?'

b. *mi lo yadata im ne'etsar al-yedei ha-xayalim?
who no you.knew if was.detained by the-soldiers
'Who didn't you know whether [he] was arrested by the soldiers?'

(10) H5 (Borer, 1995):

a. Dani muxan haya ba-zman ve-Rina muxana hayta gam.
Dani ready was on-time and-Rina ready was too
'Dani was ready on time and Rina was too.'

b. *Dani muxan haya ba-zman ve-Rina hayta gam.
Dani ready was on-time and-Rina was too
'Dani was ready on time and Rina was too.'

(11) H6 (Borer, 1995):

a. ha-ne'arot shalxu kulan mixtavey mexa'a la-memshala.
the-women sent all letters protest to.the-government
'The women all sent protest letters to the government.'

b. ??eize mixtavim ha-ne'arot shalxu kulan la-memshala?
which letters the-women sent all to.the-government
'Which letters did the women all sent to the government?'

(12) H7 (Borer, 1995):

a. ratsinu lish'ol eifo ha-ma'avak ha-axaron ihiye.
we.wanted to.ask where the-struggle the-last will.be
'We wanted to ask where the last struggle will be.'

b. *eifo ha-ma'avak ha-axaron ihiye?
where the-struggle the-last will.be
'Where will the last struggle be?'

(13) H8 (Shlonsky, 2004):

a. [ra'iti] para shveitsarit xuma.
[I.saw] cow Swiss brown
'I saw a brown Swiss cow.'

b. *[ra'iti] para xuma shveitsarit.
[I.saw] cow brown Swiss
'I saw a Swiss brown cow.'

(14) H9 (Shlonsky, 2004):

a. [hu lakax et] ha-shulxan ha-shaxor ha-arox [sheli].
[he took ACC] the-table the-black the-long my
'He took my long black table.'

b. *[hu lakax et] ha-shulxan ha-arox ha-shaxor [sheli].
[he took ACC] the-table the-long the-black [my].
'He took my black long table.'

(15) H10 (Siloni, 1996):

a. [shamanu al] ha-harisa shel ha-tsava et ha-ir.
[we.heard about] the-destruction of the-army ACC the-city
'We heard about the army's destruction of the city.'

b. *[shamanu al] ha-harisa et ha-ir shel ha-tsava.
[we.heard about] the-destruction ACC the-city of the-army
'We heard about the army's descruction of the city.'

(16) H11 (Siloni, 1995):

a. ad ha-shana she-avra kol ha-klavim ha-noshxim et ba'aley-hem hayu mumatim.
until the-year last all the-dogs the-biting ACC owners-their were killed
'Until last year, all the dogs biting their owners were killed.'

b. *ad ha-shana she-avra kol ha-klaim she-noshxim et ba'aley-hem hayu mumatim.
until the-year last all the-dots that-bit ACC owners-their were killed
'Until last year, all of the dogs who bit their owners were killed.'

(17) H12 (Preminger, 2009):

a. Dan masar et ha-ma'atafa la-mefakeax.
Dan handed ACC the-envelope to.the-supervisor
'Dan handed the envelope to the supervisor.'

b. *Dan masar et ha-ma'atafa.
Dan handed ACC the-envelope
'Dan handed the envelope.'

(18) H13 (Shlonsky, 1997):

a. ha-sfarim ne'elmu me-ha-sifriya.
the-books disappeared from-the-library
'The books disappeared from the library.'

b. *ne'elmu ha-sfarim me-ha-sifriya.
disappeared the-books from-the-library
'The books disappeared from the library.'

(19) H14 (Botwinik-Rotem, 2008):

a. ha-sefer kal li-kri'a ve-le-nituax.
the-book easy to-reading and-to-analyzing
b. *ha-sefer kal li-kri'a ve-nituax.
the-book easy to-reading and-analyzing
'The book is easy to read and to analyze.'

A.2 Control items

(20) H101 (anaphora binding):

a. im-o shel ha-tinok ra'ata oto.
mother-his of the-baby saw him
'The baby's mother saw him.'

b. *im-o shel ha-tinok ra'ata et atsmo.
mother-his of the-baby saw himself
'The baby's mother saw himself.'

(21) H102 (number agreement):

a. naflu le-Dani ha-maftexot.
fell.3PL to-Dani the-keys
b. *ha-maftexot nafal le-Dani.
the-keys fell.3SG to-Dani
'Dani dropped his keys.'

(22) H103 (the relativizer *ha* can only be used directly before the present participle, whereas the relativizer *she* can be used anywhere in the sentence):

a. hine ha-ish she-lo xoshev al kesef.
here the-man who-not think about money
b. *hine ha-ish ha-lo xoshev al kesef.
here the-man the-not think about money
'Here is the man who doesn't think about money.'

(23) H104 (resumptive pronoun in subject relative clause):

a. ze ha-ish she-ohev le-daber al politika.
this the-man who-likes to-talk about politics
b. *ze ha-ish she-hu ohev le-daber al politika.
this the-man who-he likes to-talk about politics
'This is the man who likes to talk about politics.'

B List of Japanese grammaticality contrasts

The judgments were selected from peer-reviewed papers published in various issues of *Natural Language and Linguistic Theory*, *Linguistic Inquiry*, and *Journal of East Asian Linguistics*. In addition, one book and three dissertations were included: an influential dissertation published as a book (Miyagawa, 1989) and frequently cited dissertations (Farmer, 1980; Hoji, 1985; Oku, 1998). Other judgments were taken from some Japanese-specific journals, but data points reported there are of theoretical importance.

Some of the Japanese judgments were bound to particular semantic interpretations (e.g. scope interpretations). In those cases in which acceptability of the sentences has to be evaluated given a particular interpretation, the explicit contexts were given to the participants,

and they were asked to rate the sentences under those contexts.

B.1 Critical items

(24) J1 (Miyagawa, 1989):

a. Kuruma-ga dorobou-ni ni-dai nusum-are-ta.
car-Nom thief-by two-CL steal-Pass-Past
'Two cars were stolen by a thief.'

b. *Kodomo-ga geragerato san-nin warat-ta.
children-Nom loudly three-CL laugh-Past
'Three children laughed loudly.'

(25) J2 (Kishimoto, 2001):

a. Taro-wa nani-o kai-mo si-nakat-ta.
Taro-Top anything-Acc buy-Q do-Neg-Past
'Taro did not buy anything.'

b. *Dare-ga warai-mo si-nakat-ta.
anyone-Nom laugh-Q do-Neg-Past
'Anyone did not laugh.'

(26) J3 (Miyagawa, 2001):

a. Sono tesuto-o zen'in-ga uke-nakat-ta.
that exam-Acc all-Nom take-Neg-Past
'That exam, all did not take.' (not $>$ all)

b. *Zen'in-ga sono tesuto-o uke-nakat-ta.
all-Nom that exam-Acc take-Neg-Past
'All did not take that exam.' (not $>$ all)

(27) J4 (Saito, 1994):

a. Dare-ga naze nani-o kat-ta no?
who-Nom why what-Acc buy-Past Q
'Who bought what why?'

b. *John-ga naze nani-o kat-ta no?
John-Nom why what-Acc buy-Past Q
'What did John buy why?'

(28) J5 (Tada, 1992):

a. Taro-wa migime-dake-o tumur-e-ru.
Taro-Top right.eye-only-Acc close-can-Pres
'Taro can wink his right eye.' (can > only)

b. *Taro-wa migime-dake-ga tumur-e-ru.
Taro-Top right.eye-only-Nom close-can-Pres
'Taro can wink his right eye.' (can > only)

(29) J6 (Sakai, 1994):

a. Mary$_i$-no [kanozyo$_i$-ga kik-anakat-ta] hihan
Mary-Gen she-Nom hear-Neg-Past criticism
'Mary's criticism that she did not hear'

b. *Mary$_i$-no [kanozyo$_i$-no kik-anakat-ta] hihan
Mary-Gen she-Gen hear-Neg-Past criticism
'Mary's criticism that she did not hear'

(30) J7 (Oku, 1998):

a. Taro-wa zibun-no gakusei-o home-ta. Ziro-wa home-nakat-ta.
Taro-Top self-Gen student-Acc praise-Past Ziro-Top praise-Neg-Past
'Taro praised Taro's student. Ziro did not praise Ziro's student.'

b. *Taro-wa kuruma-o teineini arat-ta. Ziro-wa araw-anakat-ta.
Taro-Top car-Acc carefully wash-Past Ziro-Top wash-Neg-Past
'Taro washed cars carefully. Ziro did not wash cars carefully.'

(31) J8 (Hiraiwa, 2010):

a. Atama-o Ken-ga omoikkiri Naomi-o tatai-ta.
head-Acc Ken-Nom hard Naomi-Acc hit-Past
'On the head, Ken hit Naomi hard.'

b. *Ken-ga omoikkiri Naomi-o atama-o tatai-ta.
Ken-Nom hard Naomi-Acc head-Acc hit-Past
'Ken hit Naomi hard on the head.'

(32) J9 (Farmer, 1980):

a. Hanako-wa Taro-niyotte sono inu-o kaw-ase-rare-ta.
Hanako-Top Taro-by that dog-Acc buy-Caus-Pass-Past
'Hanako was made by Taro to buy that dog.'

b. *Sono inu-wa Taro-niyotte Hanako-ni kaw-ase-rare-ta.
that dog-Top Taro-by Hanako-Dat buy-Caus-Pass-Past
'That dog was made by Taro Hanako to buy.'

(33) J10 (Hoji, 1985):

a. Daremo-o dareka-ga aisiteiru.
everyone-Acc someone-Nom love
'Everyone, someone loves.' (every > some)

b. *Dareka-ga daremo-o aisiteiru.
someone-Nom everyone-Acc love
'Someone loves everyone.' (every > some)

(34) J11 (Miyagawa & Tsujioka, 2004):

a. Taro-wa Hanako-ni Tokyo-ni nimotu-o okut-ta.
Taro-Top Hanako-Dat Tokyo-Dat package-Acc send-Past
'Taro sent Hanako a package to Tokyo.'

b. *Taro-wa Tokyo-ni Hanako-ni nimotu-o okut-ta.
Taro-Top Tokyo-Dat Hanako-Dat package-Acc send-Past
'Taro sent Hanako a package to Tokyo.'

(35) J12 (Boeckx & Niinuma, 2004):

a. Hanako-ga Tanaka-sensei-ni Mary-o go-syookai-si-ta.
Hanako-Nom Prof.Tanaka-Dat Mary-Acc Hon-introduce-Hon-Past
'Hanako introduced Mary to Prof.Tanaka.'

b. *Hanako-ga Mary-ni Tanaka-sensei-o go-syookai-si-ta.
Hanako-Nom Mary-Dat Prof.Tanaka-Acc Hon-introduce-Hon-Past
'Hanako introduced Prof.Tanaka to Mary.'

(36) J13 (Saito, 1992):

a. Karera$_i$-ga otagai$_i$-o hihansi-ta.
they-Nom each.other-Acc criticize-Past
'They criticized each other.'

b. *Otagai$_i$-no sensei-ga karera$_i$-o hihansita.
each.other-Gen teacher-Nom them-Acc criticize-Past
'Each other's teachers criticized them.'

(37) J14 (Watanabe, 2006):

a. Roger-wa donburi-ni yon-hai-no gohan-o tabe-ta.
Roger-Top big.bowl-Dat four-CL-Gen rice-Acc eat-Past
'Roger ate four big bowls of rice.'

b. *Roger-wa yon-hai-no gohan-o donburi-ni tabe-ta.
Roger-Top four-CL-Gen rice-Acc big.bowl-Dat eat-Past
'Roger ate four big bowls of rice.'

B.2 Control items

(38) J101 (quantifier floating):

a. Sensei-ga san-nin sake-o non-da.
teacher-Nom three-CL sake-Acc drink-Past
'Three teachers drank sake.'

b. *Gakusei-ga katudon-o san-nin tabe-ta.
student-Nom pork.bowl-Acc three-CL eat-Past
'Three students ate pork bowl.'

(39) J102 (Nom-Gen conversion):

a. Kesa ame-ga hut-ta.
this.morning rain-Nom fall-Past
'It rained this morning.'

b. *Sakuban kaminari-no nat-ta.
last.night thunder-Gen fall-Past
'It thundered last night.'

(40) J103 (double Acc constraint):

a. Taro-wa Hanako-ni sono hon-o yom-ase-ta.
Taro-Top Hanako-Dat that book-Acc read-Caus-Past
'Taro made Hanako read that book.'

b. *Taro-wa Hanako-o sono ryori-o tsukur-ase-ta.
Taro-Top hanako-Acc that dish-Acc cook-Caus-Past
'Taro made Hanako cook that dish.'

(41) J104 (nominal structure):

a. Taro-wa takusan-no hon-o kat-ta.
Taro-Top many-Gen book-Acc buy-Past
'Taro bought many books.'

b. *Hanako-wa gengo takusan-o benkyosi-ta.
Hanako-Top language many-Acc study-Past
'Hanako studied many languages.'

C Detailed procedures

We asked participants not to participate in the study if they did not satisfy the following two conditions: (1) they lived in Israel / Japan in the first 13 years of their lives, except for short breaks; and (2) their parents spoke Hebrew / Japanese to them. It was emphasized that a grammatical sentence was not necessarily one that would be approved by official language institutions, but rather one that would not sound out of place when uttered by a native speaker in a conversation.

The stimuli were divided into two blocks; each block contained one of the sentences of each contrast. Participants were not made aware of this division. The assignment of sentences to blocks was counterbalanced across participants, and the order of sentences within each block was pseudo-randomized such that no more than three consecutive sentences had the same acceptability annotation, with the exception that the uncontroversial judgments were presented first in each block, to familiarize the participants with the task by using relatively easy judgments. An additional constraint on the order of presentation was that the first three sentences did not all have the same acceptability annotation.

The order of the sentences of each contrast was also counterbalanced across participants: for a given contrast, approximately half of the participants read the unstarred version of the contrast earlier than the starred one, and the other way around for the other half of the participants. This allocation was performed under the constraint that each block should contain an equal number of unstarred and starred sentences.

Participants in the Hebrew experiment also rated a few unpaired sentences for acceptability; these sentences appeared in a middle block, between the two blocks reserved for acceptability contrasts. The ratings of these sentences are not analyzed in the current paper.

D Additional analyses

D.1 Sensitivity of tests

We defined the sensitivity of our tests as the minimal mean difference in ratings between the unstarred and starred versions of a contrast, for which our power to detect the difference with a $p < 0.05$ threshold is at the standard level of 0.8 (calculated using the `power.t.test` function in R). We estimated the standard deviation of the difference in ratings by averaging the empirical standard deviations across all contrasts within a given

language; this estimated standard deviation was 1.7 for Hebrew and 1.88 for Japanese. The resulting sensitivity estimates were 0.55 and 0.54 respectively.

D.2 Reanalysis with a smaller sample size

The number of participants in our experiments was fairly large (around twice that of Sprouse and Almeida (2012), for example). Such samples are not always easy to obtain for less widely spoken languages. To determine whether the statistical signifance of our findings crucially depended on the large sample size, we repeated our analysis on an arbitrarily selected subset of 20 participants.

With a smaller sample size, only four of the Hebrew contrasts were replicated at the conventional statistical threshold of $p < 0.05$. None of the differences in the remaining ten contrasts reached significance; four out of these were negative, and the other six were positive. In Japanese, seven of the 14 contrasts were replicated, one (J6) showed a significant difference in the opposite direction than predicted, and the remaining six contrasts did not reach significance. The detailed results of the subset analysis are shown in Table 3 for Hebrew and in 4 for Japanese.

Contrast	Starred	Unstarred	Diff	sd(Diff)	t	p	Sig.
H1	6.58	6.16	0.42	1.77	1.03	0.315	
H2	5.31	5.50	-0.19	2.07	-0.36	0.723	!
H3	6.00	2.65	3.35	1.73	7.99	< 0.001	**
H4	3.84	3.21	0.63	2.41	1.14	0.268	
H5	3.47	3.89	-0.42	1.74	-1.05	0.306	!
H6	5.89	3.50	2.39	2.09	4.85	< 0.001	**
H7	6.11	6.39	-0.28	1.13	-1.05	0.311	!
H8	6.37	5.58	0.79	1.81	1.90	0.074	
H9	6.20	6.00	0.20	1.86	0.42	0.683	
H10	4.44	1.78	2.67	1.94	5.83	< 0.001	**
H11	4.11	4.58	-0.47	1.81	-1.14	0.268	!
H12	6.95	6.74	0.21	0.54	1.71	0.104	
H13	6.89	6.37	0.53	1.12	2.04	0.056	
H14	6.79	5.95	0.84	1.54	2.39	0.028	*
H101	6.47	1.37	5.11	1.29	17.30	< 0.001	**
H102	6.35	1.45	4.90	1.48	14.77	< 0.001	**
H103	6.10	2.15	3.95	1.85	9.55	< 0.001	**
H104	6.80	1.75	5.05	1.39	16.19	< 0.001	**

Table 3: Hebrew results: first 20 participants

Contrast	Starred	Unstarred	Diff	sd(Diff)	t	p	Sig.
J1	5.43	3.86	1.57	1.45	4.05	0.001	**
J2	4.19	2.31	1.88	1.89	3.96	0.001	**
J3	3.40	3.65	-0.25	0.97	-1.16	0.262	!
J4	3.19	1.88	1.31	1.30	4.03	0.001	**
J5	5.72	4.56	1.17	1.86	2.67	0.016	*
J6	3.17	4.67	-1.50	2.02	-2.57	0.026	*!
J7	3.50	3.50	0.00	3.51	0.00	1.000	
J8	2.84	2.68	0.16	2.24	0.31	0.762	
J9	3.70	3.05	0.65	1.81	1.60	0.126	
J10	4.85	2.90	1.95	2.01	4.33	< 0.001	**
J11	3.40	2.60	0.80	1.67	2.14	0.046	*
J12	4.42	4.21	0.21	1.81	0.51	0.619	
J13	6.50	4.45	2.05	2.04	4.50	< 0.001	**
J14	5.05	2.53	2.53	2.25	4.90	< 0.001	**
J101	4.71	1.41	3.29	2.34	5.81	< 0.001	**
J102	6.78	1.50	5.28	0.83	27.09	< 0.001	**
J103	6.00	1.67	4.33	1.41	13.00	< 0.001	**
J104	6.56	3.00	3.56	1.98	7.63	< 0.001	**

Table 4: Japanese results: first 20 participants

References

Adger, D. (2003). *Core Syntax*. Oxford: Oxford University Press.

Arad, M. (2003). Locality constraints on the interpretation of roots: The case of Hebrew denominal verbs. *Natural Language & Linguistic Theory*, *21*(4), 737–778.

Armon-Lotem, S., Danon, G., & Rothstein, S. (2008). *Current issues in generative Hebrew linguistics*. John Benjamins.

Belletti, A., & Shlonsky, U. (1995). The order of verbal complements: A comparative study. *Natural Language & Linguistic Theory*, *13*(3), 489–526.

Boeckx, C., & Niinuma, F. (2004). Conditions on Agreement in Japanese. *Natural Language & Linguistic Theory*, *22*(3), 453-480.

Borer, H. (1995). The ups and downs of Hebrew verb movement. *Natural Language & Linguistic Theory*, *13*(3), 527–606.

Botwinik-Rotem, I. (2008). Object gap constructions. In S. Armon-Lotem, G. Danon, & S. D. Rothstein (Eds.), *Current issues in generative Hebrew linguistics* (pp. 77–104). Amsterdam and Philadelphia: John Benjamins Publishing.

Culicover, P., & Jackendoff, R. (2010). Quantitative methods alone are not enough: Response to Gibson and Fedorenko. *Trends in Cognitive Sciences*, *14*(6), 234–235.

Edelman, S., & Christiansen, M. H. (2003). How seriously should we take minimalist syntax? *Trends in Cognitive Sciences*, *7*(2), 60–61.

Farmer, A. (1980). *On the interaction of morphology and syntax* (Unpublished doctoral dissertation). MIT.

Featherston, S. (2009). Relax, lean back, and be a linguist. *Zeitschrift für Sprachwissenschaft*, *28*(1), 127–132.

Gibson, E., & Fedorenko, E. (2010). Weak quantitative standards in linguistics research. *Trends in Cognitive Sciences*, *14*(6), 233–234.

Gibson, E., Piantadosi, S. T., & Fedorenko, E. (2013). Quantitative methods in syntax/semantics research: A response to Sprouse and Almeida (2013). *Language and Cognitive Processes*, *28*(3), 229–240.

Hickok, G. (2010). *Weak quantitative standards in linguistics research? The debate between Gibson/Fedorenko & Sprouse/Almeida.* http://www.talkingbrains.org/2010/06/weak-quantitative-standards-in.html.

Hiraiwa, K. (2010). Spelling-Out the Double-o Constraint. *Natural Language & Linguistic Theory*, *28*, 229-240.

Hoji, H. (1985). *Logical form constraints and configurational structures in Japanese* (Unpublished doctoral dissertation). University of Washington.

Huddleston, R., & Pullum, G. K. (2002a). *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Huddleston, R., & Pullum, G. K. (2002b). *A response concering The Cambridge Grammar.* http://linguistlist.org/issues/13/13-1932.html.

Kishimoto, H. (2001). Binding of Indeterminate Pronouns and Clause Structure in Japanese. *Linguistic Inquiry*, *32*, 597-633.

Langendoen, D. T., Kalish-Landon, N., & Dore, J. (1973). Dative questions: A study in the relation of acceptability to grammaticality of an English sentence type. *Cognition*, *2*(4), 451–478.

Marantz, A. (2005). Generative linguistics within the cognitive neuroscience of language. *The Linguistic Review*, *22*(2-4), 429–445.

Miyagawa, S. (1989). *Structure and Case Marking n Japanese*. New York: Academic.

Miyagawa, S. (2001). EPP, Scrambling, and Wh-in-situ. In M. Kenstowicz (Ed.), *Ken Hale: A Life in Language* (p. 293-338). Cambridge, MA.: MIT Press.

Miyagawa, S., & Tsujioka, T. (2004). Argument structure and ditransitive verbs in Japanese. *Journal of East Asian Linguistics*, *13*(1), 1-38.

Oku, S. (1998). *A theory of selection and reconstruction in the Minimalist perspective* (Unpublished doctoral dissertation). University of Connecticut.

Phillips, C. (2010). Should we impeach armchair linguists? *Japanese/Korean Linguistics*, *17*, 49–64.

Phillips, C., & Lasnik, H. (2003). Linguistics and empirical evidence: Reply to Edelman and Christiansen. *Trends in Cognitive Sciences*, *7*(2), 61–62.

Poeppel, D. (2010). *An egregious act of methodological imperialism.* http://www.talkingbrains.org/2010/06/egregious-act-of-methodological.html.

Preminger, O. (2009). Failure to agree is not a failure – φ-Agreement with post-verbal subjects in Hebrew. In J. van Craenenbroeck (Ed.), *Linguistic Variation Yearbook 2009* (pp. 241–278). John Benjamins Publishing Company.

Saito, M. (1992). Long distance scrambling in Japanese. *Journal of East Asian Linguistics*, *1*(1), 69–118.

Saito, M. (1994). Additional-*wh* effects and the adjunction site theory. *Journal of East Asian Linguistics*, *3*(3), 195–240.

Sakai, H. (1994). Complex NP Constraint and case conversion in Japanese. In M. Nakamura (Ed.), *Current topics in English and Japanese* (pp. 179–200). Tokyo: Hitsuji Shobo.

Schütze, C. T. (1996). *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. Chicago: University of Chicago Press.

Shlonsky, U. (1990). *Pro* in Hebrew subject inversion. *Linguistic Inquiry*, *21*(2), 263–275.

Shlonsky, U. (1992). Resumptive pronouns as a last resort. *Linguistic Inquiry*, *23*(3), 443–468.

Shlonsky, U. (1997). *Clause structure and word order in Hebrew and Arabic: An essay in comparative Semitic syntax*. New York and Oxford: Oxford University Press.

Shlonsky, U. (2004). The form of Semitic noun phrases. *Lingua*, *114*(12), 1465–1526.

Siloni, T. (1995). On participial relatives and complementizer D^0: A case study in Hebrew and French. *Natural Language & Linguistic Theory*, *13*(3), 445–487.

Siloni, T. (1996). Hebrew noun phrases: Generalized noun raising. *Parameters and functional heads: Essays on Comparative Syntax*, 239–267.

Siloni, T. (2001). Construct states at the PF interface. *Linguistic Variation Yearbook*, *1*, 229–266.

Sprouse, J., & Almeida, D. (2012). Assessing the reliability of textbook data in syntax: Adger's Core Syntax. *Journal of Linguistics*, *48*(3), 609–652.

Sprouse, J., Schütze, C., & Almeida, D. (2013). A comparison of informal and formal acceptability judgments using a random sample from Linguistic Inquiry 2001-2010. *Lingua*, *134*, 219–248.

Tada, H. (1992). Nominative objects in Japanese. *Journal of Japanese Linguistics*, *14*, 91–108.

Wasow, T., & Arnold, J. (2005). Intuitions in linguistic argumentation. *Lingua*, *115*(11), 1481–1496.

Watanabe, A. (2006). Functional projections of nominals in Japanese: Syntax of classifiers. *Natural Language & Linguistic Theory*, *24*(1), 241–306.

Willis, D. (2006). Against N-raising and NP-raising analyses of Welsh noun phrases. *Lingua*, *116*(11), 1807–1839.