Improving the reliability of syntactic judgments in less studied languages by systematizing the peer review process

Tal Linzen

May 5, 2011

1 Introduction

Introspective acceptability judgments are commonly used as data for linguistic theories. This source of data has always been controversial, but has come under particulary intense attack in recent years (Edelman and Christiansen 2003; Gibson and Fedorenko 2010a,b; Wasow and Arnold 2005 and others).

Several researchers have responded to the criticism with qualitative arguments (Huddleston and Pullum 2002a; Phillips 2009; Phillips and Lasnik 2003; Featherston 2009 and others). Recently, Sprouse and Almeida (submitted) joined the debate with a large-scale study that showed that nearly all of the judgments in a minimalist syntax textbook stood up to the quantitative standards advocated by the critics. They concluded that there is no reason to doubt the theoretical results achieved over the last few decades: the judgment methodology is sound, and in the few cases where it fails (if ever), peer review processes filter out the errors.

However, Sprouse and Almeida's conclusion can be argued to be valid only for English, where there are sociological reasons to believe that a questionable judgment would not stand the test of time. These reasons do not necessarily hold true for other languages.

This paper builds on Sprouse and Almeida's work and extends it to Hebrew. Since their results show that most syntactic intuitions are replicable, this study focused specifically on judgments that were judged by the author to be questionable. An on-line experiments found that half of these judgments (7 out of 14) failed to replicate, even with a relatively large sample of 76 subjects. One of the judgments that failed to replicate was in fact rated by subjects in the opposite direction than predicted by the original author. A reanalysis of a subset of the results with a smaller sample size of 20, as common in cognitive psychology practice, resulted

in 11 replication failures out of 14 contrasts.

This paper will argue that these results are worrisome, but at the same time point at the solution to the problem. Formal experiments as advocated by Gibson and Fedorenko (2010a) are unnecessary for nearly all of the cases, and their application to languages other than English face a number of technical obstacles overlooked in their paper. The major reason for the potential difference in reliability between Hebrew and English is not, of course, that Hebrew-speaking linguists are intellectually dishonest, but that their judgments are not as carefully reviewed by peers as the judgments reported by their English speaking colleagues. In a sense, the selection of the materials for the experiment involved a peer review process, and it proved quite effective. Hence, I propose an overhaul of the peer review process, which will extend to Hebrew (and other less widely spoken languages) the benefits that this process has for the quality of English judgments.

1.1 An attack on minimalism?

Most of the critics of judgment data focus their attention on the generative literature. An outsider to the field might get the impression that linguists of other persuasions do not use judgments to justify their theories. This is far from being the case: researchers who identify themselves as typologists, functional linguists or cognitive linguists make frequent and usually exclusive use of judgment data. A good example is the position expressed by Huddleston and Pullum (2002a) in their reply to a review of Huddleston and Pullum (2002b) criticizing the fact that they did not rely on "objective" data sources:

> . . . [the] range of sources was vast: the authors' lifelong experience of the English language; the similar experience possessed by a dozen other native-speaker collaborating authors; further evidence pointed out by others. . .

The reliability of judgments is therefore not only an issue for mainstream generative grammar, but for other traditions in the field as well. The interpretation of these judgments differs between traditions: cognitive and generative linguists generally see them as a window to the speaker's mind, whereas typologists and "traditional" linguists do not attach a great deal of importance to the way the speaker's knowledge is organized, instead viewing linguistics as a social science. The question whether acceptability judgments are reliable is orthogonal to this distinction: if the factual foundation that underlies both enterprises turns out to be is shaky, they both face a problem.

2 The experiment

2.1 Selection of experimental materials

Sprouse and Almeida conducted an exhautive investigation of all of the sentences in Adger (2003) that could be easily tested in a web experiment. Their experiment entailed obtaining judgments on thousands of sentence tokens, from hundreds of subjects. Since it is harder to recruit Hebrew speakers on line than English speakers (see Section 3), an online experiment on the same scale would not be feasible.

Moreover, Sprouse and Almeida's large scale study already established that an overwhelming majority of the judgments in the literature are robust (around 97%, depending on the statistical measure used), and simply duplicating their work in a different language seems to be a futile exercise. Not all judgments in the syntax literature are equally likely to be controversial, and hence in a follow-up experiment it might be more illuminating to focus on the tough cases. As an illustration, consider the judgments that Gibson and Fedorenko (2010a) take issue with:

(1) a. The man that the woman that the dog bit likes eats fish.

b. Mary wondered what who bought where.

c. What do you wonder who say?

These are complicated sentences. It is not a-priori clear how a naive participant might judge them. On the other end of the spectrum, syntax papers and books often include judgments as a rhetorical or pedagogical aid. In a sense the theory presented does depend on the truth of these judgments, but their truth is often taken for granted rather than introduced as a data point. For example, the following contrast (Adger 2003, p. 38, originally (81) in Section 2) is cited to demonstrate that the English past tense does not trigger person agreement:

(2) a. The bear snuffled

b. *The bear snuffleds

Unsurprisingly, Sprouse and Almeida replicated this contrast with an extremely large t value of 20.69 (p. 18). The following contrast in Adger's textbook, which illustrates that pronouns are specified for case, is equally dramatic (p. 222):

(3) a. He loves him.

b. *Him loves him.

c. *His loves him.

Finally, the following judgments are given as examples of one-place predicates (4a), two-place predicates (4b) and subcategorization frames (4c), respectively:

(4) a. Alison ran. (p. 62)

b. Alison kicked the cat. (p. 62)

c. *I ate that she was happy. (p. 69)

It would indeed be astonishing to find that naive participants judge these sentences differently from Adger. For Sprouse and Almeida's purposes, it made sense to use each and every judgment in Adger (2003). They were making a quantitative point, namely that *most* judgments are reliable, not that *all* judgments are reliable. While it is true that the trivial examples given above might inflate the replication rate in favor of their position in the debate, by using every single example found in the book, Sprouse and Almeida eliminated the risk of experimenter bias, which is a concern when a subjective criterion is used as in the present study.

The experiment presented here, however, had a different purpose, for which only the subjective method was be feasible.[1] The following method was used to select the stimuli: I reviewed the generative syntax literature on Hebrew, and select the judgments I disagreed with. The literature sample contained as many peer-reviewed papers as possible, primarily 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 et al. 2008) and a frequently cited dissertation published as a book (Shlonsky 1997). Other judgments were taken from papers published in various issues of NLLT and Linguistic Inquiry. It should be stressed that the selection of judgments studied in the present paper is in no way representative: the 14 contrasts and 11 individual-sentence judgments presented to the participants were hand-picked out of the thousands of judgments presented in the aforementioned papers, as a way of putting the reliability of Hebrew judgments to the most stringent test possible.

Some of the judgments were for DPs rather than for entire sentence, such as the following judgment

[1] In a future version of this study, this risk of experimenter bias could be reduced by conducting a pre-test to winnow out the uncontroversial cases.

from Belletti and Shlonsky (1995, p. 517):

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

b. *ha-haxzara la-falastinim šel ha-štaxim
the-handing.over to.the-Palestinians of the-territories

I judged both DPs to be equally grammatical. In order to simplify the experimental task, these DPs were presented embedded in a simple sentence, for example:

(6) hitvakaxnu al ha-haxzara šel ha-štaxim la-falastinim.
we.argued about the-handing.over of the-territories to.the-Palestinians.

2.2 Stimulus presentation

The questionnaire was administered using a web site created for this purpose. Subjects were recruited through Facebook. No personal information was collected about the subjects.

The instructions were a loose Hebrew translation of the those used by Sprouse. Participants were asked not to participate in the study if they did not satisfy the following two conditions: 1. they lived in Israel in the first 13 years of their lives, except for short breaks; and 2. their parents used Hebrew to speak to them. It was emphasized that a grammatical sentence was not necessarily one that would be approved by a composition professor or by the Academy of the Hebrew Language, but one that would not sound out of place when uttered by a native Hebrew speaker in a normal conversation. The participants were requested to rate each sentence on a scale from 1 (very bad) to 7 (perfectly fine). The intermediate steps on the scale were not labeled.

The set of stimuli consisted of three types of sentences: 14 paired controversial judgments that were originally presented as contrasts; 11 unpaired judgments, where no contrast was intended in the original paper (5 of them originally judged as grammatical, 4 as ungrammatical and 2 as very marginal); and 4 paired uncontroversial judgments from the literature.

The stimuli were randomized in three blocks: paired, unpaired and paired again. The participants were not made aware of this division. The first and third block contained one sentence each out of each pair of contrasted judgments. The allocation to blocks of each contrast was randomized between subjects – that is, the grammatical sentence of the contrast was the first one presented to approximately half of the subjects,

and the ungrammatical one was the first one presented to the other half of the subjects. This allocation was performed under the constraint that each block should contain an equal number of grammatical and ungrammatical sentences. In addition, the order of sentences within each block was randomized such that no more than three consecutive sentences were of the same grammaticality status, with the exception that the uncontroversial judgments were presented first in each block, to familiarirze the participants with the task using relatively easy stimuli. An additional constraint on the order of presentation was that the first three stimuli were not all of the same grammaticality status. The participants' performance on the uncontroversial judgments was also intended to be used to gauge how well they understood the task.

The unpaired judgments were presented in the second of the three blocks. Their order was randomized between participants, again under the constraint against more than three consecutive sentences with the same grammaticality status.

2.3 Results

During the 24 hours that the experiment was online, 184 user sessions were registered, out of which 76 participants completed the entire questionnaire. Only these 76 participants were entered into the analysis. A small number of trials were skipped; the skipped trials were taken out of the analysis, as well as their contrasted counterparts if applicable. Due to time constraints and methodological uncertainties about the statistical analysis of unpaired judgments, only the paired stimuli were analyzed. The mean results are shown in Figure 1.

The control contrasts, shown on the right side of the figure (groups 101 to 104) were robustly replicated. The grammatical sentences within each contrast were ranked higher than 6 (out of 7) on average, whereas the ungrammatical were ranked under 2.5. Most of the other contrasts went in the predicted direction, even though not with uniform robustness. There is considerable variation in the ranking of the ungrammatical sentences even in the replicated contrasts: in groups 23 and 24, for example, both sentences in the contrast were ranked as highly acceptable (6 or higher), but the grammatical sentence in each contrast was still consistently given higher rankings than the ungrammatical one. Three of the contrasts – 2, 7 and 19 – show effects in the opposite direction from the predicted effect.

Figure 2 shows the repeated measures t-statistic for each contrast. Out of the 14 controversial contrasts,

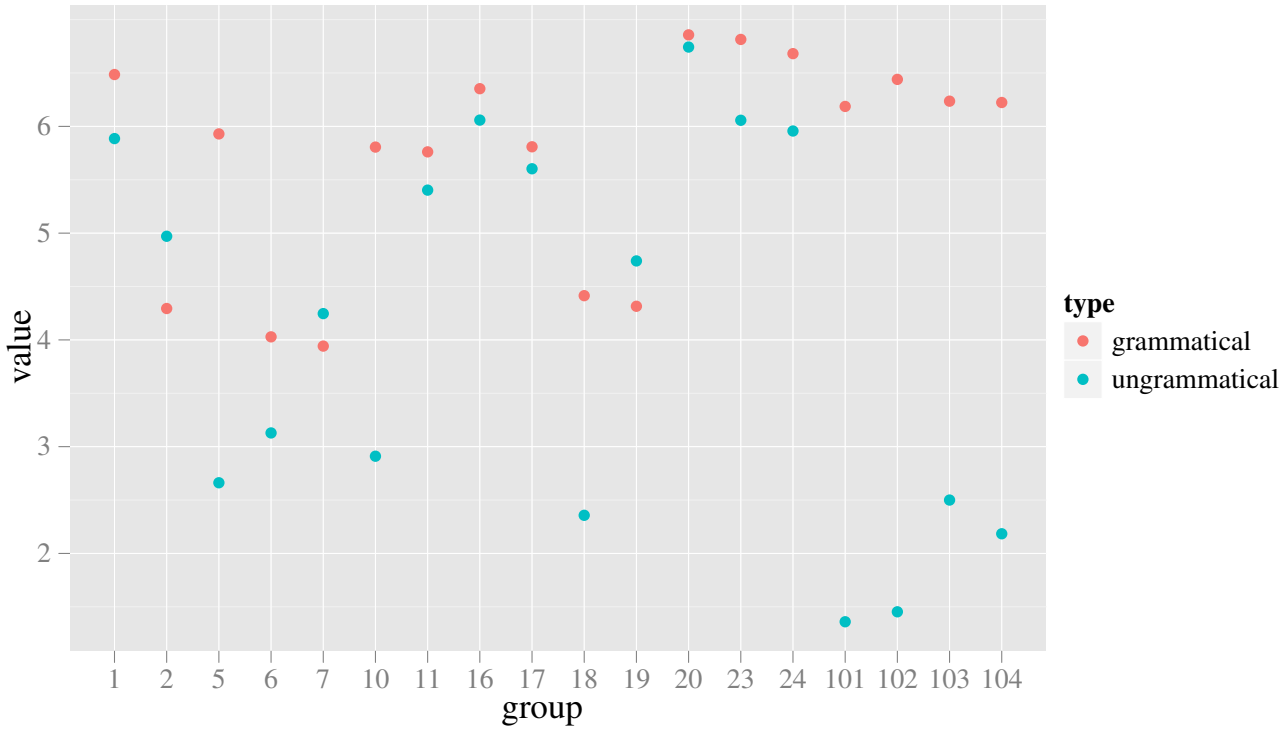

Figure 1: Mean ranking for each of the sentences in the contrasts tested in the experiment. The color of the dots reflects the judgment of the authors of the papers where these contrasts were presentes.

7 were replicated and 7 were not. Four trended in the predicted direction and three trended in the opposite direction, one of which reaching "arguable significance" (see below).

The statistical significance threshold is corrected for 19 comparisons (Bonferonni correction)[2]. The contrasts are in general independent, except for 5 and 18, which both concern the order of complements in DPs, though not the same complements; and 16 and 17, which both relate to the order of adjective within a DP, though not the same type of adjectives. In addition, it's not clear that the control contrasts should count as comparisons for this purpose. So corrections for 17 or 13 can also be justified. Incidentally, changing the number of comparisons from 19 to 13 makes the opposite-from-predicted effect in group 2 reach significance. For this reason I refer to this result as "arguably significant" in this paper.

[2]Sprouse and Almeida did not correct for multiple comparisons, even though they had a larger number of comparisons (104) than the present study. This inflates the number of replications that their study produced. For 9 out of 104 comparisons, the significance of the relevant mixed effect model coefficient was $p > 0.001$ (p. 7), or $p > 0.1$ after a Bonferonni correction. Even assuming that all of the comparisons whose p-value is marked $p < 0.001$ in the table were in fact significant at the uncorrected threshold $p < 0.0005$ as well, and would thus survive the correction, we get a replication rate of 91%. On the other hand, it is very

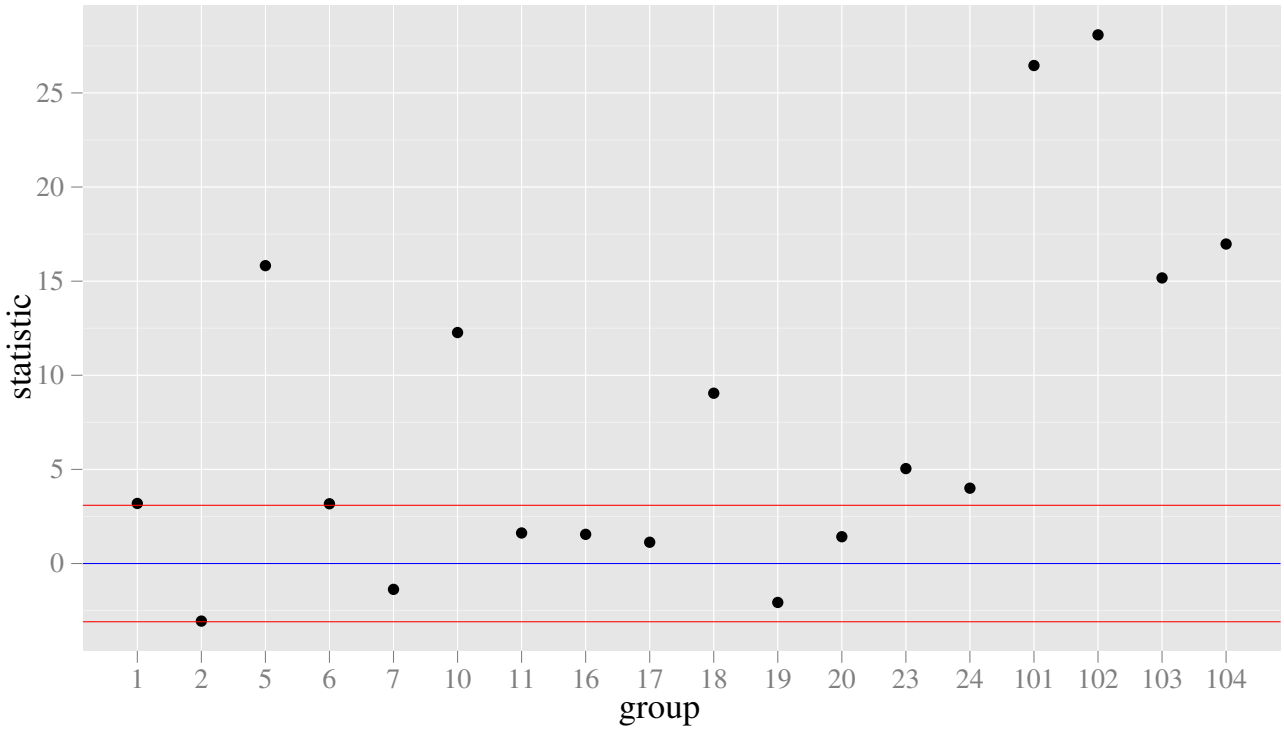

Figure 2: Two-sided t-statistics for the reliability of the difference in rankings between the two sentences of each contrast. The red lines show the significance threshold at the 5% level for a *t*-distribution with 75 degrees of freedom, corrected for 19 comparisons using the Bonferroni correction.

Even among the significant replications, the effect sizes are large only in three of the groups (5, 10 and 18). Figure 3 shows the t-statistics calculated based on the first 20 subjects, a common sample size in cognitive psychology. The three contrasts mentioned above are the only ones that survive the reduction in sample size. This means that with a modestly powered experiment 11 out of the 14 result would not be replicated.

3 On quantitative replications

The experiment described in section 2 shows that there is indeed a judgment reliability problem in the Hebrew syntax literature, at least for the handful of controversial judgments inspected in this paper. This

significant that none of their comparisons came out in the opposite direction from the one predicted. This means that while more sentences failed to replicate than Sprouse and Almeida admit, that is probably due to insufficient power, since on the whole there is a significant trend towards replication.

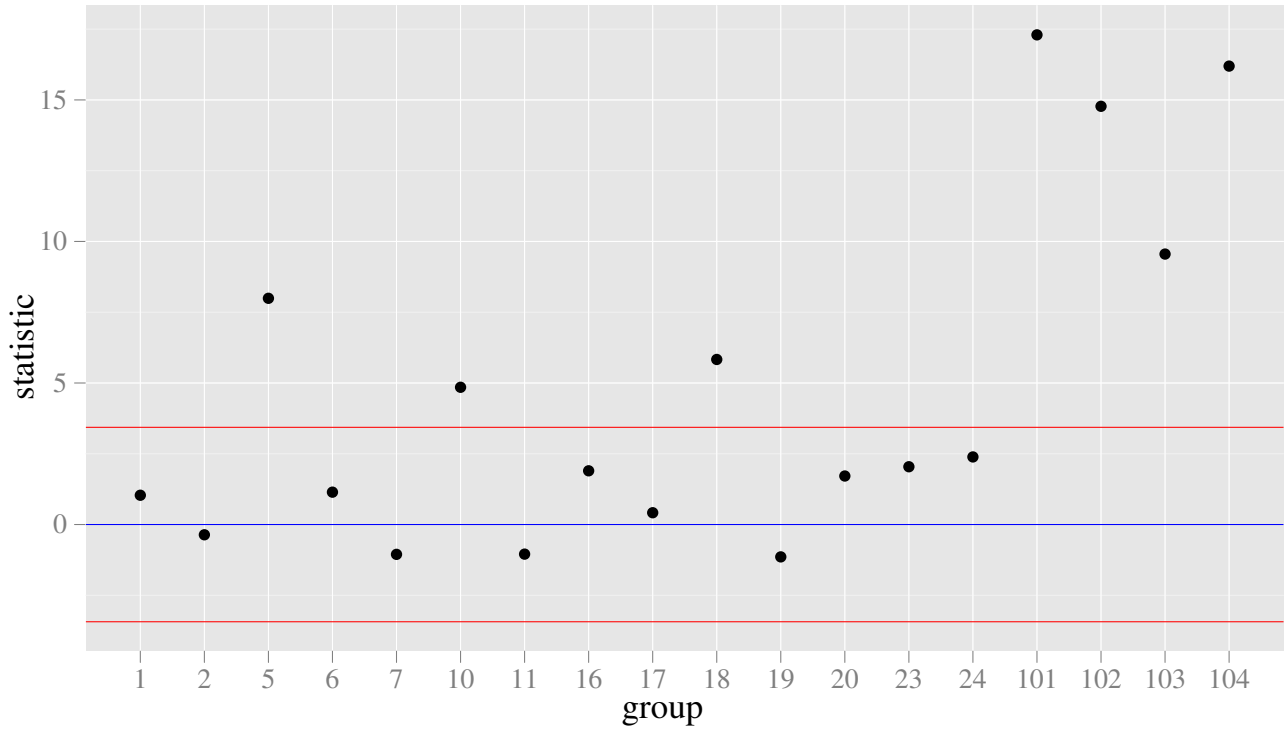

Figure 3: t-statistics in a smaller sample ($n = 20$)

may seem to vindicate Gibson and Fedorenko's (2010a) point.

Indeed, Gibson and Fedorenko argue that each and every acceptability contrast reported in a linguistics paper should be validated in a quantitative experiment: they "recommend gathering quantitative evidence for all empirical claims, because conducting experiments is so easy to do" (p. 17). There are two elements to this recommendation, a theoretical element, namely that every scientific claim must be quantitatively grounded, and a practical one, that naive subject judgments are easy to collect.

The first argument, while hard to reject in principle, is impossible to implement in practice. As Greg Hickok eloquently points out in his comment at his blog Talking Brains:[3]

> There are some things that simply don't need to be experimentally tested: that the word "plant" is ambiguous, that a Necker cube is bi-stable, that "Who did you see Mary and?" is ungrammatical whereas "Who did you see Mary with?" is perfectly fine.

[3] http://www.talkingbrains.org/2010/06/weak-quantitative-standards-in.html

> In the real world, with limited resources, we indeed need to make rational decisions about which aspects of our experiments to quantify...

More important for the present paper, however, is the fact conducting experiments gets less easy to do as soon as one leaves the world of English judgments. Gibson and Fedorenko recruit their subjects on Amazon Mechanical Turk. This platform currently can only pay workers from the US and from India[4], and an overwhelming majority of the workers come from those countries (Ross et al. 2010). Even if every country were equally represented on AMT, running experiments on medium-sized languages such as Hebrew, which have 50 times less speakers than English, would take a week or two. On smaller languages, especially where Internet access is less prevalent, this would simply be impossible.

In addition to recruiting participants, the syntactician will need to prepare experimental materials and instructions in the target language. If he is not a native speaker of the language he is researching, that will not be easy to do without a collaborator who speaks the language (a linguist or an experimental psychologist, rather than an informant). In addition, the materials will need to be matched for a reasonable selection of psycholinguistic variables, at the very least frequency – again, another technical hurdle for most non-European languages.

To sum up, for languages with a medium to small community of speakers, Culicover and Jackendoff's (2010) quote is even more pertinent than for English:

> It would cripple linguistic investigation if it were required that all judgments of ambiguity and grammaticality be subject to statistically rigorous experiments on naive subjects...

4 The judgment peer review process

Phillips's (2009) reply to the critics of intuitive judgments is that these judgments are subjected to several stages of formal and informal peer review, and hence are unlikely to be unreliable:

> If a key judgment is questionable, this is likely to be pointed out by a colleague, or by audience members in a talk, or reviewers of an abstract or journal article. If the questionable general-

[4] Yes, "you can elect to receive a check for your earnings in Indian Rupeers" (`https://www.mturk.com/mturk/help?helpPage=worker#how\_paid`). All other international Workers can only disburse to an Amazon.com gift certificate.

ization somehow makes it past that point, then it will still be subjected to widespread scrutiny before it becomes a part of linguistic lore.

The informal part of this process is clearly effective for work on English, especially when it is carried out in departments and conferences in American universities, where there are plenty of native English speakers and a questionable judgment would not go unchallenged. Indeed, this stage is crucial, since some of the most influential work in syntax is never published in peer-reviewed journals, but rather in books, dissertations, conference proceedings and informally circulated manuscripts. For languages less widely spoken than English, however, this informal process is less effective: a native speaker of Estonian working on his own language in an American department may never have to present his work to any other Estonian speaker.

The formal peer review as implemented today does not always solve the problem. Clearly, errors can slip through this process in any scientific field; however, the case of an unreliable judgments is different from, say, a misapplication of a statistical method in medical research. A reviewer does not need to suffer from cat allergies in order to evaluate the statistical methods in a paper testing a new treatment for cat allergies, but does need to be a native Chamorro speaker to be able to evaluate judgments on Chamorro sentences. The smaller a language community is, the less likely is an editor to find linguists speaking that language to review a paper. Even in the case of languages such as Hebrew, which are researched by a relatively vibrant community, a paper with dubious judgments can still make it to publication without ever being reviewed by a native Hebrew speaker – especially when the paper is not specifically about Hebrew, but contains Hebrew judgments.

The third bar that a judgment needs to cross, according to Phillips, is the *historical peer review* process (my term). I am not sure how well this stage works: once a questionable judgment has made it into the literature, it is not easy to eradicate it. For example, Shlonsky's (1997) judgment presented as example (8) below has been challenged in Siloni (2002). It is not clear how this fact might come to the attention of a linguist who is not a specialist on Hebrew, unless she is extremely diligent and takes the time to read all the 184 articles citing Shlonsky (1997) (according to Google Scholar). Another example is Landau's (1999) judgment that the Hebrew Possessive Dative does not carry the implication that the possessor was affected by the action described in the sentence. To me, this judgment is odd; indeed, Lee-Schoenfeld (2006) notes:

> ...the native speaker of Hebrew I asked to confirm Landau's judgments rejected all the exam-

ples that do not satisfy the affectedness condition. This discrepancy may be due to speaker variation or (as suggested by a reviewer) contextualization effects. Landau's judgments could be based on possible but pragmatically unusual situations. More native speakers need to be consulted to clarify this.

Landau himself has since changed his judgment (p.c.). However, his paper, which crucially relies on this judgment, remains very influential, and it is not clear how many of the authors of the 126 papers that cite it are aware of the inadequate empirical generalization that underlies it (though at least one of them, Pylkkänen (2002), seems skeptical).

In principle a paper containing a false generalization may be retracted, but in most if not all cases such an extreme step would be unjustified – it is unlikely that *all* of the judgments an analysis is based on are wrong. (My subjective impression, as well as the conclusion of Sprouse and Almeida, is that at most one judgment in twenty is even potentially contestable.)

5 Conclusion: the problem, and a sketch of a solution

This paper has shown that some of the judgments in the Hebrew syntax literature cannot be replicated even with a relatively large sample of participants. Even granting that the sentences selected for the experiment are far from a representative sample of Hebrew judgments, this is worrisome result. At the same time, it has been shown that these questionable judgments can be identified by a linguist (the author) with considerable reliability. This fact, in addition to Sprouse and Almeida's finding that virtually all English judgments that survived the "historical peer review" process are replicable given enough power, indicates that the way to improve the reliability of judgments in less studied language lies not in cumbersome formal experiments, but in the institutionalization of the peer review process in a way that remedies the differences in the size of the research community between English and those languages.

The following solution can be envisioned for languages such as Hebrew, which enjoy an active community of researchers working on them and have more than around a million speakers (most of whom have internet access). At the first stage, an online database of existing judgments will be created. At least for papers published in journals and available online, this should be an easy task to automate: judgments are

clearly set off from the rest of the text and marked with a small constant set of possible diacritics. Links between different papers that discuss a given judgment will be automatically generated, to facilitate the historical peer review procedure discussed in Section 4. A Wiki interface will be provided for adding comments to a judgment.[5] Such comments can specify the contexts where the judgment is valid, or present an attested corpus example which seems to contradict the judgment. It will be possible to leave anonymous comments, to reduce the change that a young linguist will be reluctant to disagree with a judgment made by a prominent figure in the field. In addition to text comments, a simple anonymous voting mechanism will be available, so that feedback is as easy to provide as possible. A similar system seems to work relatively well for community projects such as `www.urbandictionary.com`. This mechanism will enable linguist who is planning to use a judgment from the published literature to assess the judgment's reliability in a straighforward way. This should be particularly useful if this judgment is about a language he does not speak.

At the next stage, a peer review system for judgments will be put in place. The linguists on the site's review board will receive a weekly digest of judgments from recently submitted papers, and will vote on their validity. This should not take more a few minutes of work every week, assuming the number of linguists qualified to evaluate judgments in a given language is approximately proportional to the number of new papers published on that language every week. The same system may be used for collecting judgments during research as well, as an institutionalized version of the requests for typological data on Linguist List or in personal emails.

In addition to the peer review system, the proposed website will provide a platform for collecting acceptability judgments from naive subjects, for the handful of cases in which this will be found to be necessary – for example, if there is considerable disagreement among linguists about a crucial judgment. The site will maintain a list of speakers of various languages who are willing to participate in the experiments, optionally for a small payment. A convenient interface will be provided for entering the relevant contrasts, optionally with several lexicalizations for each contrast. In addition, since the statistical analysis of judgment experiments is very straighforward, the site will automatically analyze and report the results in a graphical format.

[5]This idea was suggested to me by Jeremy Kuhn.

Even English may benefit from this system: some of the issues presented in Section 4 apply to widely researched languages as well, though perhaps to a lesser extent. For example, once a questionable English judgment made it into a published paper, it may be used by linguists who are not native English speakers and are not aware the controversy surrounding the specific judgment. This may be useful in cases such as the non-replication by Clifton et al. (2006) and Fedorenko and Gibson (2010) of Bolinger's (1978) judgment that superiority violations can be remedied by adding a third Wh-element.

On the other hand, this system will not be effective for languages with a small number of speakers and no practicing linguists. In those languages, the only way to convince a reader of the robustness of judgments will be to run a small scale experiment. Sprouse (in his NYU talk) shows that most contrasts are robust enough to show statistically significant differences with only a handful of speakers and items. A small scale field experiment along the lines suggested by Myers (2009b,a) and Collins et al. (2009) should be sufficient. This is mostly relevant for future studies; re-examining all the judgment in the existing literature would be an expensive and unrewarding enterprise which is unfortunately unlikely ever to be carried out.

Appendix: the contrastive view of judgments and S&A's choice of task

Both in S&A and in the present paper, the judgments collected from the literature included two kind of judgments: single-sentence grammaticality judgments on the one hand, and two-sentence contrasts, where a minimal variation on a sentence is shown to change its grammaticality status, on the other. In both cases, an ungrammatical sentence is marked with an asterix. Sprouse and Almeida used two different tasks to elicit the participant's judgments in these two cases. Single-sentence judgments were elicited using a yes-no task, with the following instructions[6]:

> For each sentence, simply click the **yes** radio button if you think the sentence could be spoken by a native speaker, or the **no** radio button if you think that the sentence could not be spoken by a native speaker.

This task reflects the common way of thinking about grammaticality judgments: a sentence is either grammatical – that is, it is a part of the language – or ungrammatical, in which case it is not part of the

[6]Taken from the online materials published on Jon Sprouse's website, at `http://www.socsci.uci.edu/~jsprouse/tools/MTurk.templates.zip`

language. A judgment from the literature was taken to have been replicated if significantly more than half of the participants agreed with it.

Sprouse and Almeida's materials had to be adjusted to be used with this task. Virtually all of Adger's single-sentence examples were judged in the literature as grammatical: out of 250 single-sentence stimuli, 247 were listed as grammatical and 3 as ungrammatical. To address this imbalance, Sprouse and Almeida introduced filler items in a way that the number of unacceptable and acceptable items was equalized. This means that almost all of the filler items were expected to be judged as unacceptable. The authors do not describe how the filler items were selected (other than the fact that they were used as items in a previous study), and how their grammaticality status was established. Intuitively, if the filler items were "more ungrammatical" than the target items were grammatical, this might lead the participants to rate items as grammatical which would not be rated as grammatical if presented in a balanced design. Moreover, Sprouse and Almeida do not report the participants' performance on the filler items (or a d' score); this means that a participant who was blindly marking every stimulus as grammatical would be taken to be replicating Adger's judgments.

For the two-sentence contrasts Sprouse and Almeida used a different task: instead of requiring the participants to determine whether each of the sentences in each contrast was grammatical or not, they instructed them to give the sentences a numerical rating, using magnitude estimation. The participants were shown a reference sentence, which was presented as having a grammaticality score of 100. The instructions were then as follows:

> For each sentence after the reference, you will assign a number to show how grammatical or ungrammatical the second sentence is in proportion to the reference sentence.

A grammaticality contrast from the literature was taken to be replicated if there was a reliable difference in rating between the two sentences, and that difference went in the right direction. This conception of grammaticality is weaker than the one used for the single-sentence examples: a starred sentence is not seen as ungrammatical in itself, but only as *less grammatical* than its counterpart. The contrastive view of grammaticality fits the considerably between-item variation we see in ranking results (Figure 1) much better the traditional absolute view. However, this view is called into question in cases where two forms are both grammatical, but differ in register, markedness or naturalness. This is often the case in Hebrew judgments,

which frequently revolve around the acceptability of different word orders. Consider the following two sentences:

(7) a. etmol dani axal tapuax.
yesterday dani ate apple

b. etmol axal dani tapuax.
yesterday ate dani apple
'Dani ate an apple yesterday.'

Neither of these sentences is ungrammatical. However, the SV word order in (7a) is more common than the VS order in (7b), and hence may be expected to be rated higher in an acceptability questionnaire. If that indeed turns out to be the case, this would not warrant marking (7b) with an ungrammaticality star, not even in the context of this particular contrast. This is a particularly worrisome case, in which a grammaticality judgment could be replicated in a relative acceptability rating task, but still be false.

A case in point is the ordering of adjectives. Shlonsky (1997) presents the following contrast:

(8) a. para švecarit xuma
cow Swiss brown

b. *para xuma švecarit
cow brown Swiss
'A brown Swiss cow'

I personally agree with this relative judgment: example (8) sounds better than (8b). But I would not go so far as to say that (8b) is *ungrammatical*. Incidentally, English has an analogous phenomenon, where color adjectives are normally placed before so-called provenance adjectives (Huddleston and Pullum 2002b, p. 453). And in English as well this seems to be a tendency rather than a grammatical rule: a Google search turns up results for both *brown Swiss cow* and *Swiss brown cow*. The contrast in (8) was one of the stimuli presented to the subjects in the current experiment (group 16). Both (8) and (8b) received very high rankings (more than 6 out of 7 on average). The difference in ranking between the two sentences did not reach significance. This may be an artifact of the Likert scale used for rating the stimuli: the sensitivity of the scale declines near the edges, so when both sentences in a contrast consistently receive high rankings a large sample size is required to detect a difference between them. However, it is quite plausible that a large enough sample size would indeed replicate the judgment. That would be a case of clear clash between the

two notion of grammaticality: the absolute notion would consider both as perfectly grammatical, whereas under the relative one only will be considered grammatical.

A conceivable risk is that the starred sentence could be taken out of the context of the particular contrast and presented as ungrammatical on its own – even though in isolation it might be ranked as acceptable. This, however, is less worrisome, since it is generally understood within the field that grammaticality contrasts are relative, and hence examples cannot be used out of the original context in which they were presented (citation needed). Any deviation from this practice can be detected using the usual methods of peer review, in the same way that incorrect use of statistics should not in principle find its way into psychology journals.

References

Adger, David. 2003. *Core syntax*. Oxford University Press.

Armon-Lotem, Sharon, Gabi Danon, and Susan Rothstein. 2008. *Current issues in generative Hebrew linguistics*. John Benjamins.

Belletti, Andrea, and Ur Shlonsky. 1995. The order of verbal complements: A comparative study. *Natural Language & Linguistic Theory* 13:489–526.

Bolinger, Dwight. 1978. Asking more than one thing at a time. In *Questions*, volume 1. Springer.

Clifton, Charles, Gilbert Fanselow, and Lyn Frazier. 2006. Amnestying superiority violations: Processing multiple questions. *Linguistic Inquiry* 37:51–68.

Collins, C., S. Guitard, and W. Jim. 2009. Imposters: An Online Survey of Grammaticality Judgments. Technical report, NYU Working papers.

Culicover, P.W., and R. Jackendoff. 2010. Quantitative methods alone are not enough: Response to Gibson and Fedorenko. *Trends in cognitive sciences* .

Edelman, Shimon, and Morten H. Christiansen. 2003. How seriously should we take minimalist syntax? A comment on Lasnik. *Trends in Cognitive Science* 7:60–61.

Featherston, Sam. 2009. Relax, lean back, and be a linguist. *Zeitschrift für Sprachwissenschaft* 28:127–132.

Fedorenko, Evelina, and Edward Gibson. 2010. Adding a Third Wh-phrase Does Not Increase the Acceptability of Object-initial Multiple-wh-questions. *Syntax* 13:183–195.

Gibson, Edward, and Evelina Fedorenko. 2010a. The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes* 1–37.

Gibson, Edward, and Evelina Fedorenko. 2010b. Weak quantitative standards in linguistics research. *Trends in cognitive sciences* .

Huddleston, Rodney, and Geoffrey Pullum. 2002a. A response concering The Cambridge Grammar. `http://linguistlist.org/issues/13/13-1932.html`.

Huddleston, Rodney, and Geoffrey K. Pullum. 2002b. *The Cambridge Grammar of the English Language*. Cambridge University Press: Cambridge.

Landau, Idan. 1999. Possessor raising and the structure of VP. *Lingua* 107:1–37.

Lee-Schoenfeld, Vera. 2006. German possessor datives: raised and affected. *The Journal of Comparative Germanic Linguistics* 9:101–142.

Myers, J. 2009a. Syntactic judgment experiments. *Language and Linguistics Compass* 3:406–423.

Myers, J. 2009b. The design and analysis of small-scale syntactic judgment experiments. *Lingua* 119:425–444.

Phillips, Collin. 2009. Should we impeach armchair linguists? *Japanese/Korean Linguistics* 17.

Phillips, Collin, and Howard Lasnik. 2003. Linguistics and empirical evidence. Reply to Edelman and Christiansen. *Trends in cognitive sciences* 7:61.

Pylkkänen, Liina. 2002. Introducing arguments. Doctoral Dissertation, MIT.

Ross, J., L. Irani, M. Silberman, A. Zaldivar, and B. Tomlinson. 2010. Who are the crowdworkers?: shifting demographics in mechanical turk. In *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems*, 2863–2872. ACM.

Shlonsky, Ur. 1997. *Clause structure and word order in Hebrew and Arabic: An essay in comparative Semitic syntax*. Oxford University Press, USA.

Siloni, Tal. 2002. Adjectival constructs and inalienable constructions. In *Themes in arabic and hebrew syntax*, 161–187. Kluwer Academic Publishers.

Sprouse, Jon, and Diogo Almeida. submitted. A formal experimental investigation of the empirical foundation of generative syntactic theory .

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