Semantic fieldwork: How experimental should we be?*

Lisa Matthewson

Abstract: In recent years there has been an increase in formal semantic research that is based on quantitative data collection methodologies (a.k.a. 'experiments'). At the same time, many researchers are conducting non-quantitative fieldwork-based studies, and still other papers rely on introspective data by the authors. In this paper I focus on methodologies for hypothesis-driven semantic fieldwork. The core question is whether and how the methodologies semantic fieldworkers are currently using should change, in response to challenges from both a typological angle (e.g., Aikhenvald 2018) and an experimental angle (e.g., Gibson and Fedorenko 2013). I support Davidson's (2020) proposal that *experimental* is a gradable predicate (i.e., there is no simple two-way dichotomy between experimental and non-experimental research), and that research can often be made more robust by adopting some practices from experimental fields. However, I argue that 'more experimental' is not always 'better'.

Keywords: Salish, Tsimshianic, St'át'imcets, Gitksan, fieldwork, experiments

1 Introduction

Semantic fieldwork is the practice of establishing empirical generalizations about meaning, through one-on-one or small group work with speakers. It has been practised for over a century (e.g., Boas 1917), but has been applied within formal linguistics for only a few decades (with Bittner 1987 as a pioneering example). The methodologies used by semantic fieldworkers are actively being discussed, refined and improved (e.g., papers in Bochnak and Matthewson 2015; papers in *Semantic Fieldwork Methods*).

Hypothesis-driven semantic fieldwork is fieldwork which gathers linguistic data with the goal of testing theoretical hypotheses about meaning. This type of fieldwork, as it is commonly practised, has recently been criticized from two quite opposite directions. The two strands of criticism can be roughly be characterized as follows: on the one hand, hypothesis-driven fieldwork is believed to be too experimental, and on the other hand, it is not experimental enough.

* Heartfelt thanks to Gitksan consultants Vincent Gogag, Jeanne Harris, Hector Hill, Ray Jones, Herb Russell, and Barbara Sennott – Ha'miiyaa! – and to St'át'imcets consultants Carl Alexander, the late Beverley Frank, the late Gertrude Ned, the late Laura Thevarge, and the late Rose Agnes Whitley – Kukwstum'ckál'ap!
For helpful feedback, I am very grateful to Ryan Bochnak, Kate Davidson, Henry Davis, Aaron Malong, Jérémy Pasquereau, the participants in Henry Davis's 2020 Field Methods class at the University of British Columbia, two anonymous reviewers for *Semantic Fieldwork Methods*, and audiences at the Sinn und Bedeutung 25 Roundtable on Semantic Fieldwork Methodology, the Fieldwork Forum at the University of California, Berkeley, SemanticsBabble and the Linguistic Fieldwork Working Group at the University of California, San Diego, and the 23rd Seoul International Conference on Generative Grammar (SICOGG 23). Thanks also to the organizers of all the events just listed, to Jérémy Pasquereau for his position paper at the Sinn und Bedeutung Roundtable, and to Kate Davidson for sending me a copy of her paper, which was the inspiration for this paper (Davidson 2020).
This research was supported by the Social Sciences and Humanities Research Council of Canada (#435-2016-0381) and the Jacobs Research Funds.

In this paper I outline both of these challenges, and after dismissing the first one, concentrate in more depth on the second. The particular question I address is whether the methodologies that are currently used by semantic fieldworkers should change to adopt more of the methods used in quantitative research.

I will argue in support of Davidson's (2020) proposal that *experimental* is a gradable predicate, and that any research that produces new data – including field research – can be made more robust by adopting some practices from experimental fields. However, I argue that 'more experimental' is not always 'better'. I provide an assessment of the current level of 'experimentality' of semantic fieldwork, and I outline where we are doing well, where we could do better, and where I think it is fine to remain non-experimental. Throughout, I give examples from published papers by fieldworkers and from my own fieldwork, illustrating how the proposed methods can be implemented.

2 Semantic fieldwork: Challenges from two directions

The value of hypothesis-driven semantic fieldwork is not obvious to everyone in the world. Fieldworkers face two explicit criticisms to the legitimacy of our methodologies, from opposite directions. I call these respectively the 'Be less experimental!' challenge and the 'Be more experimental!' challenge.

2.1 Be less experimental!

Some researchers explicitly reject a hypothesis-driven approach to the collection of linguistic data. I will illustrate this viewpoint by focusing on the methodological comments in Aikhenvald's (2018) introduction to her handbook on evidentiality.[1] Aikhenvald presents an extreme version of the anti-hypothesis-driven fieldwork idea, to the extent that she rejects any kind of elicitation, and claims that fieldworkers should gather information only from naturally occurring sources such as texts and conversation.

Let's start with the following quote, to get a flavour of the idea:

> Elicitation and translation from a lingua franca (be it English, Spanish, Portuguese, or Mandarin Chinese) will not produce sensible results. Grammatical elicitation ... "should play no role whatsoever in linguistic fieldwork"—a statement by R. M. W. Dixon ([2010]: 323), to which any linguist who has ever professionally worked on a language in its entirety will subscribe. (Aikhenvald 2018:7-8)

Further statements in the same vein include that "[e]licitation is likely to produce unnatural, artificial results' (2018:8), and that "[b]y using artificially constructed examples, translations, and tests imposed on speakers, one can hardly capture the essence of a category which does not have a ready-made translational equivalent" (Aikhenvald 2018:8).

In addition to claiming that elicitation fails to give decent results, Aikhenvald also proposes that hypothesis-driven fieldworkers are prone to questionable cultural attitudes and linguistic biases. She writes:

[1] Although Aikhenvald is a typologist, not all typologists have a problem with hypothesis-driven semantic fieldwork. On the contrary, many researchers do typological work while incorporating a formal linguistic perspective. See, for example, Hawkins (2004), and see Polinsky (2010) for discussion of the relationship between typology and formal linguistics.

> Proponents of deductive approaches to any feature of the language—grounded not in empirical study but in ad hoc ideas of what a language should have—run the danger of imposing intuitions or facts of their native language onto other languages. One of the reasons for mistakenly conflating the notions of evidentiality with reliability, possibility, probability, and epistemic modality lies in the English-centric approach to those languages which have evidentials, and the pitfalls of translation. (Aikhenvald 2018:7)

This passage contains several misconceptions. First, it misses the fact that the scientific method involves making a hypothesis, testing it on data, and then keeping, discarding or revising it based on the results of the tests. The dismissal of "ideas of what a language should have" conflates the two different concepts of a starting hypothesis and a final analysis. It also assumes that the initial hypotheses of hypothesis-driven fieldworkers always involve similarity with familiar languages. This is, however, not the case, as pointed out for example by Davis et al. (2014).

Second, this statement mistakenly assumes that hypothesis-driven fieldworkers rely only on translations to produce their analyses. The exact opposite is the case: translations are a very small part of a semantic fieldworker's evidence when drawing conclusions about meaning. See for example Matthewson (2004), Bohnemeyer (2015), among many others.

Third, the above quote accuses some fieldworkers of Anglo-centrism – yet it has repeatedly been shown that hypothesis-driven fieldwork has the ability to detect subtle types of linguistic variation. It is simply not the case that hypothesis-driven fieldwork is more likely to result in claims that languages resemble English. Again, for extensive discussion of this, with examples from several case studies, see Davis et al. (2014, 2015).[2]

All of these points have been made before, so it is difficult to know what to say that might convince someone who still believes it is wrong to ask native speakers questions about their language. Perhaps a simple thought experiment is worth trying. So, here is a question for readers who are native speakers of English. (Non-native speakers can play along as well, for fun.) Are the utterances in (1)-(3) fully acceptable, or is there something odd about them? Note that the question is not *why* these sentences might sound odd. The question is simply whether they sound fully acceptable, or not.

(1) Jane ate all the cookies and saved two for later.

(2) Imagine that Violet has exactly three children, who are one-year-old triplets. A new colleague of Violet's asks her how many children she has, and Violet replies "I have two."

(3) Toby has just been born. He has done nothing in his very short life so far except cry. One minute after his birth, he stops crying and suddenly smiles. His father says "He smiled again!"

These are not basic-level elicitation prompts, by the way; they involve advanced research questions (about quantifier denotations, scalar implicature, and presuppositions). Nevertheless, I am willing to bet that most readers were able to identify that there is something odd about each of them.

[2]Strangely, even the discovery of cross-linguistic variation can still somehow suggest Anglo-centrism to Aikhenvald. When discussing Matthewson et al.'s (2007) "doubt that evidentials have a uniform denotation across languages", she writes that this claim "bears an imprint of an English-language bias" (2018:53).

Assuming that that is the case, we should next consider whether readers of this paper have this ability only because they are linguists. This is highly doubtful. A quick spot-check with two acquaintances confirmed that these utterances are easy for non-linguist English speakers to make judgments about.

Now for the final crucial question: if speakers of *English* have the ability to respond to elicitation questions as in (1)-(3), why would anyone assume that speakers of *non-English* languages would lack the same abilities to answer questions about their languages?[3]

I would like to make one last point about a methodology that relies solely on collecting spontaneous language data: it is slower and less reliable than elicitation. The appendix to Aikhenvald's paper, which contains "suggestions for fieldworkers", provides a list of 40 questions that fieldworkers should answer about evidentials – but no tips for how to get the answers, except for the statement that the fieldwork "should be based on naturally occurring texts and conversations, avoiding elicitation and translating sentences from a metalanguage" (Aikhenvald 2018:37). Examples of the questions include "Is evidentiality in the language an obligatory grammatical category?"; "Can a noun phrase be marked with an evidential?" and "Does the 'reported' term have any connotation of unreliability?" (2018:37-38).

These are great questions. Let's compare how one can find the answers using elicitation, with how one can answer them only via the gathering of texts and conversations. (See also Matthewson 2004 for similar points.) Suppose we want to find out whether a noun phrase can be marked with an evidential. The fieldworker who uses elicitation will put an evidential on a noun phrase and ask whether the sentence it appears in is acceptable. Of course, they have to make sure that the discourse context is set up appropriately, that the correct evidential is chosen to match the speaker's evidence in the context, and that they understand the morpho-syntax well enough to construct the sentence correctly, all of which can take a bit of work. And ideally one asks the question not just once, but with several different evidentials, different noun phrase types, and so on. But once those pre-conditions are met, the answer is obtained: yes, or no.

The non-elicitation fieldworker, in contrast, has to either luckily chance upon an example of an evidential on a noun phrase (from which they can conclude 'yes'), or else, to conclude 'no', they need to gather a large enough corpus that lacks any evidentials on noun phrases to be sure that their absence is not an accident. Given that evidentials on noun phrases might be optional or even rare, this would potentially have to be a very large corpus, which of course has to be transcribed and translated in its entirety. (Note that I am imagining that the language under study is not a majority language, with a massive corpus already available.) And no matter how large a corpus one can realistically gather for a minority language, one still cannot be certain that the absence of the noun-phrase evidential was not an accident.

To clarify: I am not advocating against the use of corpora in data collection! Working with corpora has led to many important discoveries; see Katz (2012) for an overview of corpus methods and how they can be used in addressing theoretical semantics questions. And sometimes corpus study is the only option (for example, when investigating historical phenomena). What I am arguing against is the restriction to *solely* corpus-based investigation as a means to answering research questions,

[3] An anonymous reviewer observes that this thought experiment is not quite parallel to what Aikhenvald is envisaging when she mentions elicitation using a 'lingua franca' (see quote above), since in my thought experiment, the contact language and the target language are the same. While use of a contact language certainly can introduce challenges, these can be mitigated; see e.g., Matthewson (2004), AnderBois and Henderson (2015), among others. Notice also that the thought experiment involves simple acceptability judgments, and the language in which the instructions are given is unlikely to affect the results.

if native speakers are available. Restricting one's fieldwork to only collecting stories and passively waiting to see if a sentence of the right type shows up is akin to studying gravity by waiting for apples to fall from trees. Apples falling from trees can be useful clues and can even lead to strokes of insight. But when used to the complete exclusion of more experimental methods, they are not a realistic or reliable way of answering research questions.

2.2 Be more experimental!

The second challenge to semantic fieldworkers – opposite to the first – is the idea that we should be as experimental as possible. According to this view, linguistic data are not fully robust unless sufficient trials are run to enable statistical analysis.

This standpoint is exemplified well by Gibson and Fedorenko (2013) (see also for example Featherston 2007, among many others). According to Gibson and Fedorenko, the typical methodology currently used in syntax and semantics research "does not allow proper testing of scientific hypotheses" (2013:88). Their main target for this criticism is research in which the data are provided by authors' own introspective judgments, rather than fieldwork; however, their proposals for what counts as good methodology also rule out most of what is typically done by fieldworkers.

Gibson and Fedorenko (2018:88) assert that "future syntax and semantics research should adopt the following standards, common in all behavioural sciences, when using acceptability judgements as evidence":

(4) a. include many experimental participants

b. include many experimental materials, in order to rule out effects due to idiosyncratic properties of individual experimental items

c. use naïve participants, all of whom are blind with respect to the research questions and hypotheses, and include distractor materials

d. present counter-balanced lists of items to different experimental participants (in different random orders) (Gibson and Fedorenko 2013:88)

Failing to follow these methods is predicted to have stark consequences: "allowing lax data collection standards has the undesirable effect that the results and claims will often be ignored by researchers with stronger methodological standards" (Gibson and Fedorenko 2013:88).

A very similar point of view is advanced by Wasow and Arnold (2005); they write that:

> consulting the primary intuitions of native speakers is a type of psychological experiment. Hence, such data collection should be subject to the usual methodological expectations of experimental psychology. In particular:
>
> - The number of subjects should be large enough to allow testing the results for statistical significance.
> - The order of presentation of stimuli (that is, linguistic examples) should be randomized.
> - The subjects should be ignorant of the hypotheses being tested, preferably with double-blind presentation of stimuli.
> - The data collected should be subjected to appropriate statistical analysis.

> Unfortunately, such basic precautions are almost unheard of in generative linguistics. Consequently, the journals are full of papers containing highly questionable data (Wasow and Arnold 2005:1483-1484).

Not all experimentalists share these views, and in fact some experimentalists have explicitly argued against the idea that data collected outside quantitative experiments is faulty; see for example Phillips (2010), Sprouse and Almeida (2013; 2018), Sprouse et al. (2013), among others. Nevertheless, amidst the ever-increasing trend towards quantitative data collection methods in linguistics, there is a need to take this kind of challenge seriously. The question I want to pose is: do semantic fieldworkers need to make our data-collection and reporting practices more robust, using insights from the quantitative literature?

My response to this question is largely to agree with the proposals put forward by Davidson (2020). In the following section I summarize Davidson's core idea, and after that I will discuss each of her specific suggestions, evaluating their relevance for a fieldwork context.

3 Davidson (2020): Is experimental a gradable predicate?[4]

Davidson begins with the observation that linguistics differs from other sciences in the relation between whether research is theoretical or experimental, and whether or not new data are offered. In other sciences, there is a clearer divide between experimenters (who produce empirical data) and theoreticians (who produce theories based on other people's data). In linguistics, in contrast, many theoretical papers present novel data, and many papers present data that are not collected via experiments.

Given this situation, one might think that linguists should collectively agree on some guidelines about when, exactly, experiments are necessary. However, Davidson argues against such an either-or option. She proposes instead that "**The question should not be when or where an experiment is warranted, but a better consideration of which aspects of experimentation should apply** to a given piece of research" (2020:126, emphasis original). Davidson presents a call to the field, stating that we need:

> a re-thinking in linguistics of the current divide between experimental and non-experimental work, so that even the reporting of armchair intuitions takes on some traditionally experimental properties, and the status of "data" is more uniform between experimental and non-experimental linguistics. (Davidson 2020:126)

A prerequisite for breaking down the experimental/non-experimental divide is that we need to understand the term *experimental* as being "a multi-dimensional, gradable predicate" (Davidson 2020:126). Davidson outlines the dimensions of experimentality as in (5), and then gives several case studies exemplifying research projects that 'make use of something "a little bit experimental" to collect a compelling set of data' (2020:126).

[4] Morzycki (2021:1072) defines "gradability" as "the ability of linguistic expressions to occur in comparatives, superlatives, equatives, and related constructions." Thus, if *experimental* is a gradable predicate, we should be able to talk about some studies being more experimental than others, and so on.

(5) Davidson's (2020:126) dimensions of experimentality:
   a. Controlled sampling of speakers/signers
   b. Controlled manipulation of context
   c. Controlled manipulation of linguistic features
   d. Controlled type of response by participants
   e. Quantitative data reporting
   f. Statistical data analysis
   g. "Open science" procedural and data transparency

In sum, Davidson's core ideas are that (a) experimentality is a gradient notion; (b) we should not automatically dismiss research that is not fully experimental; and (c) there is value in trying to be at least a little bit experimental.

I strongly agree with all of these points. It seems to be exactly right that we should avoid a strict experimental/non-experimental divide, especially if that divide is wrongly causing non-experimentally-gathered data to be automatically dismissed as non-robust. At the same time, it is undoubtedly correct that some data collected in the most 'armchair' fashion could do with an increase in robustness (see also Tonhauser and Matthewson 2016 for discussion of this).

As might be expected given the above discussion of the two opposing challenges to fieldwork methodology, fieldwork typically occupies a space somewhere in the middle zone of the experimentality scale. Davidson points this out, writing that we should "acknowledge the type of rigorous data collection already being done by fieldworkers as at least somewhat experimental, even when it is not highly quantitative" (Davidson 2020:139). My goal for the remainder of this paper is to assess the current state of fieldwork methods on each of the experimentality dimensions, and advance some proposals about whether and where we need to up the experimentality grade.

Before turning to this, however, I first give some very brief background on my own fieldwork situation. Since every fieldwork situation has unique properties, my answers below should be understood in the light of my particular context.

4 Fieldwork contexts this paper is based on

My fieldwork experience primarily involves work on two Indigenous languages of North America: St'át'imcets and Gitksan.

St'át'imcets /šƛ'æƛ'yəmxəč/ (ISO code lil) is also known as Lillooet (English term) or ucwalmícwts (literally, 'language of the people'). It is a Salish language, of the Northern Interior branch, and is spoken in British Columbia, Canada. According to Dunlop et al. (2018), St'át'imcets had approximately 98 fluent speakers in 2018. There are active community efforts towards language revitalization and retention.

'Gitksan' (ISO code git) is the English name for a continuum of Interior Tsimshianic dialects spoken in the northwest Interior of British Columbia, Canada. According to Dunlop et al. (2018), Gitksan had approximately 523 fluent speakers in 2018. There are active community efforts towards language revitalization and retention.

For each of these two languages, the majority of the fieldwork is done with a small handful of speakers, although annual trips to Gitksan traditional territory have added over a dozen occasional

speakers to the pool of consultants. All the consultants for both languages are over the age of 60. Some of the consultants are literate in their languages, and some are not. All speak English fluently and fieldwork sessions are conducted primarily in English.

The fieldwork with each speaker takes place either in a large urban centre where my university is located (either at the university or at consultants' homes), or in the local speech communities (usually at consultants' homes). The fieldwork relationship with each consultant typically lasts for a period of years, and in some cases for decades.

A final note about my fieldwork and my reporting of data in research papers is that my methods are constantly evolving, and hopefully improving, over time. My own earlier publications did not always follow the desiderata I lay out below!

Although not every semantic fieldworker works on an endangered language, many work on minority languages and/or under-described languages. Some are even working in contexts where there is no shared language between researcher and consultant. At the other extreme, fieldwork also takes place on majority languages like English, and on languages that the researcher speaks natively. It follows from Davidson's proposals (and see also Tonhauser and Matthewson 2016 for this point) that even in situations where it is easy to obtain data from oneself or a couple of one's friends, it is valuable to act a bit more experimentally – which can begin with understanding that even when collecting data from oneself or a couple of friends, one is in fact conducting fieldwork.

5 The experimental dimensions and fieldwork

In this section I consider each of Davidson's dimensions of experimentality through the lens of a semantic fieldworker. I provide an assessment of which dimensions we are already experimental on, which ones we should be more experimental on, and which ones we should *not* necessarily be more experimental on. To preview my claims, the dimensions fall into four categories, as listed in (6).

(6) Dimensions of experimentality, in relation to fieldwork methods:
   a. *Already doing* (*but stay vigilant*)*:*
      Controlled manipulation of linguistic features
      Controlled manipulation of context
   b. *Could do better* (*but don't have to go all the way*)*:*
      "Open science" procedural and data transparency
   c. *Often can't do, and don't always need to do:*
      Controlled sampling of speakers/signers
      Quantitative data reporting
      Statistical data analysis
   d. *Shouldn't necessarily do:*
      Controlled type of response by participants

5.1 Already doing (but stay vigilant)

The first two dimensions are well-known to almost all linguists, and are used to a greater or lesser extent in all kinds of research, including 'armchair' studies. Semantic fieldworkers are

usually quite 'experimental' on these dimensions, but it can't hurt to always keep a vigilant eye on whether we are consistently applying them.

5.1.1 Controlled manipulation of linguistic features

Within semantics, controlled manipulation of linguistic features means using minimal pairs (or sets) in which the linguistic form is varied while other aspects of the data (in particular, the discourse context in which the utterances take place) are kept constant. As Davidson (2020) observes, the literature on semantic fieldwork methodology already addresses the need for such controlled manipulation of linguistic features (Tonhauser and Matthewson 2016, among others).

An example of this type of minimal pair is given in (7). Here, the researchers were testing the hypothesis that in Gitksan, the 'plain future' *dim* and the 'progressive future' *yukw dim* differ in their ability to make offers (cf. Copley 2009 for English *will* vs. *be going to*). This is investigated by using a discourse context which unambiguously involves an offer, and asking for acceptability judgments on utterances containing the two ways to convey future time reference. It turns out that only the plain future, and not the progressive future, is acceptable in offering contexts.[5][6]

(7) [I am hosting a potluck dinner next week. You have no idea what you're going to bring because you haven't thought about it yet. I tell you 'Nobody has offered to bring fry bread to the meal, but I hope somebody does.' You decide to offer to bring it so you reply:]

a. **Dim** di-bagw-i-'y=hl eeja-m t'ilix.
FUT COM-arrive-TR-1SG.II=CN fry-ATTR grease
'I'll bring fry bread.'

b. #Yukw dim=in di-bakw=hl eeja-m t'ilix.
PROG FUT=1SG.I COM-arrive=CN fry-ATTR grease
'I'm going to bring fry bread.'
Consultant's comment on (b): "No, it's not offering to be the one to bring the fried bread, it's just saying that you're going to bring the fried bread." (Todorović et al. 2020)

5.1.2 Controlled manipulation of context

The other type of minimal pair (or set) involves the controlled manipulation of context. In this case, the linguistic form is kept constant. An example is given in (8)-(9). Here, the researchers were testing whether the Gitksan reportative evidential *gat* restricts the relation between the time of the event and the time of the acquisition of evidence (i.e., the report). In order to test this, the linguistic form is kept constant and there are two different versions of the context, one in which the event time (Florence's ride) *follows* the time of the report (8), and one in which the event time *precedes* the report (9). It turns out that both are acceptable, which is a piece of evidence that

[5](7)a,b are a fully minimal pair, differing only in the absence vs. presence of the progressive marker *yukw*. This is obscured on the surface by predictable differences in agreement morphology, due to the dependent clause-type induced by *yukw*.

[6]Glosses not found in the Leipzig Glossing Rules: I/II = series I/II pronoun, ATTR = attributive, COM = comitative transitive, CN = common noun connective, CNTR = contrastive, PN = proper noun, REPORT = reportative.

gat does *not* restrict the temporal relation between the event time and the evidence acquisition time.

(8) [Florence told me yesterday that she would go for a ride **later in the day**. Today, I say to another person:]
Makxw=**gat** [t=]Florence ky'oots.
ride=**REPORT** [PN=]Florence yesterday
'Florence went riding yesterday.' (Hirayama and Matthewson 2022:185)

(9) [Florence told you yesterday that she had gone for a ride earlier that day. A friend asks you what Florence did yesterday and you reply:]
Makxw=**gat** [t=]Florence ky'oots.
ride=**REPORT** [PN=]Florence yesterday
'Florence went riding yesterday.' (Hirayama and Matthewson 2022:185)

Again, the literature on semantic fieldwork methods already contains discussion of the desirability of controlling for discourse context. For example, Matthewson (2004:393) writes that "discourse contexts are necessary not just with translation but also with judgment tasks," and Deal (2015) proposes the hypothesis in (10), which makes clear that controlling for context is crucial (emphasis added):

(10) Equivalent judgments hypothesis (EJH):
In a particular context, speakers accept/reject sentences expressing the same range of propositions regardless of what language they are speaking. (Deal 2015:159)

The literature on semantic fieldwork methodology has also engaged in discussion of the best way to present discourse contexts to consultants (see for example AnderBois and Henderson 2015). One useful way to make sure the context is kept uniform across consultants, researchers, and language varieties is to use storyboards (Burton and Matthewson 2015; Vander Klok 2019; Cable 2019; Bochnak and Matthewson 2020, among many others).

Although fieldworkers are often fairly good at providing discourse contexts for utterances that are judged by consultants, contexts are still not used consistently across the board. Contexts are often used more sparingly in papers based on the researcher's own judgments. Yet as argued by Tonhauser and Matthewson (2016), contexts are important not just for official fieldworkers, but for any paper that includes semantic data. Generalizing Deal's hypothesis in (10), we can say that utterances are only true or false, felicitous or infelicitous, or acceptable or unacceptable, when embedded in contexts of utterance. It follows from that that the context should always be explicitly controlled when making claims about semantics or pragmatics.

5.2 Could do better (but don't have to go all the way)

The next category of experimentality dimensions has only one member. It is a dimension on which I think semantic fieldworkers could ideally be more experimental than we currently are. However, I also think there are reasons why we do not have to go all the way to the top of the experimentality scale.

5.2.1 "Open science" procedural and data transparency

"Open science" procedural and data transparency is when the researcher makes their procedures and data clear enough so that other researchers can attempt to replicate the data, in the same or another language.

The goal behind this is obviously a good one. For example, I believe that we should report full pieces of data, and following Tonhauser and Matthewson (2016), full pieces of data include all of the following:

(11) Full pieces of data contain:

i. a linguistic expression

ii. a context in which the expression is uttered

iii. a response by a native speaker to a task involving the expression in that context

iv. information about the native speakers who provided the responses

This type of transparency in the reporting of data is important not just for official fieldworkers, but for any paper that presents semantic data. The reader should know where the data came from. How many speakers were consulted? What dialects did they represent? Are the data from the researcher's own judgments, or somebody else's?

The reader should ideally also be told what exactly was asked of the speakers. Thus, it is not ideal to simply state that a certain utterance does or does not entail, presuppose, or implicate some proposition. In order to facilitate replication and to enable the reader to assess the robustness of the evidence, it should be stated what task the speaker(s) of the language took part in to lead the researcher to the conclusions being advanced.

It is worth pointing out that transparency in data reporting looks a little different for fieldworkers than it does for quantitative researchers. In a fieldwork study, it is almost always unreasonable to provide all the data that was collected, or even all the stimuli materials. Fieldworkers usually do not use invariant questionnaires or ask exactly the same questions of every speaker. We ask follow-up questions, depending on what happens in the elicitation session. We ask things many times, perhaps in slightly different ways. There may even be some trial and error and false starts when fieldworkers first start working on a particular topic, while we find out which elicitation scenarios/lexical items/methodologies work; this is a luxury not afforded the experimentalist due to the greater resources needed to set up a large experiment. So rather than publishing all the data, we extract generalizations from the data we have collected, and in published papers we report selected examples that illustrate the generalizations.[7]

These common fieldwork practices do lead to a lack of total transparency (in that not all data collected is included in the reporting), and they create the need for a level of trust between researcher and reader. For example, the reader needs to trust that the fieldworker has not 'cherry-picked' examples, keeping silent about other data that might have gone a different way. In my experience, this trust is justified. Fieldworkers in general – even those with theoretical preferences – are strongly committed to portraying the languages they work on accurately. This

[7]Thanks to Henry Davis (p.c.) for helpful discussion of these points.

includes that any data that are not fully consistent are reported with notes about the variation, and that if the data are very variable, papers are not written about those topics until further clarity is gained. See, for example, discussion of this in Tonhauser (2019); I return to this also in section 5.3.2 below. A related issue is whether fieldworkers should report how many pieces of data their conclusions are based on. This has not been typical practice so far, but it would be a way to increase our experimentality level on this dimension. Exact numbers are probably often not feasible given the way fieldwork works, and the sometimes months-long investigations of a single research question. But at least we could say how many speakers gave data for each generalization made, and perhaps we could say how many fieldwork sessions, or approximately how many different examples were asked.

An inspiring example of the level of transparency that is possible in fieldwork data reporting is found in Grubic's (2015) study of focus phenomena in Ngamo (West Chadic). Grubic worked primarily with two consultants using one-on-one fieldwork methods, and also collected questionnaire responses from eight speakers. Even for the one-on-one fieldwork, she reports the exact questions the speakers were asked, the response scale they were asked to use, and the numbers of each judgment that were given for each data point (see, e.g., Grubic 2015:102ff). Another nice illustration of transparency in fieldwork data reporting is found in Driemel's (2019) investigation of scope phenomena and pseudo-noun-incorporation in various non-European languages.

An interesting question when it comes to data transparency is how to deal with 'off days', or end-of-session fatigue, or any natural human event that leads the results to seem less reliable on certain occasions. Many fieldworkers find that every so often a certain session, or part of a session, produces results that are anomalous with respect to other similar questions that have been asked on other occasions. There is no disrespect to consultants in such an observation. It is merely an acknowledgment that consultants and fieldworkers are human beings, and any one of them may have a day when they have concentration lapses, or are distracted or tired.

Excluding the results obtained on those fieldwork 'off days' is very parallel to what quantitative researchers do when they routinely exclude outliers, and focus on what they consider to be more reliable data. One difference is that while a quantitative researcher might exclude an entire participants' data, fieldworkers very often do not have that possibility. We are typically working with the same group of consultants for long periods of time, so rather than excluding consultants altogether, the fieldwork equivalent could be that we exclude 'outlier sessions'. In order to provide transparency, best practice is to report that a certain number of sessions were excluded.[8]

Another difference between small-scale fieldwork and large-scale experiments is that usually, the exclusion of an outlier elicitation session is not based on numerical metrics (such as vastly slower or quicker reading times, for example), but on factors such as the fieldworker's observation of the personal dynamics of the session, the disruption of judgments which are otherwise consistent across several sessions, and so on. We should take steps to confirm that the results from this session are really one-off outliers, and we should also rule out the possibility that other factors were responsible for the judgments given (for example, that the discourse contexts presented were slightly different, or presented in a slightly different way).

Emily Clem (p.c.) points out another case where quantitative and fieldwork studies are partially parallel, but the fieldwork version is more informal. Experiments frequently begin with

[8]Thanks to Emily Clem, Kate Davidson, Ginny Dawson, and Andrew Kehler for helpful discussion of these issues.

training items, and in some experiments (particularly first-language acquisition studies), participants who 'fail' the initial items may be excluded from the results. The fieldwork version of this is that the first session with a new consultant tends to involve easy questions, to get a feel for whether the consultant enjoys the task and is able to give clear responses. If not, then the fieldworker switches to another type of task, and in some cases, abandons hypothesis-driven elicitation altogether with certain individuals (asking instead for spontaneous narratives, for example).

Finally, it is worth pointing out that the generally more casual, flexible nature of fieldwork sessions and the typically long-term working relationships that develop between linguist and consultants also give rise to some advantages over large anonymous experiments. For example, the in-depth, one-on-one conversations that take place with fieldwork consultants mean that fieldwork researchers are sometimes able to give *more* information than experimental researchers.

A recent paper by Berthelin (2020) addresses this point; she argues that transparency is increased by the inclusion of direct quotes drawn from consultants during fieldwork. She writes:

> [Q]uotes show exactly what the consultant has said in response to the stimulus that, together with other data, has led the researcher to draw the given conclusions about the meaning of the expressions under investigation ... This is in line with Cover and Tonhauser's (2015:343) call for more transparency with regards to what consultants have said in interviews. (Berthelin 2020:27)

An example of this fieldwork-style transparency is drawn from Berthelin's work on North Slope Iñupiaq (Inuktut). The research question in this case was the meaning of the suffix *-niq*. In (12), the task for the consultant was to 'describe a situation or scenario where she would utter [the] sentence to another speaker of the language.'

(12) Aalaak umiaqaġ**niq**suq!
Aalaak umiaq -qaq **-niq** -tuq
Aalaak boat have **turns.out** IND.3SG
'Aalaak does have a boat!' (Berthelin 2020:19)

The consultant commented as follows:

> The first scenario for me is: I've been wondering with someone else whether this person has a boat. I go and check to see whether that person has a boat. I see that he has a boat, 'cause I... see it. And then I go back, or I holler back to the person: *Umiaqaġ**niq**suq* 'Yes, he does have a boat!' (Berthelin 2020:19)

Comments such as this provide a rich window into the consultant's understanding of the contribution of the suffix.[9] After gathering many such pieces of data relating to *-niq* (including, of course, minimally different variants lacking *-niq* for comparison), the fieldworker will develop a generalization about the type of discourse situation *-niq* is licensed in. In addition, providing some of the consultant's comments offers transparency about what that generalization is based on. In this particular case, Berthelin argues that "*niq* marks a connection between the utterance and the previous discourse by indicating that the speaker has realized that the propositional content is a true description of the world, and that this realization is relevant to the previous discourse' (Berthelin 2020:23; see Berthelin 2012 for a full analysis).

[9]Importantly, the consultant is not asked directly about the contribution of the suffix. They are only asked for comments about full sentences containing or lacking the suffix (e.g., Matthewson 2004).

Notice that I argued for providing *some* – not all – of the consultant's comments. Giving a verbatim recording of the elicitation session would not only be unwieldy and impractical, it might even give rise to misunderstanding. The fieldworker, who knows the consultant personally and has interacted with them on many occasions, is in a position to interpret the consultant's responses in a way that the reader cannot. Thus, the same literal response from one consultant might count as an emphatic rejection, but from another as marginal acceptance. In addition, the fieldworker is privy to the consultant's gestures, body language, facial expression and tone, all of which cannot reasonably be fully conveyed in a publication.[10]

In summary, it is a good idea to strive for as much transparency as possible in the reporting of data. This can and should include, for any type of linguistic study, information about the source of the data, the number of speakers and which varieties they speak, full information about the elicitation tasks that were carried out, and information about the exclusion of outlier data. It can also include comments by consultants that transparently reveal the basis of the researcher's generalizations about meaning. However, unlike for quantitative studies, it is usually not possible or necessary for fieldworkers to present all data that was collected.

5.3 Often can't do, and don't always need to do

Now we turn to three dimensions of experimentality that are not easily achievable for many semantic fieldworkers. I argue that we should do what we can, and that what we can do is good enough. However, I also argue that for some research questions, a high experimentality grade is not necessary on these dimensions.[11]

5.3.1 Controlled sampling of speakers/signers

Fieldworkers working on languages with small speaker populations cannot do controlled sampling of speakers or signers. We simply work with consultants who are willing and able to do the work.

In many cases, we work with the only speakers who are willing and able to do the work – we have no choice, because there may only be a small handful of remaining fluent speakers. In other cases, the choice of which speakers to work with is based on a range of factors including geography, mobility, health, community politics, personality fit, enthusiasm of the consultant to do certain types of fieldwork task, and so on. Ideally, the choice of consultants would consider demographics and include a spread of genders, ages, dialects, and so on – but this may not always be possible.[12]

In spite of these practical limitations, there is also gradience on this dimension. We can and should at least be transparent about the speakers who gave our data, so the reader can judge the

[10] Thanks to Emily Clem, Kate Davidson, and Ivano Caponigro for helpful discussion of these points. Davidson points out that in order to alleviate concern about the interpretation of different consultants' comments, fieldworkers could explain the basis of their reasoning, e.g., "The consultant said *X*; *X* has been used by this consultant in this type of task to mean *Y*, so I interpret it here as meaning *Y*."

[11] Thanks to Jérémy Pasquereau and two anonymous reviewers for encouraging me to argue that experimentality is not always necessary on these dimensions.

[12] It is worth pointing out that quantitative experiments can also suffer from the problem of non-representative populations. Most quantitative studies in psychology and experimental linguistics are carried out on the sub-population of university students from WEIRD societies (Western, Educated, Industrialized, Rich, and Democratic; see Henrich et al. 2010 for discussion). Thanks to an anonymous reviewer for this point.

empirical spread for themselves.

Interestingly, the challenge of working on minority languages is, for Gibson and Fedorenko, the one argument in favour of non-quantitative data collection (out of seven considered) that they judge to be valid. They write:

> in circumstances where gathering data is difficult, it is better to gather as much data as possible, even if no statistical evaluation of hypotheses is possible (without a sufficient number of participants and/or items). For example, a speaker of a dying language[13] may only be available for testing during a brief period of time. (Gibson and Fedorenko 2013:94)

They further write, however, that "[o]f course, the conclusions that can be drawn from these data will be weaker and more speculative in nature than the conclusions based on quantitative data" (Gibson and Fedorenko 2013:94).

Even though Gibson and Fedorenko admit that gathering non-quantitative data is better than no data at all and is sometimes the only choice, it is possible to push back against the 'weaker and more speculative' rhetoric. The issue here is a divide between those who believe that even a single speaker's grammar is a valid object of study (because it reveals something about human language), and those who believe that our goal must be to capture the speech patterns of an entire speech community. I belong to the former camp: the object of study is, in the end, a grammatical system that forms part of the cognitive architecture of a speaker, and is ultimately housed in the brain of a single speaker. Drawing perhaps on the same base, Davidson (2020) also writes: "That there exist interesting linguistic arguments for which such ["armchair" introspective] intuitions can provide a full set of supporting data is not in doubt to this author."

5.3.2 Quantitative data reporting and statistical data analysis

If controlled sampling of speakers or signers is not possible due to minority language status, quantitative data reporting and statistical data analysis are likely not possible either. Note that this may be true not just for a few isolated language situations, but for the majority of the world's languages. Anyone who is interested in linguistic diversity must acknowledge the importance and urgency of research on minority languages, and therefore must be satisfied with doing without quantitative results some of the time.

However, this is also a gradient dimension: we can be a tiny bit experimental. For example, even when working on a language with a small number of speakers, we can at least do the following for each generalization made: (a) have more than one test sentence; (b) replicate the data with the same speaker at different times, using different lexical items; and (c) if possible, work with more than one speaker.[14]

Explicit discussion of this type of fieldwork-style quantitative data reporting is found in Tonhauser (2019). Tonhauser writes that "[d]ata collected in semantic/pragmatic fieldwork are typically not subjected to statistical analysis due to a lack of power: for any given hypothesis, I collect around 5 pieces of data from around 3-5 native speaker consultants" (2019:2). Tonhauser

[13] The metaphors of dying, dead and extinct languages are no longer regarded as appropriate; see for example Perley (2012) for discussion and alternatives.

[14] Even if one believes (as I do) that the grammar of a single speaker is a valid object of study, it is still useful to elicit data from more than one person. Apart from anything else, it helps confirm that the methodology being used is giving consistent and reliable results.

further writes that "[a]s a consequence, the empirical support for a hypothesis supported by such data needs to be relatively noise-free: I typically only report hypotheses where all of the consultants agreed in their responses to all of the pieces of data elicited in support of that hypothesis" (2019:2). This highlights something I touched on above: that overwhelmingly, fieldworkers attempt to convey accurate generalizations about the languages they work on. Although we may not perform statistical analysis, we try to base our hypotheses on data that has been a little bit experimental in terms of who it was collected from and how many data points the hypothesis relies on. Just like Tonhauser, every fieldworker I have discussed this with applies a similar rule: if the data are 'noisy' (i.e., messy) and refinement of the testing procedure or materials cannot produce consistent results, we either do not write a paper about that topic, or we report on the variation.[15]

Finally, there are reasons why one might want to be cautious about conducting an experiment with quantitative and statistical power, even if one is logistically able to do so. The first reason is that for some research questions, a large experiment would be a waste of time and money. Here is an example provided by a reviewer: suppose that one's research question is whether evidentials can appear on noun phrases. If the first three to five native speakers one asks immediately and without any qualification judge a sentence containing an evidential on a noun phrase to be acceptable, there would be little point in investing the resources to run a full-blown quantitative study.

Finally, caution may also be called for due to unforeseen issues that can arise when conducting quantitative experiments on non-WEIRD populations. For example, Bombi and De Veaugh-Geiss (2018) and De Veaugh-Geiss (2021) report on experiments that were designed to test definiteness presuppositions and exhaustivity in Akan (Kwa). These researchers found that applying the same methodologies as had been used in European language-experiments led to unexpected results, and moreover the experimental results were at odds with the results obtained through small-scale fieldwork (while the fieldwork results did pattern as had been expected based on other languages). Bombi and De Veaugh-Geiss discuss various reasons why the experimental results were unreliable in Akan, including orthographic variability, participants' unfamiliarity with aspects of the experimental procedure, and multi-lingualism. More generally, it seems clear that many populations across the world are not as familiar with test-like experimental protocols as are populations in North America, Europe, and East Asia, and this may impact the reliability of results.[16]

5.4 Shouldn't necessarily do

In the previous section we looked at dimensions of experimentality for which it is often impossible to get very high on the scale. Here I discuss the final dimension, and argue that it is not necessarily desirable to become more experimental in this respect, even if it is logistically possible.

[15] Henry Davis (p.c.) points out that some topics give rise to inherently messy or variable results, including for example studies on the semantics of intonation. Correspondingly, these might be more likely to require greater statistical power to obtain clear generalizations. Note however that even here, we would not want to exclude small-scale studies altogether, or we would risk losing the chance to learn anything about intonation in minority languages.

[16] Thanks to an anonymous reviewer for this point.

5.4.1 Controlled type of response by participants

In semantic fieldwork, the response expected from the consultant could in principle vary from extremely free (for example, asking the consultant to tell a story or talk about anything they want to talk about, or having a metalinguistic conversation about some phenomenon) to extremely controlled (for example, asking for numerical acceptability judgments on a Likert scale). In practice, the vast majority of hypothesis-driven fieldwork lies somewhere in between these two extremes. The most free types of response are either not perfectly suited for hypothesis-testing (asking for free discourse) or not standardly accepted as a good idea (having metalinguistic conversations; see Matthewson 2004 for arguments against this). The other extreme – where one *only* permits the consultant to respond in numerical ways – would lose many of the advantages of one-on-one fieldwork, such as were discussed in section 5.2.1.

Given this, let's assume that the option of *only* asking for fully controlled responses is not on the table for fieldworkers, and that typical elicitation sessions include some opportunity for the consultant to offer comments. The remaining question is whether there is an advantage to *also* being more experimental by asking consultants to give standardized responses, for example on a strict response scale (which could be binary, for example 'true' vs. 'false' or 'good' vs. 'bad').

One advantage of including a standardized response option, pointed out by Jérémy Pasquereau (p.c.), is that it reduces potential worry about the researcher's process of interpreting the speakers' responses, specifically any worry about whether unconscious bias could arise. Moreover, as pointed out by Emily Clem (p.c.), sometimes it can be useful to restrict consultants' response options initially, and then to invite them to give free responses in follow-up discussion. This is the case, for example, with speakers who are otherwise inclined to respond to acceptability judgment tasks by offering their preferred way to say the sentence. While such volunteered utterances give valuable information, it can be useful to use a two-step process whereby an explicit judgment on the original utterance is solicited first, and then the consultant offers their preferred alternative.

In spite of these considerations, there are also reasons in some fieldwork contexts to avoid using standardized consultant responses. In my fieldwork situation, for example, it is not appropriate to restrict the consultants' answering possibilities. When working in a close relationship with consultants for a long time, I prefer to let them say whatever they want in response to elicitation tasks. I don't ask for responses on a scale. For acceptability judgment tasks, I usually just tell the consultant that I want to know whether the utterance sounds good in the given context, or I simply ask "How does this sound?"[17]

This way of conducting fieldwork is more culturally appropriate in my fieldwork context, because it is less formal and because it shows the consultants, who are elders, that I respect their thoughts and insights and that I am willing to listen. It also allows the consultant to feel less like an experimental participant, and more like an active collaborator in the research. This in turn makes the fieldwork experience more enjoyable, and contributes to the long-lasting researcher-consultant relationships that can be so fruitful. However, I should point out once again that not

[17] Kate Davidson (p.c.) suggests that it would be beneficial to standardize the way the question is phrased across different consultants, at least for the same data that is reported as 'repeated', because small variations in asking whether something is "good", "true", or "appropriate", etc. can make a difference. I agree that asking about truth is different from asking about acceptability (and I usually don't ask my consultants about truth, because it is a loaded term). I am not yet sure that different ways of asking about acceptability make a difference in the context of a typical fieldwork session, which is a conversation. I think long-term consultants understand well what the fieldworker wants to know when they ask about acceptability.

every fieldwork situation is the same, and what is culturally appropriate differs. In some types of community, asking for numerical responses may be perfectly fine.

Apart from the human motivations for eliciting relatively unstructured responses in some fieldwork contexts, I also believe that allowing freedom in consultant responses is scientifically justified. As long as the researcher records all the comments made by consultants, no information is lost, and as outlined in section 5.2.1, in some cases more information is gleaned than would have been obtained through numerical or binary responses.

Here are some examples of the rich types of information one can gain from non-controlled responses by consultants. In example (13), the research question was whether in Gitksan, one can reply to an assertion of a proposition *p* by saying 'yes' followed by the negation of *p*. The background for this question is that according to Guntly (2021), in English a reply of *Yeah, that's not true* (or similar) is felicitous and means 'I accept that you believe that, but I don't agree.' Guntly found, for example, that a majority of experimental participants rate English conversations like (13) as either 'very appropriate' or 'appropriate'.

(13) [A and B are discussing a movie they saw. After the first two utterances, C enters the room.]
A: The car chase scene was long.
B: Yeah, it wasn't.
C: How was the movie?
B: She thinks the car chase scene was long, but I don't.

A first attempt to elicit judgments on a similar discourse in Gitksan produced the result in (14).

(14) A: Gukws jo<u>k</u>=t Clarissa <u>g</u>a'a=hl Arizona.
back live=PN Clarissa LOC=CN Arizona
'Clarissa moved to Arizona.'

B: Ee', nee=dii wil-t.
yes NEG=FOC do-3SG.II
'Yes, she didn't.'
Consultant's comment: [Laughs] "It's an agreement when you say *ee'*, that's an agreement. Yes, she didn't ... I've heard this when I was younger when the elders were talking and they were saying similar things like this ... *Ee' needii wilt.* B knows that she didn't. A believes she did. A says *Gukws jo<u>k</u>t Clarissa <u>g</u>a'ahl Arizona* but B knows that it's not right. B is agreeing about the person, not about the whole thing."

This is a preliminary result, and much more testing is required before we would conclude that Gitksan patterns like Guntly's analysis of English. However, the fact that the consultant spontaneously offered these comments appears to suggest that the hypothesis might be on the right track. These valuable clues would have been missed if the consultant had been instructed to merely choose a number on a Likert scale. And it is difficult to see what extra information would have been gleaned by asking for a numerical response in addition.

Another example is provided in (15), also from Gitksan. Here, the hypothesis being tested was that in contexts where a predicted future event is contingent on another event, the plain future

marker *dim* will be appropriate, while the progressive future *yukw dim* will be inappropriate. This hypothesis is based on observations by Binnick (1971), Klecha et al. (2008), and Klecha (2011) about English *will* vs. *be going to*. The consultant's comment not only reveals that (15b) is inappropriate in the discourse context provided, it gives additional information about what exactly has gone wrong.

(15) [There is a bomb which explodes when somebody opens the door. You warn me:]

a. Ham ji das[-t]=hl aats'ip, dim xhluxw=hl bomb!
don't.2SG IRR touch[-3.SG.II]=CN door FUT explode=CN bomb
'Don't touch the door, the bomb will explode!'

b. #Ham ji das[-t]=hl aats'ip, yukw dim xhluxw=hl bomb!
don't.2SG IRR touch[-3.SG.II]=CN door PROG FUT explode=CN bomb
'Don't touch the door, the bomb is gonna explode!'
Consultant's comment: "When you say *Yukw dim xhluxwhl bomb*, there is a certainty at some point the bomb will explode whether you touch the door or not." (Todorović et al. 2020)

Free comments of this type can be particularly useful in cases where an utterance is inappropriate in the context, as in (15b). As Matthewson (2004:375) points out, "sentences are rejected by consultants for a variety of reasons[;] this merely means that the fieldworker needs to determine which of those reasons obtains, every time s/he receives a negative judgment." Although consultants' comments don't provide an analysis, they can – unlike pure numerical responses – give clues about the reason for an utterance being rejected.

A class of cases where comments are particularly vital in explaining unacceptability is when the unacceptability is for non-linguistic reasons, such as cultural practices or stereotypes. An example provided by an anonymous reviewer involves the rejection by a speaker of a syntactically well-formed Hausa sentence meaning 'Audu [male name] is cooking beans.' Upon prompting, the consultant commented "This sentence is not acceptable because men don't cook."

A final example of the usefulness of consultant comments comes from St'át'imcets (a.k.a. Lillooet; Northern Interior Salish). The researchers were testing the effect of the frustrative marker *séna7* ([ʃɨna?]) on motion verbs. The minimal triplet in (16) contains three different motion verbs. In each case, something unexpected happened (consistent with the presence of *séna7*), but the consultant's comments provide useful clues about whether the motion (16a) has to start and also reach a destination, (16b) must start, but does not need to reach a destination, or (16c) is able to not start at all.

(16) [You were meant to be going to a gathering.]

a. **Tsícw**=kan=t'u7 **séna7**, t'u7 cw7it i=n-száyten=a.
get.there=1SG.SBJ=EXCL **CNTR** but many PL.DET=1SG.POSS-doings=EXIS
'I got there, but I had a lot to do.'
Consultant's comment: "It says you went, because of *tsícwkan*."

b. **T'ák**=kan=t'u7 **séna7**, t'u7 cw7it i=n-záyten=a.
go.along=1SG.SBJ=EXCL **CNTR** but many PL.DET=1SG.POSS-doings=EXIS
'I went, but I had a lot to do.'
Consultant's comment: "He was going, but he came back."

c. **Nás**=kan=t'u7 **séna7**, t'u7 cw7it i=n-záyten=a.
go=1SG.SBJ=EXCL **CNTR** but many PL.DET=1SG.POSS-doings=EXIS
'I was going to go, but I had a lot to do.'
Consultant's comment: "Didn't go." (Davis and Matthewson in press)

One issue with asking *only* for free responses rather than responses on a quantitative scale is that the fieldworker has work to do to interpret the results. That is, the fieldworker has to decide how to translate consultants' comments into 'judgments' such as *, ?, or # – which are, in this kind of study, not strictly judgments but rather the linguist's interpretation of the consultants' responses. Fieldworkers tend to be very good at interpreting their consultants' responses; remember, these are often relationships lasting years, and each consultant's individual response style will be well known to the linguist. However, there is certainly room for increased transparency in the use of such symbols (throughout the non-quantitative literature, not just in papers based on fieldwork).

A first step is, as advocated already, to explain the exact nature of the task the speaker was asked to perform, including whether they gave judgments on a numerical scale or not. There is also the option, as discussed above, of increasing transparency by including more of the verbatim comments in publications. Including the consultants' raw comments can help mitigate potential fears that in the process of interpreting the consultants' free responses, the researcher has let their own biases unconsciously creep in. We should also be transparent about our general linking hypotheses, which explain how the data provide support for the hypothesis. For example, we can state that if a consultant replies in such-and-such a way to such-and-such a task, we interpret this to mean that the utterance is true/false/infelicitous, and so on (see Tonhauser and Matthewson 2016 for discussion).

6 Summary: When to be experimental

In this paper, I have argued in support of Davidson's (2020) proposal that *experimental* is a multi-dimensional, gradable predicate, and that rather than perceiving a dichotomy between experimental and non-experimental research, we can view research as existing along a continuum from more to less experimental. I also agree with Davidson that any researcher, including those working on their own native language, can make their research 'a little bit experimental,' and that this often improves the robustness of the data.

My contribution here has been to go into a bit more detail about where I think semantic fieldwork should sit on the experimental scale, for each dimension. And my tweak on Davidson's proposals was to make explicit that 'more experimental' is not *always* better. Sometimes there are good reasons to be below the top of the scale.

Table 1 summarizes the claims I have made about how experimental semantic fieldworkers should strive to be on each of Davidson's dimensions. Recall that I intend to include not just official fieldworkers, but anyone who collects data about meaning.

Table 1: When and how far should we try to be experimental on each dimension?

Controlled manipulation of linguistic features Controlled manipulation of context	All the time
"Open science" procedural and data transparency	All the time, to a reasonable extent
Controlled sampling of speakers/signers Quantitative data reporting Statistical data analysis	Only as far as you can, and as far as is reasonable for the research question
Controlled type of response by participants	Not necessary (but be transparent)

I will end by saying that every claim made in this paper was my own personal opinion based on my own experience as a fieldworker. The proposals I made are *not* intended to be mandates that everyone 'should' obey. It is also important to explicitly and clearly state that there is room in linguistics for data collected in all kinds of different ways. There is no call for 'methodological monotheism' (thanks to Farrell Ackerman for this term). Indeed, any empirical generalization is likely to be made stronger if it is supported by more than one type of data: data collected from corpora and other spontaneous sources as well as fieldwork, and (if logistically possible) from more highly experimental studies as well.

References

Aikhenvald, Alexandra 2018. *The Oxford Handbook of Evidentiality*. Oxford: Oxford University Press.

AnderBois, Scott, and Robert Henderson 2015. Linguistically establishing discourse context: Two case studies from Mayan languages. In *Methodologies in Semantic Fieldwork*, ed. by Lisa Matthewson and Ryan M. Bochnak, 207–232. Oxford: Oxford University Press.

Berthelin, Signe Rix 2012. The semantics and pragmatics of the North Slope Iñupiaq postbase niq. M.A. thesis, Trondheim: Norges teknisk-naturvitenskapelige universitet [Norwegian University of Science and Technology]. https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/244045.

Berthelin, Signe Rix 2020. Semantic elicitation - A discussion of elicitation frames and their application. *Semantic Fieldwork Methods* 2(1): 1-34.

Binnick, Robert. 1971. Will and be going to. In *CLS 7: Papers from the 7th Regional Meeting of the Chicago Linguistics Society*, 40-52. Chicago: Chicago Linguistics Society.

Bittner, Maria 1987. On the semantics of the Greenlandic antipassive and related constructions. *International Journal of American Linguistics* 53(2):194-231.

Boas, Franz. 1917. Introductory. *International Journal of American Linguistics* 1:1–8.

Bochnak, M. Ryan, and Lisa Matthewson, eds. 2015. *Methodologies in Semantic Fieldwork*. Oxford: Oxford University Press.

Bochnak, M. Ryan, and Lisa Matthewson. 2020. Techniques in complex semantic fieldwork. *Annual Review of Linguistics* 6:261-283.

Bohnemeyer, Jürgen. 2015. A practical epistemology for semantic elicitation in the field and elsewhere. In *Methodologies in Semantic Fieldwork*, ed. by Lisa Matthewson and Ryan M. Bochnak, 13-46. Oxford: Oxford University Press.

Bombi, Carla, and Joseph P. De Veaugh-Geiss. 2018. Quantitative data in the field: Two case studies on Akan. Paper presented at *Linguistic Evidence*, Universität Tübingen, 15-17 February.

Burton, Strang, and Lisa Matthewson. 2015. Targeted construction storyboards in semantic fieldwork. In *Methodologies in Semantic Fieldwork*, ed. by Lisa Matthewson and Ryan M. Bochnak, 135-156. Oxford: Oxford University Press.

Cable, Seth. 2019. Describing future eventualities in Tlingit: The storyboards Hawaii Trip and Imagining the Future. *Semantic Fieldwork Methods* 1(2):1-35.

Copley, Bridget. 2009. *The Semantics of the Future*. New York: Routledge.

Cover, Rebecca T., and Judith Tonhauser. 2015. Theories of meaning in the field: Temporal and aspectual reference. In *Methodologies in Semantic Fieldwork*, ed. by Lisa Matthewson and Ryan M. Bochnak, 306-349. Oxford: Oxford University Press.

Davidson, Kathryn. 2020. Is "experimental" a gradable predicate? In *NELS 50: Proceedings of the Fiftieth Annual Meeting of the North East Linguistic Society*, ed. by Mariam Asatryan, Yixiao Song, and Ayana Whitmal, 125-144. Amherst, MA: GLSA.

Davis, Henry, Carrie Gillon, and Lisa Matthewson. 2014. How to investigate linguistic diversity: Lessons from the Pacific Northwest. *Language* 90(4):e180-e226.

Davis, Henry, Carrie Gillon, and Lisa Matthewson. 2015. Diversity driven but cognitively constrained: Boas meets Chomsky (Response to commentators). *Language* 91(3):e127-e143.

Davis, Henry, and Lisa Matthewson. in press. St'át'imcets frustratives as not-at-issue modals. *Linguistics: An Inter-Disciplinary Journal of the Language Sciences*.

De Veaugh-Geiss, Joseph P. 2021. nà-Cleft (non-)exhaustivity: Variability in Akan. *Glossa: a journal of general linguistics* 6(1):137, 1-41. https://doi.org/10.16995/glossa.5698

Deal, Amy Rose. 2015. Reasoning about equivalence in semantic fieldwork. *Methodologies in Semantic Fieldwork*, ed. by Lisa Matthewson and Ryan M. Bochnak, 157-174. Oxford: Oxford University Press.

Dixon, R.M.W. 2010. *Basic linguistic theory*. Volume 1. *Methodology*. Oxford: Oxford University Press.

Driemel, Imke. 2019. Pseudo-noun-incorporation across languages. Doctoral dissertation, Universität Leipzig.

Dunlop, Britt, Suzanne Gessner, Tracey Herbert, and Aliana Parker 2018. *Report on the Status of B.C. First Nations Languages*. Brentwood Bay, BC: First Peoples' Cultural Council.

Featherston, Sam. 2007. Data in generative grammar: The stick and the carrot. *Theoretical Linguistics* 33:269-318.

Gibson, Edward, and Evelina Fedorenko. 2013. The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes* 28:88–124.

Grubic, Mira. 2015. Focus and alternative sensitivity in Ngamo (West-Chadic). Doctoral dissertation, Universität Potsdam. https://publishup.uni-potsdam.de/frontdoor/index/index/docId/8166.

Guntly, Erin. 2021. Yeah, I doubt it.' 'No, it's true.' How paradoxical responses impact the common ground. Doctoral dissertation, Vancouver: University of British Columbia.

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

Henrich, Joseph, Steven J. Heine and Ara Norenzayan. 2010. The weirdest people in the world?

Behavioral and Brain Sciences 33(2):61-135.
Hirayama, Yuto, and Lisa Matthewson. 2022. Temporal evidentials without tense. *Journal of Pragmatics* 193:173-188. https://doi.org/10.1016/j.pragma.2022.03.004.
Katz, Graham. 2012. Semantics in corpus linguistics. In *Semantics* (HSK 33.3), ed. by Claudia Maienborn, Klaus von Heusinger, and Paul Portner, 2859-2887. Berlin: Mouton de Gruyter.
Klecha, Peter. 2011. Optional and obligatory modal subordination. In *Proceedings of Sinn und Bedeutung 15*, ed. by Ingo Reich, Eva Horch, and Dennis Pauly, 365-379. Saarbrücken: Saarland University Press.
Klecha, Peter, Joseph Jalbert, Alan Munn, and Cristina Schmitt. 2008. Explaining why gonna precedes will in acquisition. In Supplement to *the Proceedings of the 32nd Boston University Conference on Language Development*, ed. by Harvey Chan, Enkeleida Kapia, and Heather Jacob, n.p., Boston University. http://www.bu.edu/bucld/proceedings.
Matthewson, Lisa. 2004. On the methodology of semantic fieldwork. *International Journal of American Linguistics* 70:369-415.
Matthewson, Lisa, Henry Davis, and Hotze Rullmann. 2007. Evidentials as epistemic modals: Evidence from St'át'imcets. *Linguistic Variation Yearbook* 7(1):201-254.
Morzycki, Marcin. 2021. Gradable adjectives and degree expressions. In The Wiley Blackwell Companion to Semantics, ed. by Daniel Gutzmann, Lisa Matthewson, Cécile Meier, Hotze Rullmann, and Thomas Ede Zimmermann, 1-32. Hoboken, NJ: John Wiley and Sons.
Perley, Bernard C. 2012. Zombie linguistics: Experts, endangered languages and the curse of undead voices. *Anthropological Forum* 22(2):133-149. DOI: 10.1080/00664677.2012.694170.
Phillips, Colin. 2010. Should we impeach armchair linguists? In *Japanese-Korean Linguistics vol. 17*, ed. by Shoishi Iwasaki, Hajime Hoji, Patricia M. Clancy, and Sung-Ock Sohn, 49-64. Stanford, CA: CSLI Publications.
Polinsky, Maria. 2010. Linguistic typology and formal grammar. In *The Oxford Handbook of Typology*, ed. by Jae Jung Sung, 650-665. Oxford: Oxford University Press.
Sprouse, Jon, and Diogo Almeida. 2013. The empirical status of data in syntax: A reply to Gibson and Fedorenko. *Language and Cognitive Processes* 28:222–228.
Sprouse, Jon, and Diogo Almeida. 2018. Setting the empirical record straight: Acceptability judgments appear to be reliable, robust, and replicable. *Behavioral and Brain Sciences* 40:e311.
Sprouse, Jon, Carson T. Schütze, and Diogo Almeida. 2013. A comparison of informal and formal acceptability judgments using a random sample from Linguistic Inquiry 2001-2010. *Lingua* 134:219-248.
Todorović, Neda, Michael Schwan and Lisa Matthewson. 2020. Compositionally deriving the future in Gitksan. Paper presented at *SULA 11: Semantics of Under-represented Languages in the Americas 11*, El Colegio de México and Universidad Nacional Autónoma de México, 4-7 August.
Tonhauser, Judith 2019. Managing data for semantic/pragmatic fieldwork. Ms., The Ohio State University / University of Stuttgart.
Tonhauser, Judith, and Lisa Matthewson. 2016. Empirical evidence in research on meaning. Ms., The Ohio State University and University of British Columbia.
Vander Klok, Jozina. 2019. Exploring the interaction of modality and temporality through the

storyboard Bill vs. The Weather. *Semantic Fieldwork Methods* 1(1):1-29.
Wasow, Thomas, and Jennifer Arnold. 2005. Intuitions in linguistic argumentation. *Lingua* 115:1481-1496.