Is *experimental* a gradable predicate?[*]

Kathryn Davidson

Harvard University

1.   Introduction

50 years of contributions to the North East Linguistic Society, as celebrated by these proceedings, make abundantly clear the progress that can be made toward understanding human language through methods of data collection that are often much simpler than those in other social sciences or humanities disciplines. Taking advantage of the fact that the linguist and their audience are, as humans, already in possession of some intuitions about the languages that they speak or sign, linguists have made great use of "armchair" introspective intuitions to revolutionize the way that we think, and report, about how language works and what is possible and not possible in human language (Chomsky 1959, Chomsky 1965). That there exist interesting linguistic arguments for which such intuitions can provide a full set of supporting data is not in doubt to this author, and will not be argued with here. However, it is one thing to say that there exist groundbreaking arguments for which armchair intuitions provide sufficient support; it is of course quite another to say that for any individual argument, armchair intuitions suffice.

Moreover, as linguistic theory has become more dependent on increasingly nuanced data, other "non-armchair" methods for data gathering have become increasingly common over these same 50 years of contributions to NELS, generally categorized into corpus data, fieldwork, and experimental data. This has lead to lively discussion of the necessity, or not, of controlled quantitative experiments to linguistic theory (see Sprouse and Almeida 2013, Sprouse et al. 2013, Gibson and Fedorenko 2013 for discussion within syntax and semantics). That debate takes place in the negative space of this one, focusing on the queston of whether linguistic theories should ever be built on data which has only been collected via armchair intuitions. As stated above, I take the position here that they can in some cases,

but my goal in this paper is to argue that the path forward should not always take the form of determining what questions deserve to be experimentally studied or not, but rather a rethinking 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 nonexperimental linguistics.

What precisely does it mean to argue against an experimental/non-experimental binary in the language sciences? Linguistics is certainly not the only academic field to differentiate between experimental and non-experimental approaches to its topic: there are neurobiologists running "wet" experimental labs to provide data to computational neurobiologists modeling the brain, and there are experimental physicists collecting data that is used by theoretical physicists who model the results; such a theoretical/experimental divide is not in any way uncommon in academic pursuits. What is perhaps unusual about Linguistics, and what I'm arguing should be reconsidered, is that the traditional "theoretical/experimental" alignment within Linguistics actually *doesn't* correlate well with a data/modeling divide: plenty of linguistics papers presented here at NELS and elsewhere analyze linguistic data that they are also presenting as new, and yet will not claim to have conducted any experiments. The more usual alignment in other fields, then, of (a) theoreticians who build models based on others' data contrasting with (b) data-gathering experimentalists simply doesn't accurately describe the field of Linguistics, where "data" can take the form of observing one's own linguistic behavior or informally checking with your colleague, and thus data collection is considered accessible without explicit experimental training. In fact, creating and presenting informally gathered data is often considered an important skill for purely "theoretical" linguists (consider an admiring comment that someone "came up with such a clever data point in that question period!")

How should we, as linguists, then think about when and how to really carry out "experiments"? What is, and isn't, an experimental linguistics investigation? **I want to argue that the question should not be when or where an experiment is warranted, but rather better consideration of which aspects of experimentation should apply** to a given piece of research. Under this view, *experimental* is a multi-dimensional, gradable predicate. I will lay out several dimensions of "experimentality" and explain how different studies make use of something "a little bit experimental" to collect a compelling set of data. These include:

- Controlled sampling of speakers/signers

- Controlled manipulation of context

- Controlled manipulation of linguistic features

- Controlled type of response by participants

- Quantitative data reporting

- Statistical data analysis

- "Open science" procedural and data transparency

Is experimental *a gradable predicate?*

Several, if not all, of these choices have been discussed in various methodological reflections in linguistics. For example, fieldworkers collecting data to test theories of formal semantics have focused explicitly on the value of controlled manipulation of several of these features, especially the context of evaluation (Bochnak and Matthewson 2015). Tonhauser and Matthewson (2018) clearly lay out desiderata for a fully explicit piece of semantic **data**, which they state must include (a) a context of evaluation, (b) a linguistic expression, (c) a response to the expression in that context, and (d) information about the language user who provided the response. These correspond exactly to four of the above properties of experiments: controlled context, controlled linguistic features, controlled type of response, controlled sampling of speakers/signers, plus general methodological transparency. Tonhauser and Matthewson (2018) argue that if one of these are missing, the data are less "stable, replicable, and transparent". The discussion I have in mind, then, has in part already been happening in semantic fieldwork and elsewhere, especially as regards accepting and *reporting* "data", and was explicitly noting that most of these features were borrowed from quantitative linguistic studies.

What I want to further argue is that there are many situations in which adding even more aspects of quantitative experimentation to what are typically not considered experimental studies will be beneficial, and on the flipside, that the nature of linguistic data is indeed such that semi-experimental investigations can accelerate our understanding of human language, in contrast to other social sciences where such valuable data points are less accessible. To be clear, the goal is not to condone poorly designed experiments! When a research question calls for a carefully designed quantitative experiment then one must be done, and there are massive resources available to learn the relevant methodology created both within linguistics, as well as borrowed from neighboring fields like psychology, neuroscience, and other cognitive sciences. Many, if not all, linguistics departments are finding it critical to provide their students with training in experimental practices, both for students' own projects and for their ability to consume and evaluate experimental work by others. The more interesting claim that I want to make here is that this training can also benefit linguists by elucidating practices common in quantitative experimentation which can be added to improve linguistic data collection of a more traditional sort.

As many an emerging linguist is reminded, the simplest experiment in linguistics is a basic minimal pair. Consider (1), presented in a context in which there is a familiar or uniquely salient paper, to an adult native speaker of English who is also a linguist (perhaps the author of the paper), and who is tasked with providing an acceptability judgment.

(1)   a.   Alex read the paper.
       b.   Alex read paper the.

In this case, the controlled feature is the linguistic expression. Most likely, an English speaking participant, asked to decide which of these expressions is an acceptable English sentence, will find (1)a acceptable and (1)b unacceptable. This is a single controlled variation along a single dimension (controlled manipulation of linguistic features), something I would be happy to call "a little bit experimental", or perhaps "the smallest bit experimental". Such basic mini-experiments are the bread and butter of the theoretical linguist,

although too often various controlled factors are not made explicit enough. For example, it's not uncommon to leave out details of the context of evaluation and the prosody when discussing the syntactic well-formedness in (1). Usually the lack of explicitness of the context, prosody, and other factors (e.g., co-speech gesture, speaker/addressee relationship, etc.) is intended to implicate that there exists no value for these variables in which the given judgment is different. Such an inference may be true in the case of highly robust data like that in (1), but commonly fails to hold under further investigation. For example, consider a similar pair in (2) : here the judgment is at first quite parallel between (2)a-(2)b, but a simple change in vowel quality, prosody, and perhaps a gesture of the hands changes the meaning such that what had been interpreted as an indefinite article is rather interpreted as the first letter of the alphabet, for a now grammatical sentence in (2)c.

(2)   a.   Alex read a paper.
       b.   Alex read paper a.
       c.   Alex read paper "a".

Here it's clear that pronunciation and context (i.e., if there are multiple papers in the preceding discourse which have been previously labelled "a", "b", etc.) matter. When a data point doesn't contain this information, it is open to a universal inference that *all* context, prosody, etc. will make the judgment hold, and this is rarely true. Quantitative experiments are of course in theory open to such fallacious inferences too, but they nearly always include more detail about these features in their reporting and in doing so limit these kind of incorrect universal inferences.

Quantitative studies always include information about participants, which is surprisingly often not made explicit in theoretical linguistics. This, in turn, leads readers of many linguistic papers to the frequently incorrect universal inference that there are no likely participants for whom the judgment would be different. At NELS 50, the conference at which this paper was presented, 31 of the abstracts (upon which acceptance for talks were based, and which were published in the online program) presented new data. That is, 31 papers were not entirely theoretical proposals based on existing data, and as such might be considered "experimental" in other scientific fields. Of these 31 abstracts presenting new data, 11 reported data from English and/or German, 7 reported data based on full-fledged quantitative "experiments", and 13 reported data from understudied languages. The 11 abstracts reporting data on English and/or German would all have been strengthened by being more explicit about the source of the reported data: in some cases it seems likely to be the authors', but not in all cases. However, these authors also might reasonably expect that the reader of the abstract could have an intuition or a way to verify the data given how many linguists speak these languages and that the functional language of the conference was English. On the other hand, readers of the 13 abstracts which reported data from understudied languages would be much less likely to be able to confirm the data themselves. Of these 13 abstracts, only 3 reported the source for their data, either from an existing grammar (Kouneli 2019), a native speaker author (Georgi and Amaechi 2019) or consultants and a survey (Driemel et al. 2019). More following in their footsteps, whether focusing on understudied languages or on English, would be an easy way to raise the quality of the data

upon which are theories are built, without any additional burden or training on the part of the linguists reporting data. In fact, after this paper was presented at the conference, several subsequent talks reported such participant information in explicit response to this call, showing how easy this is to change, even if the answer to naming the participants turns out to be as simple as "judgements from one native speaker linguist who we know". More generally, more complete transparency along other dimensions, including even for examples in a highly studied language like English, will go a long way toward holding linguistics to a higher standard without actually engaging in a qualitatively different kind of research gathering, merely a higher level of best practices in research reporting. In addition, adding in select aspects of experimentation to "theoretical" studies along the different experimental dimensions above can lead to stronger data more generally.

To make concrete what selecting these dimensions of experimentation can look like, I will briefly walk through three examples from my own lab and explain how (both unsuccessful and successful) choices were made in the hopes that they will highlight some possibilities in semi-experimental approaches to linguistic data gathering.

1.1   A little bit experimental in visual language

I will focus on three studies from our lab that involve sign languages, although the claim at the heart of this paper, that sometimes linguistic investigations benefit by being "a little bit experimental", is not at all particular to sign languages. However, there are several reasons that sign languages are especially illustrative for making this point, and I think it is not coincidental that sign language linguistics has been an area where approaches to data gathering have been especially under the microscope. We can categorize the issues and the change in practices in response to them in sign linguistics under categories we're already familiar with: variation in **participants**, **context**, and **response type**.

Regarding the gathering of data from a reasonable sampling of **participants** and reporting in a methodology section, sign language linguistics as a field suffers from having far too few deaf signers in a linguistics audience, or as reviewers of a journal article, who can act as a second level of varification of linguistic data that are presented, in the way that is often assumed for readers of a journal article to be highly fluent speakers of English. This is also true for many spoken languages that aren't well represented in academia, but it is additionally critical in sign languages because even among competent signers who are reviewing academic work in sign linguisitcs, the makeup of language users is so varied: only a small percentage of deaf children are born into families who already know a sign language, so for many deaf signers the langugae is learned later in life; similarly, hearing signers can be native signers born into deaf signing families but have varying competencies as heritage signers; yet again a huge number of sign language users (including deaf signers) learn a sign language as a second or later language. Whose linguistic intuitions "count" varies enormously by community and research group, and thus reporting detailed information about participants has been and continues to be crucial for evaluating sign language linguistics research. In addition, many sign linguists have some training in psychology or in education, where more detailed information about participants is regularly required, so including this information has been practice in this subfield of linguistics. This is not to

say that there aren't sign linguistics papers submitted that mention that "3 signers" were consulted for the data point, with no further information about their relationship to the language, or sometimes even nothing at all, but in general there are expected best practices for including this information in a way that is more similar to experimental studies.

Second, carefully controlling the **context** in sign languages often requires more detailed experimental design than providing one or two written sentences as background, and so providing a controlled context presented through pictures or video, in the same way to several signers as done in psycholinguistic and acquisition experiments, is not uncommon in sign linguistics. One reason that a few sentences aren't enough is that it could be critical to the acceptability of a sentence whether and how participants in a dialogue were situated with respect to each other and to their surrounding space, most especially (but not only) in instances of depictive language like that discussed below. This means that judgments of linguistic form need to be given very rich contextual backgrounds. In spoken language linguistics, one property of a context can be changed by changing minimal parts of a prior discourse: a clear example is the study of projective semantic context in Tonhauser et al. (2013). However, for sign languages, it can be less obvious how to diagnose a "minimal" change in context, given how much might be situational and analog (regarding space and depiction). Therefore, sign linguistics has favored designs that share more in common with psycholinguistic studies by using a set of controlled stimuli to elicit a response, for example the exact same picture/automatic presentation to provide the context for a linguistic judgment.

Finally, using more implicit **response types** borrowed from psycholinguistics instead of acceptability judgements has proven useful in sign linguistics. Judging the acceptability of a particular sentence in a given context requires high metalinguistic awareness of the language, a significant issue when there are far fewer opportunities for formal learning about ASL, and thus less formal training in providing metalinguistic judgements than in a language widely used in testing and education like English. In addition, it can be quite difficult to separate the sentence itself from the rest of the preceding and present context (e.g., arrangement of interlocutors to themselves and their environment). Finally, fieldwork in sign languages (outside of national sign languages like ASL) often occurs in the context of no shared language between the fieldworker and the consultant: due to systemic educational inequities for signers of non-national languages there may be no one fully fluent in both a sign language of fieldwork and in a language familiar to a researcher, so asking for acceptability judgments simply won't work; in this case, picture matching tasks or language production tasks (e.g., describing a picture using a storyboard, as done in many fieldwork contexts) will likely lead to more successful data gathering, even if it ends up looking "a little more" experimental and having a less direct route to acceptability judgments than the traditional fieldwork toolbox in spoken languages.

Of course, many of the properties of sign language linguistics that lead it to be "a little bit experimental" are also found in spoken language linguistics: a speaker population with high variation in experience with the language and a lack of training in metalinguistic judgments are certainly commonplace in most languages of the world, and languages with primarily oral traditions will have a harder time with a "minimal change" context when heavily embedded with prosody and gesture. It is precisely why, I suspect, linguists doing

fieldwork have in many ways collected data in methods that are "a little bit experimental", and why I hope that in sharing three case studies from sign linguistics, this paper can provide some guidance for similar semi-experimental work in any language.

2. Case study 1: Progress but a pitfall

As a first example, I'll discuss my own study on the expression of logical conjunction and disjunction in American Sign Language, (Davidson 2013), an investigation that took a more traditional approach of separating "theory" from quantitative experimentation, and which I argue could have benefitted from a more gradiently experimental investigaton.

The project grew out a simple question: what do scalar implicatures look like in sign languages? Davidson (2014) shows that when a pragmatic "scale" in American Sign Language has similar properties to a scale in spoken English, scalar implicatures arise at similar rates in both cases. This is the case for the lexical scale $< all, some >$: describing a scene in which all cans are red with an underinformative description like {CAN, SOME RED (ASL)/*Some of the cans are red* (English)} is rejected, since while it's true that there are some cans that are red, it implies that not all are red, since one could just as easily have said {CAN, ALL RED (ASL)/*All of the cans are red* (English)}. However, that paper doesn't discuss any ASL equivalent to a well-known scale in English $< and, or >$ because it wasn't clear that expressions for *and* and *or* even contrasted with each other in ASL: no particular sign for either of these expressions is used commonly as a connective in the way it is in English. Expressions in ASL which are glossed as AND and OR tend to be more restricted in use, to especially emphatic contexts or to translations of English.

Davidson (2013) set out first, then to understand more about the expression of conjunction and disjunction in ASL. The first section of the paper is based on data gathered from discussions with three Deaf, native signing, consultants, in regular meetings in order to describe different expressions for *and* and *or* and their distribution in a typical fieldwork procedure with acceptability judgements. In this stage, it became clear that conjunction (*and*) has several expressions, including two "general use coordinators" which can also be interpreted as disjunction (*or*); in both cases, nonmanual markings like eyebrow raising, mouth movements, and head tilting distinguish conjunctive from disjunctive interpretations.

What followed involved testing various hypotheses about the nature of these generalized connective expressions and how their disjunctive and conjunctive meanings arose from their parts. For example, are they underlyingly disjunction, as has been later suggested for other systems, including child language (Singh et al. 2016, Tieu et al. 2017b)? Are they underlyingly conjunctive, as in the system in Cheyenne (Murray 2017)? I ultimately argue in Davidson (2013) that their basic meaning is to simply collect alternatives, and that existential or universal quantification over those alternatives occurs in an additional piece moderated by factors like nonmanual marking or properties of the context. This was and is the best explanation I know of for the data pattern found in ASL, and later in other sign languages such as Japanese Sign Language (Asada 2019). It also raises an intriguing question: what does the compositional nature of these connectives mean for their pragmatic effects in terms of scalar implicatures? Early discussions of scalar implicature were based on armchair intuitions (Grice 1989, Horn 1984), but since then a very lively subfield of

experimental pragmatics has investigated differences in scalar implicature calculation by different populations of participants (e.g., children vs. adults)(Noveck 2001, Papafragou and Musolino 2003), different linguistic expressions (Degen and Tanenhaus 2016), or different experimental contexts or responses (Katsos and Bishop 2011, Jasbi and Frank 2017, Guasti et al. 2005). It was then natural to investigate how the scalar structure of conjunction and disjunction in ASL affects scalar implicature calculation via quantitative experimentation, so the second half of Davidson (2013) describes an experimental study.

In the experiment, scalar implicatures based on two scales in ASL were compared to each other, and compared to two scales in English. Deaf native signers of ASL participated in the ASL version of the experiment, and hearing native speakers of English participated in the English version of the experiment. The English scales were based on quantifiers (*some, all*) and connectives (*or, and*), while the ASL scales were based on quantifiers (SOME, ALL) and connectives, which were distinguished by the nonmanual movements (of face and body). This experiment varied, then, both the **participants** and the **linguistic expressions** between languages, while keeping constant both the **contexts** and the **response type** across languages. The results showed that scalar implicatures based on the quantifier scale (some, all) were similar in ASL and in English: this is a "control" that was kept constant between this and previous studies of implicatures in these two languages. However, the scalar implicatures based on *coordinators* were different in English and in ASL, where there were fewer implicatures calculated for the connectives than in English connectives, and also fewer than the ASL quantifiers. The conclusion, then, was that the combination of general connective plus nonmanual marking in ASL did not lead to the same kinds of pragmatic inferences (scalar implicatures) that other scales did, both other scales in ASL (with the same participants) and in English (with the kinds of meaning).

Fast-forward a couple of years, and someone comments that while it's true that nonmanual marking distinguishes disjunctive from conjunctive meaning in connectives in ASL, it's not always consistent between signers *which* of the two sets of nonmanual markings are used for which of the two meanings (conjunction vs. disjunction). The result of this is that the semantics need not change (determining conjunctive or disjunctive meaning is still contextual), but the conclusions from the experimental pragmatic component regarding scalar implicatures are much more difficult to interpret. Is it simply the nature of nonmanual marking to, perhaps like spoken prosody, allow more variation and also weaker pragmatic inferences, or are there fewer implicatures due to signer variation and thus weaker inferences? Clearly there are worthwhile followups to this particular study, but there is also a more general point important for our current context: in this case a "traditional" linguistic divide between working with language consultants to establish semantic facts, and then administering an experimental pragmatics study second, could have been improved by semi-experimental informal surveys throughout to include looped feedback about pragmatic inferences and more signers, without "waiting until the right moment" to conduct the full experiment. Note in this case that, like in many linguistic studies on underrepresented languages, the participant pool is more limited than the vast online recruitment techniques available for a language like English, which also meant that a series of several formal experiments at each stage (as is sometimes done in piloting/norming studies in psycholinguistics) is far less feasible.

Is experimental *a gradable predicate?*

In general the "establish theory then conduct experiment" structure to this study offers what is hopefully a cautionary tale of thinking about these methods as disjoint (and as an even more important point for a different day, of not including Deaf collaborators in the design of the first round of collection of new sign language data). Important new linguistic insights about semantic building blocks involved in general use coordination still stand, but without some kind of intermediate surveys an equally important insight was missing that affected the interpretation of the experimental pragmatic component. In this case there was variation among **participants** that semi-experimental approaches could have identified earlier, and this is a common reason that many theoreticians assume that one might do an experimental study. Quantitative studies in sociolinguistics and acquisition report data from varieties of participants to understand how and why there is variation, while quantitative studies in psycholinguistics report data from varieties of participants to infer conclusions about populations of which their participants are a representative sample. Theoretical linguists will often claim they are interested in what is simply possible in one person's mind (as a "complete" grammatical system), except we know that a single individual's language is fluid and changes across contexts and interlocutors, and we can end up making incorrect universal inferences about something as core as pragmatic inferences if we don't gather and report appropriate variation even among a single participant in different contexts. Moreover, participant variation is only the beginning: we emphasized earlier that other experimental dimensions can be just as valuable to import to semi-experimental approaches, and we turn to those next, where we focus on controlled variation of **context** and **linguistic expression**.

3.   Case Study 2: Telicity and sign language verbs

Psycholinguistic studies have long varied linguistic expressions in a controlled way while also varying participants, in an effort to understand the structure of the lexicon, so certainly varying linguistic forms is nothing new in experimental linguistic studies. At the other end of the spectrum, minimally varying a linguistic form, presented to a single speaker, is the definition of a minimal pair seen in any theoretical linguistic study. Understanding how to vary linguistic expressions in a semi-experimental way that covers a wider swath of the lexicon can be useful practice for theoreticians. In our lab, a study on the semantics of verbs in sign languages (Davidson et al. 2019) provides an illustrative example of a "little bit experimental" approach to testing a theoretical claim across a wide variety of forms in an understudied language.

The topic of this study is telicity in sign languages. An exciting proposal put forth in recent years in sign linguistics is the *event visibility hypothesis*, the proposal that event structure in sign languages is visible in linguistic forms: verbs expressing bounded (telic) predicates differ from those expressing unbounded (atelic) predicates by having clear boundaries in their form (Wilbur 2003, Wilbur 2008, Malaia et al. 2013). For example, the sign PLAY typically forms an atelic predicate and contains no constrained boundaries in its production (it involves a constant internal movement) but the sign STEAL typically forms telic predicates and has a clear boundary point in its form, which involves a single directed movement accompanying a handshape change: a "visible" manifestation of the semantic boundary.

Our project was inspired by this work, and grew from two foundational observations. First, it has long been noticed that there are paralellisms between the notions of countability in the nominal domain (e.g., mass/count distinction) and boundedness in the verbal domain (e.g., atelic/telic distinction) (Bach 1986); yet although many linguistic features (e.g., plurality, reference to substances vs. objects, etc.) track with the semantic properties of mass/count, no language marks that distinction morphemically across a lexicon. Along these same lines, we don't find (many, if any) examples of telicity morphemes in spoken languages, so finding telicity directly marked at all (much less "iconically" via a visible boundary) would be surprising. Our second observation was that several possible counterexamples to the event visibility hypothesis appear, at least on their face, to exist in a sign language familiar to us, and we wanted to understand how these fit in to the larger pattern (were these exceptions or part of a pattern themselves?), and what they mean for a theory of event visibility both within sign languages and outside of them (Strickland et al. 2015).

Our original two-part methodological plan was to conduct interviews with Deaf native signing consultants that would then inform an experiment in the second stage. The goal was to understand what kind of variation of form/meaning exists in the lexicon: not an exhaustive corpus/dictionary study of all lexical items, but rather a study with enough coverage that we could uncover systemtic counterexamples. For example, in the mass/count domain, many English nouns that refer to objects use count syntax (e.g. *cat, box*), and many that refer to substances use mass syntax (e.g. *water, milk*) but there are mass nouns with atomic reference (e.g. *furniture, cutlery*) and alternators that can do both (e.g. *stone, string*) (Barner and Snedeker 2005). Moreover, in spoken languages, semi-experimental approaches have proved highly fruitful for understanding the lexical variation (or uniformity) in mass/count expressions even in highly understudied languages (Lima 2014, Deal 2016). Since most discussions of telicity in sign languages discuss only a very small collection of verbs with regard to the event visibility hypothesis, we wondered: what kind of lexical variation do we find in form and telicity when it comes to sign language verbs?

We set out to create stable contexts and responses for a given participant, thus just varying the linguistic form. What became quickly clear after our first interview, however, is that contexts are slippery beasts for licensing verb forms in sign languages for several reasons. For one thing, many forms can be construed in a flexible way, just as the mass noun *water* can be coerced to count (e.g. *She gave me many* [containers/kinds of] *waters*). For another thing, many verbs can be modified depictively as event modifiers (Davidson 2015), so getting truly minimal pairs *in verbal/predicate forms* is more challenging than in other areas of the sign language lexicon (e.g. compared to nouns). Third, even non-depictive modification is not yet understood morphologically in sign languages: it wasn't clear to us what changes in a verb's form should count as a completely new form or not for even a basic verb like READ. This all lead to a much more interesting, and exciting, picture, but not at all one suited to creating an experiment after one or two consultations.

In the end, we settled for three in-depth interviews with Deaf native signers with lots of experience providing metalinguistic comments in order to investigate a reasonable variety of verbal forms, along the lines of Deal (2016) for mass/count in Nez Perce. This time, our research team already included a Deaf signer as a collaborator from the beginning. What we found was a strong trend supporting the event visibility hypothesis, plus important

	TELIC	ATELIC
ENDSTATE	STEAL, DIE, DESTROY, LEARN, READ_1, WRITE_1, DRIVE_1, TYPE_1, SKI_1, DRINK	DRINK, SLEEP, IMAGINE, THINK
NO ENDSTATE	SWIM, SKATE, LECTURE, PAINT, DANCE, BUILD	PLAY, STORYTELL, STUDY, BREATHE, READ_2, WRITE_2, DRIVE_2, TYPE_2, SKI_2, SWIM, SKATE, LECTURE, PAINT, DANCE

Figure 1: Event structure (telic/atelic) and verb form (endstate/no endstate) in several verbs in American Sign Language, from Davidson et al. (2019)

patterned exceptions. For one thing, we found evidence of a class of alternators, verbs like READ and DRIVE which have telic interpretations when their form has a boundary point, and atelic interpretations when it does not (Figure 1). This was, in a way, even stronger evidence for morphemic marking of telicity in ASL than had been given previously. On the other hand, we also found entire classes that didn't fit the generalization, such as verbs in the PAINT class which take the same (boundary-less) form but can have either telic or atelic interpretations.

The result is a study that is only "a little bit experimental", showing how careful variation of forms, contexts, and some variation in participants can lead to a better understanding of the phenomenon. We can measure it along several dimensions of being experimental. The context was controlled to a medium degree: we created a clear test case for determining both form and meaning, but did not report the entire context in our paper, as would have been ideal or is typically expected for an experimental paper. We had a high degree of controlled manipulation of the linguistic expression, since different forms and their meaning formed the primary purpose of our study. Our sample size of participants was low on the experimental spectrum but interviewing three signers and reporting details is higher than many theoretical studies (including past studies on the same topic). However, there was no quantitative data reporting outside of a table (Figure 1) and no statistical data analysis, and due to privacy reasons the full videorecordings are not available online. No one who regularly collects data via quantitative experimentation would call this an experimental study, and yet it borrowed some aspects from psycholinguistic studies to, at least from our perpsective, improve our understanding of the variation between signers and between lexical items, critical for understanding the nature of the event visibility hypothesis itself. In summary, probing a wide variety of predicates across a medium variety of signers provides:

- Controlled sampling of speakers/signers: *low-medium - three signers*

- Controlled manipulation of context: *medium*

- Controlled manipulation of linguistic features: *high*

- Controlled type of response by participants: *low*

- Quantitative data reporting: *low*

- Statistical data analysis: *low*

- "Open science" procedural and data transparency: *medium*

4. Case Study 3: Depiction and negation

Finally, a third study from our lab that provides a useful illustration of some benefits to a somewhat experimental approach is an investigation of the relationship between sign language classifier predicates and negation (Zlogar et al. 2019, Zlogar 2019). Classifier predicates in sign language are interesting for linguistic theory not only because they exhibit properties of verbal classifiers seen in spoken language (in which their form depends on the noun class of one of their arguments), but also because they are used for iconic purposes of depiction: the manner of window opening that is conveyed by a co-speech gesture in English scan be conveyed with a classifier predicate in ASL. That classifiers are more closely integrated into the grammar of sign languages than are co-speech gestures in spoken languages is clear, but precisely how their depictive elements enter into the semantics is an open question. Some have taken them to be primarily gestural (Cogill-Koez 2000), others entirely morphemic (Supalla 1983). Some have suggested that iconic functions occur throughout grammar, and classifiers is just one of many such places (Schlenker et al. 2013); others have suggested that classifiers obligatorily take a depiction, or demonstration, as an argument (Davidson 2015, Zucchi et al. 2011).

One way to approach the question of the linguistic structure of classifier depiction is to ask how the depictive elements in classifier predicates interact with other logical operators in the grammar, such as negation. We had observed that there were very few examples in published literature of classifiers in a negative sentence in American Sign Language. Of the few that did exist, the negation was actually expressed with a negative quantifier (e.g. NONE), and not sentential negation (Benedicto and Brentari 2004, Wood 1999). We wondered, then, how sentential negation with classifiers might be expressed and interpreted, and compare to that with negative quantifiers. To our knowledge, there had been no targeted investigations of the interaction of negation with the depictive content of classifier predicates. We therefore decided to compare depiction in sign languages with depiction in co-speech gesture accompanying spoken language.

Our team, which was comprised of one Deaf and two hearing collabators, created stimuli that paired a pictured context with a video that played a pre-recorded sentence (either in English or in ASL) (Figure 2). Participants were 7 (hearing) English speakers and 7 Deaf ASL signers, and the response was a binary acceptability judgment: is the sentence acceptable given the pictured context? The question we were interested in was whether the depictive element (size and shape and manner depictions) would be affected by the negative operator (which varied between negative quantifiers and sentential negation). In the example in Figure 2, the depiction is of the horizontal manner of window opening, and the description contains sentential negation.

Of seven participants in the English half of the study, all seven showed agreement across conditions by *not accepting* the English sentences that we had designed to test depictive co-speech gestures targeted by negation, across both kinds of depiction and types of negation.
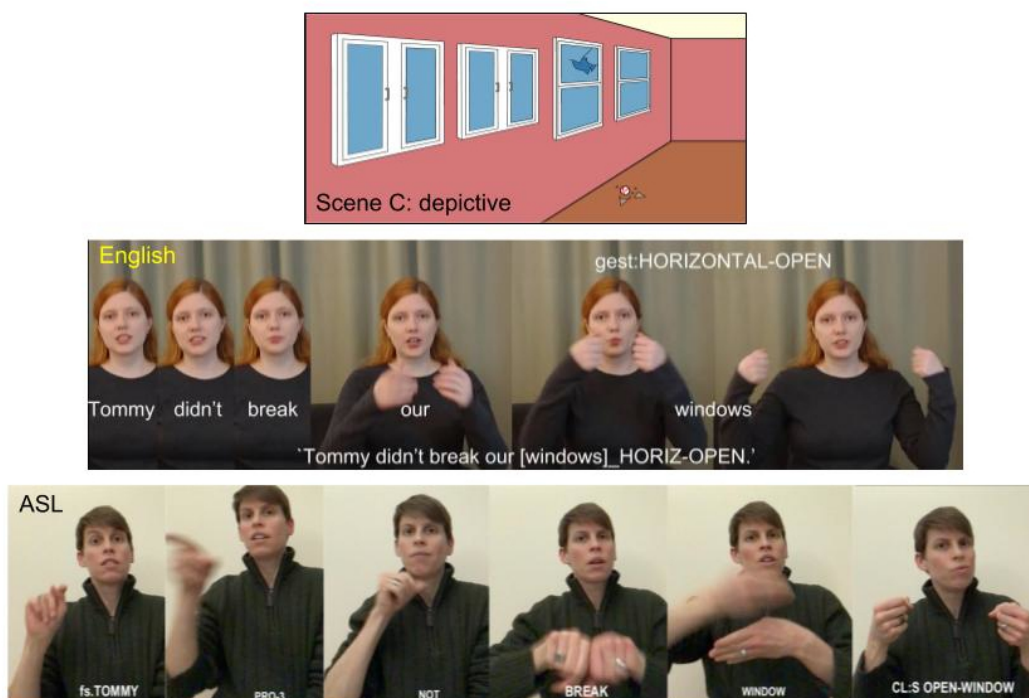
Figure 2: A context-setting scene (top), and sentences to be evaluated in that context in English (middle) and ASL (bottom), from Zlogar et al. (2019). The gesture HORIZ-OPEN is produced during the English text in brackets, namely *windows*.

Of the seven participants in the ASL half of the study, there was widespread agreement in exactly the opposite direction: six of seven *accepted* the ASL sentences which were designed to test the same depictions via classifier predicates targeted by negation, across both types of classifier predicates and types of negation. With robustly different patterns between the two languages, while maintaining highly controlled contexts, forms, and responses, we can conclude that these pieces (depiction and negation) interact in different ways in the two languages. It's not clear that further quantitative analysis is necessary with regard to this particular question, at least at this preliminary point: further investigation of more complex interactions, lexical items, participant variation (as in the one outlier in ASL) etc might call for a full experimental investigation, but we can already conclude much from this simple "semi-experimental" study. It is also similar to other semi-experimental studies that ask for acceptability judgements given carefully controlled contexts and linguistic expressions, across a small to medium sample of signers, which are not uncommon in sign language linguistics (e.g. Hübl et al. 2019) but could be easily used in any other languages. In summary, asking for carefully controlled stimuli-based acceptability judgments from a medium variety of signers provides:

- Controlled sampling of speakers/signers: *low-medium*

- Controlled manipulation of context: *high*

- Controlled manipulation of linguistic features: *high*

- Controlled type of response by participants: *high*

- Quantitative data reporting: *medium*

- Statistical data analysis: *low*

- "Open science" procedural and data transparency: *medium*

5.   Further dimensions of experimentation

As mentioned above, sign language linguistics has several reasons to focus especially intensely on methodology for data collection in theoretical investigations. I am in general in this paper leaving aside corpus methodologies, which have been increasingly used for some of the best work in sign language linguistics and can be combined with elicitation as well (Kimmelman et al. 2018). Another method used recently in sign language linguistics that is worth mentioning in the context of semi-experimental approaches is the "playback method" in which a participant produces several linguistic expressions/sentences (based on translation and/or in response to a prompt and context discussion) and then is asked, later, for an acceptability judgment of their own production (Schlenker et al. 2013, Kuhn 2016). This has primarily been done with the same signer, so in this case the **linguistic form** changes across different items within a contrast set/minimal pair, and so does the **response**: it begins as production and then ends as an acceptability judgment, while the signer remains constant. In practice this has advantages over more traditional methods of note-taking during a consultant meeting, since all of the data and judgments are recorded and available for replication studies, although it fails to provide information about a signer's changing judgments across different contexts: are they due to a changing mind of a speaker or a slightly different context of elicitation? In summary, the playback method provides:

- Controlled sampling of speakers/signers: *usually low*

- Controlled manipulation of context: *low*

- Controlled manipulation of linguistic features: *high*

- Controlled type of response by participants: *high*

- Quantitative data reporting: *medium*

- Statistical data analysis: *low*

- "Open science" procedural and data transparency: *high*

6.   A little bit experimental in gestural studies

Sign language linguistics may be a rich source of semi-experimental approaches, but there are of course also countless illustrative examples of research which is "a little bit experimental" in spoken languages, including and perhaps especially in the study of co-speech gesture. Sometimes these arise through similar issues to those in sign languages regarding fully explicit contexts and controlled stimuli for gesture studies: Ahn and Davidson (2018) investigate demonstrative expressions in spoken Korean and spoken English with and without co-speech pointing and Tieu et al. (2017a), Tieu et al. (2018), and Zlogar and Davidson (2018) use semi-experimental studies to investigate basic linguistic properties like presupposition projection in English co-speech depictive gesture. Research on gesture lends itself well to quantitative research in order to control both **context** and **form**, since we still don't understand well what aspects of a context are relevant to gestural production, and understanding a "minimal difference" in gestural forms is likely not even possible. In the case of Zlogar and Davidson (2018) this is especially apparent, as the study is essentially a gestural counterpart of Tonhauser et al. (2013)'s approach to projective content, and yet the gestural modality was investigated in a "more experimental" way in at least two dimensions (in participant sample which was easier to vary in English, and in the control of linguistic forms, which were pre-recorded).

7.   Discussion

In this brief space I've argued for a multi-dimensional gradable notion of "experimental" linguistics, and illustrated this specific proposal through a few case studies in sign language linguistics. One way to paraphrase this reframing would be to say that all data collection in linguistics is at least a little bit experimental. A possible outcome of this new perspective could be, then, to encourage all linguists to become at least somewhat familiar with experimental methodology, on the one hand, and on the other hand to acknowledge the type of rigorous data collection already being done by fieldworkers as at least somewhat experimental, even when it is not highly quantitative.

Why does all of this matter? First, I hope that it highlights the value of training graduate students in highly controlled **experimental design**, and also highlights how this is dissociable from the equally complex and valuable tools of quantitative **data analysis**. Many fieldworkers and theoreticians will have high need for training in experimental design for data collection, which can be shared with those getting trained in psycholinguistics, although only the latter will have the high need for quantitative training, which they may share with those focusing on corpus data. Second, I've witnessed several recent examples where linguistics departments have categorized "experimental linguistics" as a category, either in an academic job search, in a graduate training program, or in departmental policies about funding, etc. Understanding that those interested in linguistic theory should be interested in at least some, and perhaps all, aspects of experimentation for data collection makes this kind of category less useful than, say, training in psycholinguistics (which is a particular set of questions, not necessarily methods), or in quantitative data sampling and analysis.

As far as the benefits to research go, many pieces of data used in linguistic theory are less precisely controlled and reported than data used in quantitative linguistic studies thanks to a fundamental disagreement about the potential relevance of these factors. Theorists may argue that they are interested in linguistic competance and not performance (Chomsky 1965) but separating competence from performance is, even if possible in *theory*, not possible *a priori*: we can't know before even begining an investigation that, say, prosody or particular facets of a context may or may not interact with grammar, measured indirectly as a participant's response to a linguistic form; to the extent that we make that assumption, it is often a fallacy from our assumptions around language as text.

Related to the competance/performance issue and gradable notions of experimentation, a reader of this paper is also strongly encouraged to read more about *linking hypotheses* in research on meaning. What do we mean by a *linking hypothesis*? The idea is that we are interested in studying properties of the human mind, something which we'll never have direct access to. We can only infer properties about it via observation of various behaviors (of a human, of neurons, etc.) Any given research project needs to be explicit about the link between behavior and underlying meaning, and good resources about this within semantics have been provided by Jasbi et al. (2019), Waldon and Degen (2020), Qing et al. (2018), and Tonhauser and Matthewson (2018), all of which provide ways of thinking about (either quantitative or fieldwork) approaches to this foundational experimental question with a focus on the **response**: how do properties of the task relate to conclusions that we can draw about linguistic knowledge? Regarding student training and the separation of experimental design (for fieldworkers and quantitative experimentalists) and data analysis (for quantitative experimentalists and corpus linguists), we can explicitly separate a third skill that all students must have, namely linking linguistic behavior to linguistic theory.

Other dimensions of experimentation in linguistics featured in this paper include **lexical coverage** and **speaker/signer coverage**. In our study on telicity in ASL in Davidson et al. (2019), signers easily converged on their agreement of possible verb forms (what expressions were well-formed verb plus derivational morphology combinations) but there was much wider disagreement when it came to matching form to interpretation. A similar pattern was found for variation of form/meaning pairs among coordination plus nonmanual marking pairs (Davidson 2013). Understanding possible forms across morphemes and across signers turned out to be critical in a way that was easy to see when a semi-experimental methodology was conducted. Neither of these studies focus on development or change of language systems across time or space, but of course studies that do necessarily require sampling across these dimensions.

To the extent that this paper has ignored corpus linguistics it is unfortunate, because data from linguistic corpora comprise a large portion of data in both sign language linguistics (especially in Europe where there are more widely available corpora) and in written language linguistics, and there are two important ways in which corpus linguistics intersects with the current discussion. First, methodological transparency in data collection is a key property of experimental linguistics as well as corpus linguistics, and such a habit shouldn't be given exception for "theoretical" talks that actually do present new data. Second, studies in corpus linguistics tend to have a similarly high expectation for statistical models that quantitative experimental linguistics do. Those who collect data in both of

these fields should, in my opinion, acknowledge that a theoretician who looks through a corpus to find an example, or creates a tiny survey for field work, should not always be expected to do the same level of quantitative analysis that could have been done on languages with larger corpora or more easily recruited experimental participants.

8.   **Conclusions**

NELS 50 the conference was held in the invigorating environs of the MIT campus and brought together linguists from all over the world. In that moment, the mood was that linguistics had every possibility of growing and maturing in new exciting ways. Only months later, these proceedings were written and collected in the subsequent months during the COVID-19 global pandemic, with all similar in-person festivities of academic conferences cancelled for some time. Fieldwork and lab-based data collection are similarly affected, and may be again for new reasons yet to be known. It is my hope that the perspective I have advocated for here will, thus, be timely as well, where some of the deeply cherished in-depth personal consultations that fieldworkers have become used to can be supplanted by remote methods which make up in some dimensions like population variation and controlled context and form what they miss in deep consultation; similarly, labs may be restricted in psycholinguistic measures for at least some time, and I hope that this highlights the significant data that can be gathered with what may appear to be less quantitative experimental methods.

Going forward, we can ask each other whether a particular paper or project is primarily theoretical or experimental (or corpus-based, or fieldwork-based), or perhaps whether one's training is primarily in one or the other of these backgrounds, but in discussing the gradability of experimentation I hope it is clear that doing one without a clear view of the other is a challenge: theorists depend on data gathered in a rigorous way just like they do in any academic endeavour, and of course experimental design without a clear view of important theoretical questions is equally as problematic in linguistics as it is in any other field. It's not a practical approach to train everyone equally in these areas, but the goal in this paper is to argue that it is eminently practical to provide support via training and publication for a great deal of empirical work which is "a little bit experimental".

References

Ahn, Dorothy, and Kathryn Davidson. 2018. Where pointing matters: English and Korean demonstratives. In *Proceedings of North East Linguistic Society 48*, 15–24.

Asada, Yuko. 2019. General use coordination in Japanese and Japanese Sign Language. *Sign Language & Linguistics* 22:44–82.

Bach, Emmon. 1986. The algebra of events. *Linguistics and philosophy* 9:5–16.

Barner, David, and Jesse Snedeker. 2005. Quantity judgments and individuation: Evidence that mass nouns count. *Cognition* 97:41–66.

Benedicto, Elena, and Diane Brentari. 2004. Where did all the arguments go?: Argument-changing properties of classifiers in ASL. *Natural Language & Linguistic Theory* 22:743–810.

Bochnak, M Ryan, and Lisa Matthewson. 2015. *Methodologies in semantic fieldwork*. Oxford University Press, USA.

Chomsky, Noam. 1959. A review of BF Skinner's Verbal Behavior. *Language* 35:26–58.

Chomsky, Noam. 1965. *Aspects of the theory of syntax*. MIT Press.

Cogill-Koez, Dorothea. 2000. A model of signed language 'classifier predicates' as templated visual representation. *Sign language & linguistics* 3:209–236.

Davidson, Kathryn. 2013. 'And' or 'or': General use coordination in ASL. *Semantics and Pragmatics* 6:4–1.

Davidson, Kathryn. 2014. Scalar implicatures in a signed language. *Sign Language & Linguistics* 17:1–19.

Davidson, Kathryn. 2015. Quotation, demonstration, and iconicity. *Linguistics and philosophy* 38:477–520.

Davidson, Kathryn, Annemarie Kocab, Andrea D Sims, and Laura Wagner. 2019. The relationship between verbal form and event structure in sign languages. *Glossa: A journal of general linguistics* 4(1):1–36.

Deal, Amy Rose. 2016. Do all languages make countability distinctions? Evidence from Nez Perce. In *Proceedings of Sinn und Bedeutung*, volume 20, 180–197.

Degen, Judith, and Michael K Tanenhaus. 2016. Availability of alternatives and the processing of scalar implicatures: A visual world eye-tracking study. *Cognitive science* 40:172–201.

Driemel, Imke, Ahmet Bilal Özdemir, and Marie-Luise Popp. 2019. Adyghe cislocative: Canonical inverse but non-canonical PCC marker. Talk presented at NELS 50.

Georgi, Doreen, and Mary Amaechi. 2019. On two types of resumption in Igbo and the nature of islands. Talk presented at NELS 50.

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

Grice, H. Paul. 1989. *Studies in the way of words*. Harvard University Press.

Guasti, Maria Teresa, Gennaro Chierchia, Stephen Crain, Francesca Foppolo, Andrea Gualmini, and Luisa Meroni. 2005. Why children and adults sometimes (but not always) compute implicatures. *Language and Cognitive Processes* 20:667–696.

Horn, Laurence. 1984. Towards a new taxonomy of pragmatic inference: Q-based and R-based implicatures. In *Meaning, Form, and Use in Context*, ed. Deborah Schiffrin, 11–42. Georgetown University Press.

Hübl, Annika, Emar Maier, and Markus Steinbach. 2019. To shift or not to shift: Indexical attraction in role shift in German Sign Language. *Sign Language & Linguistics* 22:171–209.

Jasbi, Masoud, and Michael C Frank. 2017. The semantics and pragmatics of logical connectives: Adults' and children's interpretations of and and or in a guessing game. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society*, 576–581.

Jasbi, Masoud, Brandon Waldon, and Judith Degen. 2019. Linking hypothesis and number of response options modulate inferred scalar implicature rate. *Frontiers in psychology* 10:1–14.

Katsos, Napoleon, and Dorothy VM Bishop. 2011. Pragmatic tolerance: Implications for the acquisition of informativeness and implicature. *Cognition* 120:67–81.

Kimmelman, Vadim, Ulrika Klomp, and Marloes Oomen. 2018. Where methods meet: Combining corpus data and elicitation in sign language research. In *Proceedings of 8th Workshop on the representation and processing of Sign Languages: Involving the linguistic community*, 95–100.

Kouneli, Maria. 2019. The singulative of abundance: when one means a lot. Talk presented at NELS 50.

Kuhn, Jeremy. 2016. ASL loci: Variables or features? *Journal of Semantics* 33:449–491.

Lima, Suzi. 2014. The grammar of individuation and counting. Doctoral dissertation, University of Massachusetts- Amherst.

Malaia, Evie, Ronnie B Wilbur, and Marina Milković. 2013. Kinematic parameters of signed verbs. *Journal of Speech, Language, and Hearing Research* .

Murray, Sarah. 2017. Complex connectives. In *Semantics and Linguistic Theory 27*, 655–679.

Noveck, Ira A. 2001. When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition* 78:165–188.

Papafragou, Anna, and Julien Musolino. 2003. Scalar implicatures: experiments at the semantics–pragmatics interface. *Cognition* 86:253–282.

Qing, Ciyang, Daniel Lassiter, and Judith Degen. 2018. What do eye movements in the visual world reflect? a case study from adjectives. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, 2297–2302.

Schlenker, Philippe, Jonathan Lamberton, and Mirko Santoro. 2013. Iconic variables. *Linguistics and philosophy* 36:91–149.

Singh, Raj, Ken Wexler, Andrea Astle-Rahim, Deepthi Kamawar, and Danny Fox. 2016. Children interpret disjunction as conjunction: Consequences for theories of implicature and child development. *Natural Language Semantics* 24:305–352.

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, 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.

Strickland, Brent, Carlo Geraci, Emmanuel Chemla, Philippe Schlenker, Meltem Kelepir, and Roland Pfau. 2015. Event representations constrain the structure of language: Sign language as a window into universally accessible linguistic biases. *Proceedings of the National Academy of Sciences* 112:5968–5973.

Supalla, Ted Roland. 1983. Structure and acquisition of verbs of motion and location in American Sign Language. Doctoral dissertation, University of California, San Diego.

Tieu, Lyn, Robert Pasternak, Philippe Schlenker, and Emmanuel Chemla. 2017a. Co-speech gesture projection: Evidence from truth-value judgment and picture selection tasks. *Glossa: a journal of general linguistics* 2.

Tieu, Lyn, Robert Pasternak, Philippe Schlenker, and Emmanuel Chemla. 2018. Co-speech gesture projection: Evidence from inferential judgments. *Glossa: A journal of general linguistics 3(1): 109* 1–21.

Tieu, Lyn, Kazuko Yatsushiro, Alexandre Cremers, Jacopo Romoli, Uli Sauerland, and Emmanuel Chemla. 2017b. On the role of alternatives in the acquisition of simple and complex disjunctions in French and Japanese. *Journal of Semantics* 34:127–152.

Tonhauser, Judith, David Beaver, Craige Roberts, and Mandy Simons. 2013. Toward a taxonomy of projective content. *Language* 66–109.

Tonhauser, Judith, and Lisa Matthewson. 2018. Empirical evidence in research on meaning. *Ms., The Ohio State University and University of British Columbia* .

Waldon, Brandon, and Judith Degen. 2020. Modeling behavior in truth value judgment task experiments. *Proceedings of the Society for Computation in Linguistics* 3:10–19.

Wilbur, Ronnie. 2003. Representations of telicity in ASL. In *Proceedings from the annual meeting of the Chicago Linguistic Society 39*, 354–368.

Wilbur, Ronnie B. 2008. Complex predicates involving events, time and aspect: Is this why sign languages look so similar. *Theoretical issues in sign language research* 217–250.

Wood, Sandra. 1999. Semantic and syntactic aspects of negation in ASL. Master's thesis, Purdue University.

Zlogar, Christina. 2019. An experimental pragmatic investigation of depictive co-speech gestures. Doctoral dissertation, Harvard University.

Zlogar, Christina, and Kathryn Davidson. 2018. Effects of linguistic context on the acceptability of co-speech gestures. *Glossa: A journal of general linguistics 3(1)* 1–28.

Zlogar, Christina, Kate Henninger, and Kathryn Davidson. 2019. Negating depictive modifiers in sign language and in co-speech gesture. Poster at the LSA Annual Meeting, New York, NY, https://osf.io/nf76b.

Zucchi, Sandro, Carlo Cecchetto, and Carlo Geraci. 2011. Event descriptions and classifier predicates in sign languages. *Presentation FEAST in Venice* 21.

Kathryn Davidson
kathryndavidson@fas.harvard.edu