Empirical evidence in research on meaning[*]

Judith Tonhauser[•] and Lisa Matthewson[◦]

[•]The Ohio State University, *judith@ling.osu.edu*
[◦]University of British Columbia, *lisa.matthewson@ubc.ca*

November 7, 2015

Abstract

Empirical evidence is at the heart of research on natural language meaning. Surprisingly, however, discussions of what constitutes such evidence are almost non-existent. The goal of this paper is to open the discussion by advancing a proposal about the nature of empirical evidence in research on meaning. Our proposal has three parts. First, we propose that a piece of data in research on meaning consists of a linguistic expression, a context in which the expression is uttered, a response by a native speaker to a task involving the expression in that context, and information about the native speakers who provided the responses. Second, we argue that some response tasks, including acceptability and implication judgment tasks, are better suited than others (e.g., paraphrase and translation tasks) for yielding robust, replicable and transparent pieces of data. Finally, we propose that empirical evidence for a hypothesis about meaning consists of — depending on the hypothesis — a positive piece of data, a negative piece of data, or two pieces in minimal pair form, together with a statement about how the pieces of data provide support for the hypothesis.

1 Introduction

Research on meaning has been thriving for decades. However, even though empirical evidence for hypotheses about meaning is at the very heart of this research, there is almost no discussion in the literature of what constitutes such evidence. Introductory textbooks and handbook articles also largely remain silent on the matter. This paper begins to fill this gap by discussing the nature of empirical evidence in research in

[*] For helpful discussions and constructive comments, we thank Ryan Bochnak, Emmanuel Chemla, Henry Davis, Judith Degen, Ashwini Deo, Thomas J. Heins, Angelika Kratzer, Kilu von Prince, Agata Renans, Craige Roberts, Kevin Scharp, Mandy Simons, Hubert Truckenbrodt, Bob van Tiel, as well as the students in Tonhauser's *Semantic fieldwork methods* courses at the 2015 Linguistic Institute at the University of Chicago and at the University of Georgia in Athens. None of these individuals necessarily share the views we advance here. We are very grateful to our past and present St'át'imcets, Gitksan and Paraguayan Guaraní consultants for sharing their knowledge of their languages with us. All previously unpublished Gitksan data in this paper are from Barbara Sennott, Vincent Gogag and Hector Hill; ha'miiyaa! We also thank Joash Gambarage for his assistance with the bibliography. Tonhauser gratefully acknowledges fellowship support from the American Council of Learned Societies (2013/14) and from the Alexander von Humboldt Foundation (2014/15), and research funding from the National Science Foundation (BCS 0952571, 2010-2015). Matthewson gratefully acknowledges support from SSHRC (#410-2011-0431) and from the Jacobs Research Fund.

semantics and pragmatics. The paper advances a three-part proposal about the nature of empirical evidence, specifically about what a piece of data is, which tasks native speakers can most usefully be asked to perform, and which types of (minimal pairs of) data provide evidence for which types of hypotheses about meaning.

As we show throughout the paper, there are many different practices in contemporary research on meaning regarding the components of a piece of data, the tasks posed to native speakers, and the use of minimal pairs in providing empirical evidence. This heterogeneity suggests that it is time for a discussion and comparison of these different practices. In this paper, we hope to kick off a collaborative process of developing consistent standards for empirical evidence about meaning. As we discuss in section 2 where we review prior literature on empirical evidence, the proposal we advance is heavily informed by our and our colleagues' experiences in conducting research on languages we do not speak natively. In this sense, this paper expands on Matthewson's (2004) discussion of semantic/pragmatic fieldwork methodology. Furthermore, some parts of our proposal are already implemented in quantitative research on meaning. This paper thus synthesizes and builds on insights from various strands of research on meaning in discussing and developing standards for what counts as empirical evidence. We also intend for this paper to be a resource for those starting to undertake research on meaning.

Our three-part proposal focuses on issues that are fundamental to any research project involving empirical evidence about meaning, regardless of whether the evidence is collected through one-on-one elicitation with native speakers (a.k.a. "fieldwork"), through the researcher's judgments about utterances in their language (a.k.a. "introspection") or through quantitative research (a.k.a. "experiments") using offline measures.[1] We propose in section 3 that a piece of data has four parts: a linguistic expression, a context in which the expression was uttered,[2] a response by a native speaker to a task about that expression uttered in that context, and information about the speakers who responded. Second, we argue (section 4) that some response tasks, including acceptability and implication judgment tasks, are better suited than others (e.g., paraphrase and translation tasks) for yielding robust, replicable and transparent pieces of data. Our third claim (section 5) is that empirical evidence for a hypothesis about meaning consists of one or more pieces of data (possibly in minimal pair form) together with a statement about how the pieces of data provide support for the hypothesis about meaning. We show that different types of minimal pairs provide evidence for different types of hypotheses about meaning. The paper concludes in section 6 with a summary.

2 Previous discussions of the nature of empirical evidence

Empirical evidence for a particular hypothesis about meaning cannot be directly read off of an expression. Rather, which meaning a speaker conveys when uttering an expression is only indirectly revealed, for example through responses to tasks about the utterance in context (for this point, see also Matthewson 2004,

[1] Empirical evidence may also come from corpora. An expression attested in a corpus may, for instance, constitute a positive piece of data under the assumption that the expression is implicitly judged to be acceptable. However, pieces of data collected from corpora have a statistical quality since corpora may include errors and corpora need not include all acceptable linguistic expressions; for discussion, see de Marneffe and Potts to appear. Since our focus in this paper is on empirical evidence involving native speakers' responses, we ignore how empirical evidence can be established through corpus studies; see e.g., Kennedy and McNally 2005, Deo 2012 and Degen 2015 for illustrative examples of how corpus data can inform research on meaning.

[2] The term 'uttered' includes cases in which the linguistic expression was spoken, signed or written.

Bohnemeyer 2015, Deal 2015). Although textbooks in semantics and pragmatics are an excellent resource on a wide variety of empirical phenomena and theoretical approaches at the heart of research on meaning, none of them contains a substantial discussion of the nature of empirical evidence, let alone a comprehensive guide to semantic and pragmatic methodology.[3] This lacuna is particularly surprising given that many textbooks point out the central importance that empirical evidence plays in scientific research on meaning. Dowty et al. (1981:2), for example, write that "[i]n constructing the semantic component of a grammar, we are attempting to account [...] for [speakers'] judgements of synonymy, entailment, contradiction, and so on". Larson and Segal (2005:9) assert that "[s]emantic facts...are verified by the judgments of native speakers" and Hurford et al. (2007:7) point out that "[n]ative speakers of languages are the primary source of information about meaning". Cruse (2011:15) proposes that "native speakers' intuitions are centre stage, in all their subtlety and nuances: they constitute the main source of primary data". And Chierchia and McConnell-Ginet (2000:5f.) call speakers' judgments "the core of the empirical data against which semantic theories must be judged". Thus, overall, semantics/pragmatics textbooks acknowledge the central importance of empirical evidence, but provide no systematic discussion of what constitutes such evidence.

Volumes about research methods, including fieldwork methods, also fail to discuss what constitutes empirical evidence in research on meaning.[4] Beyond the lexicographic realm, semantic/pragmatic topics are rarely discussed. Rather, these works focus on topics such as the practicalities of collecting data (funding, recording equipment, databases and archiving, etc.), transcription methods, ethical issues and preparation with speakers and communities, broad categories of data-collection tasks (translation tasks, judgment tasks, text collection, etc.), and qualitative and quantitative data analysis. They also make suggestions about specific elements to elicit in the fields of phonetics, phonology, morphology, syntax, and lexicography. A number of these resources discuss tasks that native speakers can or should be asked to perform, and these discussions relate — albeit implicitly and indirectly — to what we argue is a component of a piece of data, namely a native speaker's response to a task.[5] However, these resources do not discuss in any detail what constitutes empirical evidence in research on meaning. And although there is frequently mention of the elicitation of minimal pairs in these resources, these seem to be always invoked in the context of phonetics or phonology, not of research on meaning, where minimal pairs are more complex, as we discuss in section 5 (e.g., Crowley 1999:110, Bowern 2008:38, Chelliah and de Reuse 2011:258). An exception to the general absence of discussion of the nature of empirical evidence is Beavers and Sells 2014. In their presentation of how to develop and support hypotheses in phonology, morphology and syntax, they define a piece of data as consisting of a linguistic expression and a native speaker judgment (p.398f.). We argue in section 3 that a piece of data in research on meaning has two additional parts, namely a context and information about the

[3] The works on which we base this claim are Dowty et al. 1981, Hurford et al. 2007, Frawley 1992, Cann 2007, Lyons 1995, Heim and Kratzer 1998, de Swart 1998, Chierchia and McConnell-Ginet 2000, Allan 2001, Portner 2005, Larson and Segal 2005, Saeed 2009, Riemer 2010, Cruse 2011, Elbourne 2011, Kearns 2011, Zimmermann and Sternefeld 2013 and Jacobson 2014.

[4] We base this claim on Samarin 1967, Kibrik 1977, Payne 1997, Vaux and Cooper 1999, Newman and Ratliff 1999, Crowley 1999, Bowern 2008, Chelliah and de Reuse 2011, Thieberger 2011, Sakel and Everett 2012 and Podesva and Sharma 2014.

[5] Chelliah (2001:158), for example, proposes "to take sentences from texts, create minimal pairs or sets by substituting words or morphemes, and then ask consultants what the sentence meant once the change had been carried out". Bowern (2008:103) likewise suggests that researchers ask native speakers to discuss whether a sentence can have particular meanings. It is not clear which specific response tasks these authors advocate for in exploring meaning. See section 4 for characterizations of tasks.

speaker(s) that provided the judgment.

Works specifically devoted to the methodology of research on meaning have only begun to appear within the past decade, primarily from authors collecting data through one-on-one elicitation with speakers of languages not spoken natively by these authors. The handful of available resources includes Matthewson 2004, 2011b, Hellwig 2006, 2010, Krifka 2011, Tonhauser 2012, Tonhauser et al. 2013 and the papers in Bochnak and Matthewson 2015. Several of these works already make points that we wish to reinforce in this paper and integrate into a general discussion of the nature of empirical evidence in research on meaning. For example, the importance of presenting a context as part of a piece of data, which we argue for in section 3, is pointed out in Matthewson 2004 and Cover and Tonhauser 2015. Targeted discussions of the role of translations and native speaker responses in providing empirical support for a hypothesis are provided in Matthewson 2004, Deal 2015 and Bohnemeyer 2015. This literature also includes diagnostics for investigating particular semantic/pragmatic topics that can be reliably applied with native speakers without theoretical training (see e.g., Tonhauser 2012 on not-at-issueness, Tonhauser et al. 2013 on projective content, and the papers in Bochnak and Matthewson 2015 on a variety of topics). We hope to bring the advances made in this literature about empirical evidence in research on meaning to the attention of the wider community.

Like fieldwork-based research, quantitative research is also a comparatively recent development in research on semantics and pragmatics. Quantitative research on meaning builds on the principles of experimental design, methodology, and quantitative analysis used in research in the cognitive and social sciences, and some parts of the proposals we advance here are already established practice in quantitative research on meaning. For instance, quantitative research already considers speakers' responses and information about the response task as components of a piece of data. Such research also typically involves minimal pairs simply by virtue of the fact that such research compares responses to one piece of data to responses to another, minimally different piece of data. Quantitative research also engages in discussions about suitable experimental designs, including the tasks that speakers are asked to respond to (for an example, see Geurts and Pouscoulous 2009). With this paper, we hope to engage the wider community of researchers in a discussion about what counts as empirical evidence about meaning.[6,7]

3 Pieces of data in research on meaning

In this section, we propose that a piece of data in (offline, response-based) research on meaning has four components: a linguistic expression, a context in which that expression is uttered, a response to a task about the utterance of that expression in that context, and information about the speakers who responded. Our objective in making this proposal is for pieces of data that inform theories of meaning to be *robust*,

[6] We limit ourselves to research on meaning conducted through offline measures, to the exclusion of, e.g., response time or eye movement measures, and to quantitative research on meaning comprehension, to the exclusion of research on production.

[7] An important question is whether some research methodologies provide more robust support for hypotheses than others, e.g. by virtue of relying on larger numbers of speakers, larger numbers of pieces of data, and quantitative analysis. For debate see Jacobson ms. and, primarily in the domain of syntax, e.g. Wasow and Arnold 2005, Gibson and Fedorenko 2010, 2013, Culicover and Jackendoff 2010, Sprouse et al. 2013, Davis et al. 2014. We sidestep this question here since the questions addressed in this paper — namely, which components make up a piece of data and how pieces of data provide evidence for hypotheses — arise regardless of which methodology is used to collect the pieces of data.

replicable and *transparent*. After characterizing the four components of a piece of data in section 3.1, we argue in section 3.2 that pieces of data that include the four components are more likely to be robust (by controlling for the context- and speaker-dependency of natural language interpretation), replicable, and transparent (by making fully explicit the piece of data that supports the hypothesis).

3.1 The four components of a piece of data

A piece of data in research on meaning is complex. The four components of a piece of data are characterized in the following subsections.

3.1.1 The context of a piece of data

The interpretation of natural language expressions is context-dependent. The utterance context, i.e., information about the speaker, the addressee(s), and the time and the location of the utterance, plays a role e.g., in the interpretation of deictic expressions like the English pronouns *I* or *you*, which denote the speaker and the addressee(s) of the utterance. Prior linguistic context, i.e., utterances previously made by the interlocutors, is involved in interpreting the referent of the English definite noun phrase *the cup* as the cup introduced in the first sentence in the two-sentence discourse *Joan dropped a cup and a spoon. The cup broke.* The context also includes information about the structure of the discourse that a linguistic expression is part of (e.g., Roberts 2012), such as information about the topic of conversation (also called the question under discussion) as well as the goals and intentions of the interlocutors. For instance, a speaker who utters *It's raining* intends a different meaning depending on whether the topic of conversation was whether to go for a walk (in which case the speaker may be signaling unwillingness to go) or whether to water the yard (in which case the speaker may be signaling that it is not necessary to water the yard).

Given the complexity of the context that plays a role in natural language interpretation, it is clear that the context that is presented as part of a piece of data typically cannot be the entire context in which the linguistic expression of the piece of data is uttered and interpreted. Rather, the context of a piece of data only captures a very limited set of features of the context in which the expression is uttered and interpreted, because resources are limited: e.g., the cognitive capacities of the speakers who have to understand the description of the context, the time it takes to present the context, or the space in a publication. The features of the context that are included in a piece of data are those that the researcher hypothesizes to be relevant for the current investigation. For example, the context of B's utterance in (1), from Hausa, is a single question that specifies the relevant individuals (Audu and Binta) and a topical time (yesterday, when the addressee called them). The context of B's utterance in (2), from Mbyá Guaraní, consists of a question inquiring about an individual, together with a description of the situation in which the question is uttered.[8]

[8] We follow the Leipzig glossing conventions (*https://www.eva.mpg.de/lingua/resources/glossing-rules.php*) to gloss our unpublished data; published examples from other authors are presented as published. The following additional glosses are used: ¬PPS = ¬p in projected set, A = series A cross-reference marker, ADHT = adhortative, ANA = anaphoric expression, ATTR = attributive, BDY = information structure boundary, CF = counterfactual, CIRC.POSS = circumstantial possibility modal, CL.CNJ = clausal conjunction, DM= determinate marker, II = series II pronoun, INFER = inferential evidential, MUST = necessity modal, PRON = pronoun, PROSP = prospective aspect, QUDD = Question Under Discussion downdate, SNV = sensory non-visual evidential, TOP = topical object marker.

(1) A: “What were Audu and Binta doing yesterday when you called them?”

B: Su-nā̀ màganā̀.
3PL-CONT talk

‘They were talking.’ (Mucha 2013:388)

(2) Context: A is visiting B’s community. A notices a man who is addressing a small group of villagers; he asks:

A: Mava’e pa kova’e ava?
who Q this man
‘Who is this man?’

B: Ha’e ma ore-ruvicha o-iko va’e-kue. Aỹ, porombo’ea o-iko.
ANA BDY 1.PL.EXCL-leader 3-be REL-PST now teacher 3-be
‘He was our leader. Now, he is a teacher.’ (Thomas 2014:394f.)

The example in (3) illustrates a context which establishes information about the prior discourse structure. The linguistic expressions in (3b) are uttered and judged in the context of the discourse in (3a).

(3) Rojas-Esponda 2014:8

a.	i.	*A: Möchtest du ein Glas Wein?*	A: Do you want a glass of wine?
	ii.	*B: Nein, Danke.*	B: No, thank you.
	iii.	*A: Hättest du gerne ein Bier?*	A: Would you like a beer?
	iv.	*B: Nein.*	B: No.
b.	i.	*B: #Ich möchte überhaupt kein Bier.*	B: #I want *überhaupt* no beer.
	ii.	*B: Ich möchte kein Bier.*	B: I want no beer. (I don’t want beer.)

The context of a piece of data may also be used to establish facts about the world, e.g., who slapped who in the two contexts in (4a).

(4) Cable 2014:2

a. Reflexive and Reciprocal Scenarios

i. Reflexive scenario: Each boy slapped himself. Dave slapped himself. Tom slapped himself. Bill slapped himself.

ii. Reciprocal scenario: Each boy slapped some other boy. Dave slapped Tom. Tom slapped Bill. Bill slapped Dave.

b. French reflexives and reciprocals with plural antecedents

i. Les étudiants se sont frappés.
the students REFL AUX slap
‘The students slapped themselves.’
Judgment: Can truthfully describe both [(4ai,ii)].

ii. Les étudiants se sont frappés l'un l'autre.
the students REFL AUX slap the.one the.other
'The students slapped each other.'
Judgment: Can truthfully describe only [(4aii)].

Given that the context of a piece of data captures features of the context that the researcher hypothesizes to be relevant for the particular investigation, there are no hard and fast rules about which features of the context to include. Of course, it may turn out later that some feature of the context was important for the particular investigation, but was not appropriately controlled for, or that some other feature of context was not, ultimately, relevant, but was included nevertheless. In such cases, subsequent investigation builds on the previous investigation by adapting the context of the piece of data.

As discussed in Matthewson 2004, AnderBois and Henderson 2015 and Bohnemeyer 2015, the context may be described to the speakers in the language under investigation or in the contact language; it may also be acted out, drawn or presented in writing. In publications, the context of a piece of data may be presented in the language of the publication (e.g., English), or in the language under investigation, as in (2), especially when linguistic properties of the language in which the context was presented are relevant to the hypothesis to be supported. Ideally, the context of a piece of data presented in a publication is identical to the context that was used during data collection. In practice, this is not always feasible, e.g., when the context was presented to the speakers in a language other than the language of the publication or when the context was acted out. When the context was presented in slightly different ways to different speakers, only one of those variants is presented in the publication, under the hypothesis that essential features of the context remained the same across the speakers.

3.1.2 The linguistic expression of a piece of data

The linguistic expression of a piece of data in research on meaning can be any linguistic expression that a native speaker of the language of the expression can write, sign or verbally utter in the context of the piece of data and give a response to. Although much research on meaning involves pieces of data with declarative sentences as the linguistic expression, as in the examples in (1), (3) and (4), other possible linguistic expressions are sentences in the interrogative or imperative moods, multi-sentence utterances as in (2), or sub-sentential expressions as in (5B).

(5) A: Who smokes?
B: Only John. (Coppock and Beaver 2014:401)

When a linguistic expression was signed or spoken, but is reported in writing, information about the prosodic realization of the utterance is typically not reported, except for utterances in tone languages or when the prosodic realization of the utterance is relevant to the hypothesis under investigation. In the latter case, it is customary to represent (the stressed syllables of) prosodically prominent expressions in small caps or capital letters (e.g., *Jack only drinks HOT coffee* indicates that the adjective is prosodically prominent). Written representations of linguistic expressions (typically, implicitly) adopt the hypothesis that the prosodic realization of the linguistic expression is not relevant to the hypothesis under investigation, or only relevant insofar as the relevant prosodic properties can be indicated with capital letters.

3.1.3 The response task and response

A linguistic expression together with a context in which it is uttered does not yet make for a piece of data in research on meaning. What is missing — as evidenced also by the quotes from the textbooks given in section 2 — is a native speaker's response, e.g., a judgment that the expression is acceptable in the context, a judgment that the expression is false in that context, or a translation of the expression into a different language. As Bohnemeyer (2015) puts it: "The response is a communicative action in the broadest sense. It may be a target language utterance, a contact language translation, a metalinguistic judgment, or any nonlinguistic action that solves the task, for example by pointing out a possible referent, demonstrating an action that would instantiate a given description, etc." (p.20).

Even when we limit our attention to offline, response-based research, research on meaning is conducted using a wide variety of response tasks, including acceptability judgment tasks, implication judgment tasks, translation tasks and paraphrase judgment tasks.[9] (We introduce and discuss these tasks in detail in section 4.) We refer to these tasks in the plural form since many of them can be implemented in several different ways. For instance, acceptability judgment tasks can differ in which specific question is asked (e.g., *Does this utterance sound good to you?* or *Is this utterance appropriate in this context?*) and in the response option provided to the native speaker (e.g., forced-choice binary responses, responses on a Likert scale, or magnitude estimations; see Schütze and Sprouse 2014 for an overview). Given the large variety of response tasks in research on meaning, the response component of a piece of data can take many forms, including 'yes', '3 out of 5', 'probably not' or 'Jane didn't read all the books', depending on the specific response task used. It follows that a speaker's response to a linguistic expression can only be understood in relation to the task that was used to elicit the response. For instance, it is only possible to understand whether a 'yes' response means that the speaker judges the example to be acceptable, unacceptable, or true if the particular response task that was used is identified. It thus also follows that a piece of data includes a speaker's response as well as information about the response task to which the speaker responded.[10]

Works reporting results from quantitative research typically include information about the response task in a methods section. In works that present pieces of data collected through introspection or one-on-one elicitation, such information is sometimes included as part of the piece of data. For instance, the piece of data in (6), from Hausa, includes information about the linguistic expression, the context, and also the question which was posed to the Hausa speakers (as confirmed by Anne Mucha, p.c.).

(6) Context: For lunch, Hàwwa cooked beans and ate them. Audu is cooking beans for dinner right now. Is it appropriate to say:

[9] See Krifka 2011 and Bohnemeyer 2015 for broad overviews of response tasks and methods in research on meaning that includes online measures and corpus-based research.

[10] When responding, speakers in one-on-one elicitation sometimes volunteer comments. These comments can provide clues about the meaning of the expression as well as reveal what the actual judgment is, as discussed in Matthewson 2004 and Matthewson 2015. As far as we know, the practice of including relevant comments as part of the data originated with Matthewson 1999.

#Hàwwa dà Audu sun dafà wākē yâu.
Hàwwa and Audu 3PL.COMPL cook beans today

Intended: 'Hàwwa and Audu cook/cooked beans today.'
Comment: The reading is not suitable for Audu. (Mucha 2013:385)

Other researchers opt to describe the type of judgment that was elicited and the speakers' responses in the text preceding the piece of data. In general, response tasks may be described in varying levels of detail, as summarized in (7):

(7) **Description of the response task**
Useful information about a response task includes

a. the instructions given to the native speaker about the response task,

b. the specific question posed to the native speakers,[11]

c. how the linguistic expression, the context and the question were presented to the speakers (e.g., in writing or verbally), and

d. the response options given to the native speakers, including information about whether the response was given verbally, in writing, or through some other means.

3.1.4 Information about the native speakers who responded

It is generally acknowledged in linguistic research that native speakers of a given language may give different responses to the same prompt. Native speakers may disagree, for instance, about whether a particular utterance is appropriate in a particular context. Variation in speakers' responses can be due to a variety of factors, including age, gender, dialect, socio-economic class, linguistic training, etc. Since native speakers may vary in their responses, information about the speakers that provided the responses for a particular piece of data are an integral part of the piece of data. In addition to information about the speakers' language background, age, linguistic training, etc., information about the number of speakers that provided judgments is also useful to report. In research works, such information can be included in the acknowledgment footnote or a separate footnote, or in the main body of the text. In quantitative research, such information is typically provided in a methods section.

3.2 Robust, replicable and transparent pieces of data

In the preceding section, we proposed that a piece of data in research on meaning has four components:

(8) **Pieces of data in research on meaning**
A piece of data in research on meaning consists of

a. a linguistic expression of language L,

b. a context in which the linguistic expression was uttered,

[11] Motivation for including the specific wording of the question comes from the finding that slightly different question formulations may result in different responses, cf. e.g., Clark and Schober 1992.

c. a response by a native speaker of language L to the task posed for the expression in a. in the context in b., with information about the response task, and

d. information about the native speaker(s) who responded.

We characterized each of these components and provided motivation for their inclusion in the definition of a piece of data. What we put forward as a proposal for a piece of data is already established practice in parts of the contemporary literature on meaning. It is most consistently practiced in quantitative research, but it is also practiced in some research based on introspective judgments, as illustrated with the example in (3) from Rojas-Esponda 2014, as well as in some research based on one-on-one elicitation, as illustrated with the example in (4) from Cable 2014. (Both authors provide information in their papers about who provided the relevant judgments: the author herself in the case of (3) and a French speaker for (4).)

But empirical evidence in some contemporary research on meaning — including some of our own work — also relies on pieces of data consisting of only some of the components identified in (8). A survey of 40 recent journal articles we conducted[12] established that almost half of the papers either exclusively or almost exclusively presented pieces of data consisting only of a linguistic expression (usually a sentence). It was also not rare to find pieces of data consisting of a linguistic expression and a context, but no information about the response (task), and it was common to find pieces of data consisting of a linguistic expression and a response (task), but no context. Finally, we found that there is no standard practice in research on meaning about what to report about the native speakers who provided the responses. Our survey revealed that only papers that presented results from quantitative research consistently include such information. In fact, the majority of papers in our survey did not include any information about the speakers whose responses were relied on. This practice is especially pervasive when the languages under investigation are languages widely spoken by linguists, such as English, German, Greek, Spanish, Korean, etc. (whether or not the authors of the paper are native speakers of the language under investigation).

It is this heterogeneity in what is taken to be a piece of data in research on meaning that, in part, motivated us to write this paper. Our proposal that a piece of data has the four components characterized in (8) is guided by the objectives that pieces of data that inform theories of meaning ideally are robust, replicable and transparent:

(9) **Objectives:** A piece of data in research on meaning

a. is **robust** if it explicitly controls for factors that may lead to variation in speaker judgments,

b. is **replicable** if it maximally facilitates replication in the same or another language, and

c. is **transparent** if it makes fully explicit how it supports the hypothesis.

In the remainder of this section, we argue that pieces of data that include all four components best satisfy the objectives in (9).

[12] We surveyed 40 journal articles published between 2012 and 2015 in the four leading journals in research on meaning: *Natural Language Semantics, Linguistics & Philosophy, Journal of Semantics*, and *Semantics & Pragmatics*. We selected ten articles published in each of these journals within the aforementioned timeframe, excluding papers that primarily relied on secondary sources. These 40 articles cover a wide range of empirical phenomena and include data collected through introspection, one-on-one elicitation and quantitative research. We examined each article for what is considered a piece of data and the response tasks used.

3.2.1 Robust pieces of data

The first objective in (9a) is for pieces of data to be robust, i.e., to control for factors that may lead to variation in speakers' responses. Two factors that are well-known to influence a speaker's response to a task about a linguistic expression are the context in which the linguistic expression is uttered, and the speaker herself. Given the wide range of linguistic phenomena that are context-dependent, including nominal, temporal, modal and aspectual reference, presuppositions, implicatures, discourse particles, and information structure, the context in which a linguistic expression is presented undoubtedly influences the response by the native speaker. As discussed in Schütze 1996:§5.3.1, even the extent to which a particular string is judged to be an acceptable, i.e., syntactically well-formed, sentence is affected by context. Speakers may also vary in their responses depending on, for instance, their dialect (e.g., Szmrecsanyi 2015), whether they have had linguistic training (Schütze 1996:§4.4.1) and their literacy and education (Schütze 1996:§4.4.2). Since context and the native speakers are factors that may lead to variation in a speaker's response, we argue that pieces of data that include a context and information about the speakers are more robust than pieces of data that do not include these components.

In research on meaning, there are (at least) two research questions that are addressed using pieces of data that involve speakers' responses to linguistic expressions presented without a context, i.e., in a null or so-called 'out-of-the-blue' context. The first is the question of what the context-independent meaning of an expression is. For instance, in research on temporal reference, the interpretation of decontextualized sentences is sometimes taken to identify the default temporal reference of sentences, as illustrated in the following excerpt from Smith et al. 2007:59:

> We begin with canonical examples of Navajo. [(10a)] . . . [has] Imperfective viewpoint [and is] taken as present in the absence of contextual information to the contrary. [(10b)] has the perfective viewpoint and is taken as past. The translations reflect the default temporal interpretations:
>
> (10) a. Jáan Tségháhoodzánídi naaghá
> John Window.Rock-in around-3subj-impf-go
> 'John is hanging out at Window Rock.'
>
> b. Shimá ch'iyáán ła' bá naháłii'
> 1-mother groceries some 3-for pref-1subj-perf-buy
> 'I bought some groceries for my mother.'

Another example comes from the literature on scalar implicatures. van Tiel et al. 2014 asked native speakers of English to judge whether de-contextualized sentences with weak scalar expressions (e.g., *She is intelligent*) give rise to a conversational implicature that denies a semantically stronger expression (e.g., 'intelligent, but not brilliant'). By presenting the sentences out of context, van Tiel and his colleagues established context-independent differences between different types of scalar expressions.

One issue with asking speakers to respond to expressions in null contexts is that the task is rather unnatural: utterances are not typically made in a completely empty context, devoid of any information about e.g., the interlocutors and the situation in which the utterance occurs. It is possible that speakers who are asked to respond to expressions in a null context imagine a context in which the expression could be uttered (or could not be uttered) and that their response is influenced by that context. In this case, their response does not reflect the meaning of the expression in the null context provided by the researcher but rather in the context they imagine. The problem is that the researcher is not privy to this context and hence does not know which features of the context may have led to the response. As Crain and Steeedman (1985) put it: "The fact that the experimental situation in question makes a null contribution to the context does not mean that the context is null. It is merely not under the experimenter's control ... the so-called null context is in fact simply an *unknown* context" (p.338, italics in original). Consequently, getting a judgment in a null context does not necessarily reflect the context-independent meaning of the expression. See Tonhauser 2015:144 for a critique of using null contexts in research on temporal and aspectual reference.

The second research question that is often investigated with de-contextualized examples uttered in out-of-the-blue contexts is the question of whether a linguistic expression is judged to be acceptable at the beginning of a discourse and, if yes, what the expression means at the beginning of a discourse. The following excerpt from Kripke 2009:373 illustrates this practice:

> (14) Sam is having dinner in New York tonight, too.
>
> Imagine (14) as uttered out of the blue; no context is being presupposed in which we are concerned with anyone else having dinner in New York. [...] it is obvious that the *too* here is particularly bizarre. The hearer will say, "'Too'? What do you mean, 'too'? What person or persons do you have in mind?"

Researchers who work on their own language and provide judgments themselves can, of course, take care that the relevant expressions are judged in null contexts. But otherwise, the same worry as mentioned above arises: When speakers respond to linguistic expressions presented without a context, the researcher has no control over whether they make up a context that the researcher is not privy to. One way to address this worry for the second type of research question is to present the expression in a context that makes clear that the expression is supposed to be uttered as the first or one of the first utterances of a discourse. We call such contexts 'discourse-initial':

(11) **Discourse-initial context**
The context of a piece of data is a discourse-initial context when it describes a situation in which the target utterance is the first or one of the first utterances of a (possibly, one-turn) discourse.

One example of a discourse-initial context is in (12) from Gitksan (Tsimshianic). The hypothesis that was tested with this example was that the discourse particle *=ist* 'QUDD' indicates a downdate of the question under discussion (Gutzmann and Castroviejo Miró 2011), i.e., is infelicitous in a context in which the prejacent implication of *=ist* 'QUDD' (here, the proposition that Betty worked in Abbotsford) does not answer the current question under discussion (Ginzburg 1996, Roberts 2012). The context in (12) establishes that the speaker and the addressee know each other (they are married) but also, crucially, that there is no prior linguistic context: Adam and Betty have not yet raised a topic of conversation and, in particular, nothing about where Betty worked has been part of the conversation so far between the two. (A native speaker of Gitksan judged this example to be unacceptable. We return to this example in section 5.)

(12) Context: Adam and Betty are married. Betty is a traveling saleswoman and she works in a number of different towns in the surrounding area. The two are having dinner and nobody has said anything yet. Betty suddenly says:

#G̲a'a=hl Abbotsford win ahle'lsd-'y**=ist**.
LOC=CN Abbotsford COMP work-1SG.II=QUDD

'I worked in Abbotsford today.'

It may, in some cases, not be plausible for the target utterance to be the very first utterance of a discourse. It is for this reason that the definition of a discourse-initial context in (11) allows for the relevant linguistic expression to be the first or one of the first utterances. For instance, in (13), an example from Paraguayan Guaraní (Tupí-Guaraní) that was judged to be acceptable by four native speakers, the relevant linguistic expression is uttered only after the mother has apologized on behalf of her daughter. The hypothesis that was explored with this piece of data was that sentences with the verb stem *–kuaa* 'know' are acceptable when the content of the complement clause is not something that both the speaker and the addressee know. Thus, a crucial feature of the discourse-initial context of (13) is that the addressee does not know the speaker and, therefore, that the addressee does not know that the girl has to use glasses to drive.

(13) Context: A girl backs out of a driveway and hits Susi's car. A woman comes running out of the house, apologizes that her daughter hit Susi's car, and says:
Ha'e oi-kuaa o-moĩ-va'erã-ha i-lénte o-maneja-ha-guã.
PRON.S.3 A3-know A3-put-MUST-NOM B3-glasses A3-drive-NMLZ-PURP

'She knows that she has to use her glasses to drive.' (adapted from Tonhauser et al. 2013:80)

In sum, while some research questions are sometimes investigated in null contexts, care should be taken to ensure that the native speaker is not silently enriching the utterance with her own imagined context. One way the likelihood of this can be lowered is by establishing some minimal context, rather than presenting utterances completely out of the blue. In general, pieces of data which include a context and provide information about speakers are relatively more robust, because they control for elements which are known to affect speaker judgments.

3.2.2 Replicable pieces of data

The second objective in (9b) is for pieces of data to be replicable, i.e., to maximally facilitate replication of the data in the same language from other speakers (e.g., to explore inter-speaker variation) or from speakers of another language (to explore cross-linguistic variation). In order to study inter-speaker or cross-linguistic variation, it is vital that the same piece of data is collected, modulo the speakers or the language of the linguistic expression. If a researcher attempting replication of a piece of data does not have access to the context or the response task of the piece of data, then the replicating researcher may use a different context or employ a different kind of response task in their replication. An example of a piece of data that is difficult to replicate is given in (14), which Moltmann (2013:36) argues "does not sound right" and marks with '??'.

(14) ??Socrates is a man. (Moltmann 2013:36)

The piece of data in (14) does not include a context or information about the response task. Instead, we are left to infer that the expression was judged to be less than acceptable under the assumption that *Socrates* refers to the classical Greek philosopher and that the time at which (14) is uttered is a time after this philosopher's death. Under the assumption of a different context, e.g., one in which *Socrates* refers to a man called Socrates who is alive at the utterance time, or one in which *Socrates* refers to the philosopher but the example was uttered by a contemporary of his, the example is judged to be acceptable. The fact that (14) does not include a context that fixes the referent of the name *Socrates* or the time at which (14) was uttered means that a researcher attempting to replicate this piece of data may use a different context, and hence obtain a different response. To illustrate, the second author of this paper attempted to replicate (14) in Gitksan and found that a speaker of the language judged the Gitksan variant of (14) given in (15) to be acceptable:

(15) Gyat=t Saklatiis.
man=DM Socrates
'Socrates is a man.'

We now have an unfortunate situation on our hands since it is not clear whether the difference in judgments obtained for English (14) and Gitksan (15) is due to linguistically interesting variation (e.g., perhaps Gitksan does not have life time effects?) or merely due to the English and Gitksan speakers having given their responses relative to different contexts. For example, perhaps the Gitksan speaker in (15) silently imagined one of the contexts outlined above in which the English (14) would also be acceptable. In sum, we argue that a piece of data that includes a context is more replicable than a piece of data that doesn't.

3.2.3 Transparent pieces of data

The third objective in (9c) is for pieces of data to be transparent, i.e., to allow readers of the work in which the piece of data occurs to understand what the piece of data is that provides empirical support for the hypothesis. Take, for instance, the Japanese piece of data in (16). This piece of data is provided in support of the hypothesis that present tense utterances with *motto* give rise to a degree reading but not to a negative reading (under which the example would imply that the cake is not delicious):

(16) ??Kono mise-no keeki-wa motto oishii.
this store-GEN cake-TOP MOTTO delicious
'This store's cake was still much more delicious than a contextually-determined store's cake.'
(only degree reading available) (Sawada 2014:208)

A reader who is a native speaker of Japanese may well be able to construct a context for the linguistic expression and judge the acceptability of the expression in that context in order to verify that a negative reading is not available for (16). But there is no guarantee that a native speaker of Japanese would construct a context that leads them to give the same judgment as reported in (16). And a reader who is not a native speaker of Japanese cannot construct a context for (16) and give a response to (16) that would allow them to verify that (16) provides empirical support for the hypothesis. Thus, pieces of data that lack a context or information about the response task do not provide full access to the piece of data that is taken to support the hypothesis. In other words, including a context and information about the response task in the piece of data makes the piece of data more transparent.

Including information about the response and the response task is currently not the standard practice in the field, except in the quantitative literature. Importantly, the diacritic that accompanies the linguistic expression (or the absence of such a diacritic) does not convey which task was responded to. Rather, diacritics indicate the researcher's interpretation of a speaker's response. If, for example, a judgment of acceptability is elicited for a linguistic expression and that expression is judged to be unacceptable by a native speaker, then the researcher may choose to mark the example with an asterisk (*) if she hypothesizes that the unacceptability is due to syntactic reasons, or with a hash mark (#) if she hypothesizes that the unacceptability is due to semantic/pragmatic reasons. The problem with relying on the diacritics is compounded by the fact that it is not always clear how a particular diacritic is used in a particular paper (for issues about the use of diacritics see also Schütze 1996:ch.2.3.3). The asterisk, for example, though widely used to indicate syntactic ill-formedness, is also used to indicate unacceptability in particular contexts or under particular interpretations (e.g., Nicolae 2014, Henderson 2014). Likewise, the hashmark is often used to indicate that an expression is taken to be semantically or pragmatically anomalous, but is also used, e.g., in Coppock and Beaver (2014), to indicate that an expression is not a paraphrase of another. Thus, the diacritic does not replace information about the response task or the response, and a piece of data is more replicable and transparent if it includes information about the response and the response task.

3.3 Summary

In this section, we proposed that a piece of data in research on meaning consists of four components, summarized in (8): a context, a linguistic expression, a response by a native speaker to the expression in that context, and information about the speaker(s). We characterized these four components in section 3.1 and argued in section 3.2 that pieces of data that include all four components are more likely to be robust, replicable and transparent than pieces of data that lack a context, a response or information about the speakers.

4 Response tasks in research on meaning

Having established the four components of a piece of data, we now characterize the main types of response tasks used in offline research on meaning (section 4.1). We then argue that some response tasks, including acceptability and implication judgment tasks, are better suited than others (e.g., paraphrase and translation tasks) for yielding robust, replicable and transparent pieces of data (section 4.2).

4.1 Characterization of tasks

We adopt the convention of referring to a task that asks a native speaker to respond to a question about X as an 'X judgment task'. For example, in an acceptability judgment task a speaker is asked to judge the acceptability of an utterance, and in a truth value judgment task a speaker is asked to judge the truth value of an utterance.[13]

4.1.1 Acceptability judgment tasks

In an acceptability judgment task, a native speaker of a language judges the acceptability of an utterance of a linguistic expression of that language in a context. For example, in (6) above, native speakers of Hausa were asked to judge whether the given sentence is appropriate to say in the context provided. Other questions that might be posed to the speaker include 'Does this sound good to you?' or 'Would you say this?' (see Bohnemeyer 2015:36 for further examples).[14] Both binary and non-binary response options, including responses on a Likert scale or magnitude estimations, are possible (see e.g., Schütze 1996, Matthewson 2004, Schütze and Sprouse 2014 and Sprouse et al. 2013 for discussion). In one-on-one elicitation, speakers may indicate their choice using assent or dissent particles ('yes' or 'no'), or by providing some other verbal indication of assent or dissent ('That sounds good/bad'), possibly in combination with non-verbal cues like nodding, frowning or head-shaking (see Tonhauser et al. 2013:fn.13 for a brief discussion).

Two assumptions are generally made about this task. The first is that native speakers judge a linguistic expression uttered in a context to be acceptable if and only if the linguistic expression is syntactically well-formed, felicitous[15] and has truth conditions which are compatible with that context. Thus, an acceptability judgment is elicited for a linguistic expression that is uttered in a context: a de-contextualized linguistic expression might be judged to be unacceptable because a felicity condition is violated or because the speaker imagines a context in which the linguistic expression is false.

A second assumption is that native speakers judge the utterance to be unacceptable if the utterance is syntactically ill-formed, or if it is infelicitous even if it is syntactically well-formed and has truth conditions which are compatible with the context, or if it is false even if it is syntactically well-formed and felicitous, or any combination of syntactically ill-formed, false and infelicitous. Thus, a judgment of acceptability supports the hypothesis that the utterance is syntactically well-formed, felicitous and has truth conditions compatible with the context, but a judgment of unacceptability does not by itself provide insight into why the utterance was judged so (see also Chomsky 1977:4 and Matthewson 2004:409). Consider, for example, the English example in (17), which the second author judged to be unacceptable in the context in which it is presented. We use the diacritic '×' here to indicate that the sentence was judged to be unacceptable in the context in which it was uttered and to remain neutral about whether this judgment is due to the sentence

[13] We thereby expand on Carson Schütze and his colleagues' recommendation (Schütze 1996:ch.2, Schütze and Sprouse 2014:27, Sprouse et al. 2013:§2.1) that one not refer to a task in which speakers are asked to judge the acceptability of a string for the purpose of establishing whether the string is syntactically well-formed as a 'grammaticality judgment' task since speakers are not asked to judge grammaticality but acceptability.

[14] The instructions that precede the elicitation of judgments, including acceptability judgments, provide guidance to native speakers about how to interpret these questions. In general, researchers use control examples, e.g., with undeniably acceptable or undeniably unacceptable expressions, to identify whether the native speakers have interpreted the questions appropriately. But, of course, the question of whether different variants of these questions may result in different responses is an important one. The fact that this is still an open issue motivates including detailed information about the response task, as we argued in section 3.

[15] An utterance is felicitous in a context if and only if its felicity conditions are satisfied in the context. An example of a felicity condition is the requirement of a definite noun phrase like *the dog* for a salient discourse referent that denotes a dog to exist in the context (e.g., Heim 1982).

being syntactically ill-formed, infelicitous or false.

(17) Context: John came to Hamburg yesterday.
× He arrives yesterday.

From this judgment alone, we do not know whether (17) is syntactically ill-formed, infelicitous, false, or a combination of the three; it is up to the researcher to determine the reasons for the unacceptability judgment, in conjunction with the hypothesis under which (17) was elicited (see also Matthewson 2004:375). We discuss in section 5 how minimal pairs of data with acceptability judgments can be used to this effect.

4.1.2 Implication judgment and related tasks

In an implication (or, inference) judgment task, a native speaker of a language is asked to judge whether the utterance of a linguistic expression of that language in a context gives rise to a specific implication.[16] We distinguish between direct and indirect implication judgment tasks. In a direct implication judgment task, the native speaker responds to a question about the implication that the researcher is interested in. For example, Geurts and Pouscoulous (2009) were interested in whether utterances of French sentences with *certains des* 'some' implicate the denial of the stronger alternative *tous* 'all'. In one of their experiments, native speakers of French were presented with French versions of the English sentence *Betty thinks that Fred heard some of the Verdi operas* and they were then asked the following question in French: 'Would you infer from this that Betty thinks that Fred didn't hear all the Verdi operas?' (with response options 'yes' and 'no'). This task is a direct implication judgment task because native speakers are directly asked about the implication of interest ('Fred didn't hear all the Verdi operas'). Another piece of data with a direct implication judgment task is (2), repeated below:

(2) Context: A is visiting B's community. A notices a man who is addressing a small group of villagers; he asks:

A: Mava'e pa kova'e ava?
who Q this man
'Who is this man?'

B: Ha'e ma ore-ruvicha o-iko va'e-kue. Aỹ, porombo'ea o-iko.
ANA BDY 1.PL.EXCL-leader 3-be REL-PST now teacher 3-be
'He was our leader. Now, he is a teacher.' (Thomas 2014:394f.)

Thomas (2014) writes about this example that "[a]fter reading this discourse, consultants were asked whether they think that the man A is asking about is still the leader of the village" (p.394). (Thomas reports that all consultants judged that this man is no longer the leader.) For other uses of the direct implication judgment task, see e.g., van Tiel et al. 2014.

In an indirect implication judgment task, in contrast, the native speaker is asked a question seemingly unrelated to the implication of interest. However, the answer to this question allows the researcher to draw a

[16] The term 'implication' encompasses any kind of inference, including entailments, conversational implicatures, conventional implicatures, and presuppositions.

conclusion about the implication. This task was used in Tonhauser et al.'s (2013) investigation of projective content in Paraguayan Guaraní. Consider the examples in (18):

(18) Context: There is a health program that gives medicine to everybody who has ever smoked or currently smokes. Maria is administering the program in a particular town; since she doesn't know the people in the town, she is being assisted by Mario, a local townsman, who tells her the following about Marko:

a. Márko nd-o-pita-vé-i-ma.
Marko NEG-A3-smoke-more-NEG-PRF
'Marko doesn't smoke anymore.' (adapted from Tonhauser et al. 2013:88)

b. Márko nd-o-pitá-i araka'eve.
Marko NEG-A3-smoke-NEG never
'Marko never smoked.'

The implication of interest was that Marko used to smoke in the past. Rather than directly asking Paraguayan Guaraní speakers whether they would infer from (18a) or (18b) that Marko used to smoke in the past, speakers were asked to judge whether Maria would give the medicine to Marko. The assumption was that if speakers responded in the affirmative, i.e., that, yes, Maria would give the medicine to Mario, they would take the uttered sentence to mean that Marko smoked in the past; if, on the other hand, speakers responded in the negative, then they would not take the uttered sentence to mean that Marko smoked in the past. Since Paraguayan Guaraní speakers consistently responded, upon hearing (18a), that Maria would give the medicine to Marko, Tonhauser et al. (2013) concluded that the implication of interest arises from (18a). Speakers do not, however, respond that Maria would give the medicine to Marko upon hearing (18b), which provides evidence that the implication of interest does not arise from that utterance.

Entailment judgment tasks An entailment judgment task is a variant of the implication judgment task. In an entailment judgment task, a native speaker of a language is asked to judge whether an utterance of a sentence of the language has a particular entailment. Thus, this task is a variant of the implication judgment task because the speaker is asked to judge whether an utterance of the sentence gives rise to a particular implication and also whether that implication is an entailment. Crnič (2014), for example, states about (19) "that John read the book once is entailed by the proposition that John read the book twice" (p.176), thereby (presumably) illustrating the results of an entailment judgment task.

(19) a. John read the book once.
b. John read the book twice. (Crnič 2014:176)

Paraphrase judgment tasks A paraphrase judgment task is another variant of an entailment judgment task. In the paraphrase judgment task, a native speaker of a language is presented with a linguistic expression of their language, and is then either asked to judge whether another linguistic expression of their language is a paraphrase of the first expression (i.e., whether the two expressions convey the same meaning; presumably at least have the same truth conditions), or asked to identify a linguistic expression that paraphrases the first expression. For example, Coppock and Beaver (2014) write about the examples in (20) that "when *mere* occurs in an argumental noun phrase, it can be paraphrased with *just* and *merely*, but resists being paraphrased with *only*, and cannot be paraphrased with *exclusively* or any of the other exclusives that allow only complement exclusion readings" (p.374).

(20) a. The **mere** thought of food makes me hungry.
b. **Just** the thought of food makes me hungry.
c. **Merely** the thought of food makes me hungry.
d. **Simply** the thought of food makes me hungry.
e. ?**Only** the thought of food makes me hungry.
f. #**Exclusively** the thought of food makes me hungry.
g. #**Purely** the thought of food makes me hungry.
h. #**Solely** the thought of food makes me hungry. (Coppock and Beaver 2014:374)

Similarity of meanings judgment tasks A variant of paraphrase judgment tasks is the similarity of meanings judgment task (e.g., Degen 2015, Matthewson 2015). This task requires speakers to judge the similarity of the meanings of utterances of two sentences. Degen (2015), for example, asked native speakers of English to judge the similarity of the meanings of naturally occurring examples with *some* and their constructed counterparts where *some* was replaced by *some, but not all*:

(21) Degen 2015:17
a. You sound like you've got **some** small ones in the background.
You sound like you've got **some, but not all,** small ones in the background.
b. I like **some** country music.
I like **some, but not all,** country music.

Unlike paraphrase judgment tasks, similarity of meanings judgment tasks do not require speakers to judge whether the two utterances have the same truth conditions. A negative response (as with, e.g., (21a)) constitutes a clue that the speaker perceives the two expressions to differ in meaning. This clue can then provide the impetus to developing a more refined hypothesis about how the two expressions differ in meaning, e.g., whether they differ in their truth conditions, their felicity conditions, or in some other way. A positive response (as with, e.g., (21b)) provides a clue that the responding speaker did not take the two expressions to differ in meaning. However, a positive response does not warrant the assumption that the two expressions have the same truth and felicity conditions.

4.1.3 Truth value judgment and related tasks

The truth value judgment task was illustrated with the examples in (4) above from Cable 2014 (Seth Cable confirmed in p.c. the use of a truth value judgment task in this example). In this task, a native speaker of a language is asked to judge the truth value of an utterance of a declarative sentence of the language in a context. Speakers can be asked to respond to questions like 'Is this sentence true?' or be asked to indicate non-verbally whether the sentence is true. Native speakers are typically asked to give a forced choice binary response (e.g., 'yes', 'no'), though truth value judgment tasks with non-binary responses have also been used (see e.g., Chemla and Spector 2011). For use of truth value judgment tasks with children see Crain and McKee 1985 and Crain and Thornton 1998.

Inherent to the truth value judgment task is that it can only be applied to declarative sentences, which denote true or false propositions, as opposed to interrogative or imperative sentences. Furthermore, a theoretical assumption about the truth values of utterances is that only utterances whose felicity conditions are satisfied in the context in which the utterance is made have a truth value. For additional discussions of this task see e.g., Matthewson 2004, Krifka 2011 and Bohnemeyer 2015.

Ambiguity judgment task The ambiguity judgment task is a variant of the truth value judgment task: for a native speaker to judge that an expression is ambiguous, they have to identify a context in which one of the two meanings of the expression is true and the other one is false, and vice versa. One example comes from Alrenga and Kennedy (2014), who state that the example in (22) "is ... ambiguous" (p.4) and then describe the two readings:

(22) More students have read Lord of the Rings than have read every other novel by Tolkien.
(Alrenga and Kennedy 2014:4; attributed to Bhatt and Takahashi 2011:fn.18)

"Under one of its readings, [(22)] conveys that for each Tolkien novel *x* other than *Lord of the Rings*, the number of students who have read *Lord of the Rings* exceeds the number of students who have read *x*. This is a kind of $>_{max}$ reading for [(22)], since it follows that *Lord of the Rings* has more readers than the most-read other Tolkien novel. Under another reading, [(22)] instead conveys that the number of students who have read *Lord of the Rings* exceeds the number of students who have read all of the other Tolkien novels."

4.1.4 Translation task

In a translation task, a native speaker of a language provides a translation of a linguistic expression of the language (possibly presented in a context) into another language that they are a native speaker of (or at least have some fluency in), or vice versa. As noted in Deal 2015, an assumption that underlies this task is that "[t]he input to translation and the output of translation are equivalent in meaning" (p.158). For example, Tonhauser (2011) writes that, in Paraguayan Guaraní, "[i]n subordinate clauses, unmarked verbs are compatible with future time reference" (p.209), pointing to the Paraguayan Guaraní example in (23) with its English translation for evidence.

(23) Re-karú-ta re-jú-rire.
A2sg-eat-FUT A2sg-return-after
'You will eat after you return.' (Tonhauser 2011:210)

4.1.5 Interim summary

The preceding sections characterized the response tasks most frequently used in contemporary (offline) research on meaning. All of these tasks were used in the papers in our survey. We found, however, that authors do not always identify the response task that was used. In some cases, authors referred to a specific task, e.g. "grammaticality judgment" or "felicity judgment", but those task descriptions presumably often did not reflect what speakers were asked to judge. To give an example from our own work: Matthewson (2004) writes about the elicitation of "felicity judgments" (p.380), even though acceptability judgments were elicited. In other cases, no response task was identified but authors mentioned something akin to a speaker's response, e.g., that an expression was judged as "(in)coherent", "(im)possible", "odd", "problematic" or "contradictory". In a third type of case, no indication about the response task or the response was given. In such cases, a hypothesis about meaning was asserted and typically accompanied by a de-contextualized linguistic expression. This situation occurs often with particular linguistic phenomena. For scope, for example, a relatively prevalent pattern is for an author to state that an element X scopes over an element Y, and to present as the total support for this claim a de-contextualized sentence containing the expressions X and Y. Similarly for presuppositions, we found authors presenting de-contextualized sentences and simply asserting that they do or do not have some presupposition. Again, it is this heterogeneity in what is taken to constitute empirical evidence in research on meaning that motivated us to write this paper. As mentioned above, we advocate for the explicit identification of the response task that was used, and for referring to the task with a name that reflects what the speaker was asked.

4.2 Evaluation of tasks

We now evaluate the response tasks we just characterized with respect to whether they lead to robust, replicable and transparent pieces of data.[17] We argue that the tasks are qualitatively different with respect to this objective, as illustrated in Figure 1:

$$\text{translation} \quad < \quad \left\{ \begin{array}{l} \text{paraphrase} \\ \text{entailment judgment} \\ \text{ambiguity judgment} \end{array} \right\} \quad < \quad \text{truth value judgment} \quad \overset{?}{<} \quad \left\{ \begin{array}{l} \text{acceptability judgment} \\ \text{implication judgment} \\ \text{similarity judgment} \end{array} \right\}$$

$$\xrightarrow{\text{more robust, replicable and transparent pieces of data}}$$

Figure 1: Evaluation of response tasks in research on meaning

Translation task We start our evaluation with the translation task, at the very left in Figure 1. It is obvious that any speaker who is competent in two languages can produce translations. These translations, however, are at best clues to meaning, as argued extensively in Matthewson 2004, Krifka 2011, Bohnemeyer 2015 and Deal 2015, among others, since translations need not preserve truth or felicity conditions. For instance, one language could lack easy means to express truth conditions which are easily expressible in the other

[17] Schütze 2008 discusses several other tasks with respect to whether theoretically untrained speakers can reliably perform them.

(this in spite of strong claims about universal translatability by Jakobson 1959, Katz 1976; see von Fintel and Matthewson 2008 and Deal 2011, 2015 for discussion). Also, one language may neutralize a distinction which is encoded in the other, leading to stronger or weaker truth conditions depending upon the direction of translation. Such differences between languages also show that even translations offered by theoretically trained native speakers cannot be assumed to adequately capture all truth and felicity conditions of the original. Translations can also fail to preserve presuppositions or implicatures, or introduce new ones. And, as exemplified in Deal 2015, speakers may volunteer translations which have different truth conditions than the original in order to avoid incorrect pragmatic inferences which would arise from a more literal translation. In sum, cross-linguistic equivalence can fail in a translation task in a number of ways, and hence translations do not provide reliable evidence in research on meaning. In particular, the translation of an expression into English, as the dominant language of publications, does not provide evidence that the two expressions have the same truth conditions or give rise to the same inferences. This is not to say that the translation task has no place in research on meaning. On the contrary: in many instances, translations are a first step towards developing a hypothesis about meaning. Regardless, translations are at best a clue.

Paraphrase, entailment judgment and ambiguity judgment tasks We contend that paraphrase, entailment judgment and ambiguity judgment tasks cannot be reliably performed by speakers without linguistic training. This is evidenced by their characterizations in the previous sections: an understanding of truth conditions is required to be able to assess whether two utterances have the same truth conditions (i.e., are paraphrases of one another), whether one has stronger truth conditions (i.e., entails the other), or whether an expression has two distinct sets of truth conditions (i.e., is ambiguous). (Anybody who has had the experience of teaching students the concepts of equivalence, entailment and ambiguity can attest to the fact that these concepts require training.) Pieces of data that involve response tasks that can be performed reliably by speakers regardless of their linguistic training are more replicable than pieces of data that involve tasks that can only be performed reliably by speakers with linguistic training. Because pieces of data that involve these three types of task are less replicable, they are placed towards the lower end of our evaluation of response tasks in Figure 1. See Sprouse et al. 2013:§2.2 for the argument that tasks that require linguistically trained speakers are not ideal.

In general, a piece of data that includes a response that requires the responding speaker to perform linguistic analysis is less transparent than a piece of data that includes a response that does not require the speaker to perform linguistic analysis. For example, a piece of data that includes the response that the expression is ambiguous is less transparent than a piece of data that includes the response that the expression is acceptable in the context provided: a speaker is able to give the former response only after performing linguistic analysis, which is not the case for the latter response. Since this linguistic analysis is typically not presented as part of the piece of data, pieces of data that involve paraphrase, entailment judgment and ambiguity judgment tasks are less transparent.

A final reason why the entailment and ambiguity judgment tasks occur towards the left in Figure 1 is that pieces of data involving these tasks do not have contexts, by definition. That is, a sentence entails another one if and only if the latter is true whenever the former is true, and whether an expression is ambiguous, i.e. has two distinct sets of truth conditions, is determined without regard to a context in which that expression occurs. Since pieces of data involving these tasks do not involve a context, they are less robust.

We note here that pieces of data that support hypotheses about truth-conditional equivalence, entailment or ambiguity can, of course, be robust, replicable and transparent, as soon as other types of response tasks are used. As discussed in e.g., Crain and McKee 1985:104, to support a hypothesis about ambiguity the researcher can present the speaker with contexts that make one of the hypothesized meanings true and the other one false, and elicit judgments of acceptability of the relevant expression in these contexts from the speaker. Similarly, Bohnemeyer 2015:34f. points out that hypotheses about entailment can be empirically supported if the researcher constructs contexts and asks the speaker to judge whether the relevant expressions are true in the contexts: "the researcher is not asking for a direct judgment of entailment, but rather for a series of judgments about the truth of a pair of utterances in a series of scenarios" (p.35). Bohnemeyer also points to the possibility of supporting hypotheses about entailment using judgments of contradictions: "Speakers appear to be able to tell relatively immediately whether two statements are logically consistent or not. Consequently, one method for testing whether an utterance has a given entailment is by combining it with a second utterance, which negates the hypothetical entailment. If in the speaker's judgment the conjunction of the two utterances may be true in the same scenario, this suggests that the proposition negated by the second utterance is not an entailment of the first. But if the speaker judges the utterances to be inconsistent, this supports the entailment analysis" (p.35). See also de Marneffe and Tonhauser 2015 for the use of contradiction judgments.

Truth value judgment tasks In truth-conditional research on meaning, a sentence is true if and only if its truth conditions are fulfilled at the world and time of evaluation. Since the truth value judgment task asks a speaker to identify the truth value of a sentence, it cannot be reliably applied with theoretically untrained speakers, who have not learned to distinguish the truth conditions of an utterance from other conditions on its felicitous and pragmatically unmarked use. For instance, anyone who has taught undergraduate semantics and has had to explain why a sentence like *John arrived or Mary arrived* is true in a situation in which both John and Mary arrived will appreciate that theoretically untrained speakers often do not differentiate between truth conditions and implicatures. Similarly, Soames (1976:169), von Fintel (2004) and Abrusán and Szendrői (2011) argue that speakers may judge utterances that are infelicitous to be false, even though they are assumed not to have a truth value. Thus, when theoretically untrained speakers evaluate a sentence as true, it can indeed be assumed that its truth conditions are fulfilled, but when such speakers evaluate a sentence as false, it may indeed be false, or it may be infelicitous, or it may be true but pragmatically odd due to a conversational implicature, as in the disjunction example above. From this perspective, the truth value judgment task cannot be reliably applied with theoretically untrained speakers and, hence, pieces of data that involve a truth value judgment task are less replicable and transparent.[18]

The reason that truth value judgment tasks are nevertheless placed quite far to the right in Figure 1 is that there are at least two ways in which such tasks can be reliably applied with theoretically untrained speakers.

[18] Truth value judgment tasks are used in quantitative research on scalar implicatures despite these concerns. Bott and Noveck (2004), for example, asked speakers to judge the truth value of sentences like *Some elephants are mammals* and a judgment of 'false' was taken as evidence that an inference was drawn that makes the sentence false (*Some but not all elephants are mammals*). In fact, however, speakers may just be unwilling to judge such pragmatically underinformative utterances as true.

A first possibility is to interpret speakers' 'false' responses differently: rather than interpreting them as providing evidence that the truth conditions are not fulfilled, such responses are merely taken to provide evidence that the speaker finds something about the sentence objectionable in the context in which it was presented, be it that the sentence is false, infelicitous or pragmatically odd due to some implicature. Under this interpretation of a 'false' response, truth value judgment tasks can be reliably applied with theoretically untrained speakers, including children (see Crain and McKee 1985 and Crain and Thornton 1998), and pieces of data that involve this response task are replicable and transparent. The second way in which such tasks can be reliably applied with theoretically untrained speakers is to ask speakers about the truth value of a sentence that the researcher hypothesizes to be syntactically well-formed, felicitous and pragmatically unmarked in the context in which it is presented. A speaker's 'false' response to such a sentence can then be reasonably taken to be due to unfulfilled truth conditions. For instance, Cable's 2014 example (4bii) was shown to be syntactically well-formed (since it was judged to be true in another context, namely (4aii)), and it was hypothesized to be felicitous and pragmatically unmarked in the context in which it was judged. Hence, a 'false' response to (4bii) can be taken as evidence that the truth conditions of the sentence in (4bii) are not fulfilled in the context in (4ai). The same holds for Syrett & Koev's (2014) experiment 4, where theoretically untrained speakers' 'no/false' responses were taken as evidence that the utterances that were judged were false. In sum, whether pieces of data with responses to a truth value judgment task are replicable, transparent and reliable depends on the interpretation of the 'false' response and/or on whether the utterances have independently been ascertained to be syntactically well-formed and felicitous.

Acceptability, implication and similarity judgment tasks At the right end of the spectrum, we find acceptability, implication and similarity judgment tasks. Each of these tasks taps into properties of utterances that do not require training in linguistics to be reliably made. Acceptability judgment tasks tap into a property of utterances that speakers have conscious access to, namely whether an utterance sounds good (see also Sprouse et al. 2013:220). We also assume that speakers have (at least partially) conscious access to what is meant, i.e. the pragmatically enriched meaning of an utterance, and it is this awareness of what is meant that the implication judgment task taps into. Since similarity judgment tasks ask speakers to assess the similarity of what is implied by two utterances, this task can also be performed without linguistic training. As a consequence of being applicable with theoretically untrained speakers, these three tasks lead to pieces of data that are replicable. By virtue of being collectible with theoretically untrained speakers, pieces of data collected through acceptability, implication or similarity judgment tasks also satisfy the transparency objective, since they do not require the speakers to conduct analysis in giving their responses. Finally, pieces of data involving one of these three response tasks may include a context and information about the responding speakers, and can thus lead to pieces of data that are robust.

4.3 Summary

In this section, we characterized the response tasks most frequently used in contemporary research on meaning, and pointed out that they vary in the extent to which they lead to robust, replicable and transparent data. In particular, we argued that acceptability, implication and similarity judgment tasks lead to more robust, replicable and transparent pieces of data than translation, paraphrase, entailment judgment and ambiguity judgment tasks. We also showed that truth value judgment tasks can lead to robust, replicable and transparent piece of data.

5 Turning pieces of data into empirical evidence

Just like the Frog Prince is just a frog until he is kissed by the princess, pieces of data are just pieces of data until a researcher transforms them into empirical evidence. The power to transform pieces of data into empirical evidence comes from the hypothesis about meaning that the researcher is exploring: by stating how the pieces of data provide empirical support for or against the hypothesis, the pieces of data transform into empirical evidence, namely empirical evidence for or against that particular hypothesis.

Our goal in this section is to illustrate how hypotheses about meaning are empirically supported, i.e., which pieces of data provide empirical evidence for which hypotheses. This task may at first seem daunting since semanticists entertain as many hypotheses about meaning as there are frogs in the world. Ultimately, however, there are only four types of pieces of data that can support hypotheses: positive pieces of data, negative pieces of data, and two types of minimal pairs of pieces of data. We discuss these in the following.

5.1 Evidence from positive pieces of data

A positive piece of data is one in which the speaker's response to the task about the linguistic expression was positive. What counts as 'positive' is determined relative to the response task: e.g., a positive response to a binary acceptability judgment task is a judgment of acceptability, to a binary truth value judgment task is a judgment of truth, and to a direct implication judgment task is a judgment that the implication arises from the linguistic expression.[19] The connection between a positive piece of data and the hypothesis about meaning that the piece of data is taken to support is provided by statements like those in (24). Following the quantitative literature, we refer to such statements as linking hypotheses.[20]

(24) a. **Linking hypothesis for a judgment of acceptability:** If an expression is judged to be acceptable in a context, the expression is syntactically well-formed, it is felicitous in that context and its truth conditions are compatible with that context.

b. **Linking hypothesis for a judgment of truth:** If an expression is judged to be true in a context, the truth conditions of the expression are compatible with that context.

c. **Linking hypothesis for a judgment that an implication arises:** If an expression is judged to give rise to an implication in a context, the expression gives rise to that implication in that context.

[19] Throughout this section, we limit our discussion to response tasks that were identified in section 4 as leading to maximally robust, replicable and transparent pieces of data.

[20] In the words of Tanenhaus et al. (2000), "[t]he interpretation of all behavioral measures depends upon a theory, or "linking hypothesis," that maps the response measure onto the theoretical constructs of interest" (p.564f.); linking hypotheses "are a necessary part of the inference chain that links theory to data" (p.565). Thus, linking hypotheses are general hypotheses about how to interpret the responses given by speakers to pieces of data. They are distinct from the specific theoretical hypotheses that semanticists explore, and from the predictions which derive from the theoretical hypotheses being tested.

These linking hypotheses for positive data with acceptability judgments, as in (24a), for positive data with truth value judgments, as in (24b), and for positive data with implication judgments, as in (24c), are an important part of transforming positive pieces of data into empirical evidence for particular hypotheses about meaning.

To illustrate how positive pieces of data provide evidence for hypotheses, consider the hypothesis (from Mucha 2013) that temporally unmarked Hausa sentences are compatible with past temporal reference. From this hypothesis, one can derive the prediction that B's utterance in (1), repeated below, is felicitous in the context of A's question. Given the linking hypothesis in (24a) for a judgment of acceptability, the positive piece of data in (1) provides empirical evidence that B's utterance is felicitous in the context of A's question (and also that B's utterance is syntactically well-formed and that its truth conditions are compatible with the context). Thus, it is under the linking hypothesis in (28a) that the piece of data in (1) is transformed into a piece of evidence for Mucha's hypothesis about temporal reference.[21]

(1) A: "What were Audu and Binta doing yesterday when you called them?"

B: Su-nā̀ màganā̀.
3PL-CONT talk

'They were talking.' (Mucha 2013:388)

In research on meaning, it can be useful to group together two or more minimally different, positive pieces of data. Take, for example, the hypothesis that Gitksan bare verb forms, like *ha'wits'am* 'crush' in (25), can denote habitual states in the actual world as well as habitual states only found in possible, non-actual worlds. From this hypothesis we can derive the prediction that the linguistic expression in the examples in (25) is true in the context in (25a), which describes a situation in which the machine regularly crushes oranges in the actual world, and in the context in (25b), which describes a situation in which the machine has not yet crushed an orange.

(25) a. Context: This machine regularly crushes oranges.

Ha-'wits'-am olents tun=sa.
INS-squeeze-ATTR orange DEM=PROX

'This machine crushes oranges.'

b. Context: This machine was built to crush oranges, but has not crushed any yet.

Ha-'wits'-am olents tun=sa.
INS-squeeze-ATTR orange DEM=PROX

'This machine crushes oranges.'

Given the linking hypothesis in (24a), the fact that the linguistic expression in (25) was judged to be acceptable by a native speaker of Gitksan in the context in (25a) and in the context in (25b) supports the conclusion that the linguistic expression is true in both contexts, thus supporting the aforementioned hypothesis. Thus, grouping together positive pieces of data that differ only in the context in which an expression is judged can

[21] We emphasize that the gloss and translation of non-English language examples, like (1), are not part of the piece of data but merely help the reader understand the important features of the expression that is responded to.

provide empirical evidence that the meaning of the expression is compatible with the contextually conveyed meanings.

Conversely, grouping together positive pieces of data that differ minimally in the linguistic expressions can provide empirical evidence that the linguistic expressions are all compatible with a particular meaning. Consider the hypothesis that the exclusives *only* and *alone* are both compatible with rank-order interpretations (Coppock and Beaver 2014). The context in (26) establishes that the fire alarm was not sounded because there was an actual fire emergency. Native speakers of English are taken to know that an actual fire emergency outranks a fire drill on, for example, a scale of danger. The fact that both (26a) and (26b) are judged to be acceptable in this context, and hence by the linking hypothesis in (24a) can be true in this context, shows that both exclusives are compatible with the so-called 'rank order' interpretation of exclusives.

(26) Context: Susan works at a school. She is in charge of testing whether the teachers are aware of the fire safety procedures. One day, she sounds the fire alarm and observes how the teachers guide their students to safety. Once they are all gathered outside, she informs everybody that this was not an actual fire emergency...

a. It was **only** a drill.

b. It was **just** a drill.

In sum, under the linking hypotheses in (24), positive pieces of data can provide empirical evidence for a wide range of hypotheses about meaning, namely any hypothesis about a felicity condition of an expression, as in the examples in (1) or (3), about the truth conditions of an expression, as in examples (4bi), (25) and (26), or about the implications that an expressions gives rise to, as in example (2). Under other linking hypotheses, positive pieces of data may provide empirical evidence for other types of hypotheses. Crucially, however, positive pieces of data alone cannot provide empirical evidence for hypotheses about which part of an expression contributes a meaning, or about which feature of context an expression is sensitive to. For these hypotheses, we need minimal pairs, which are discussed in section 5.3.

5.2 Evidence from negative pieces of data

A negative piece of data is one in which the speaker's response to the task about the linguistic expression was negative. What counts as 'negative' is again determined relative to the response task: e.g., a negative response to a binary acceptability judgment task is a judgment of unacceptability, to a binary truth value judgment task is a judgment of falsity, and to a direct implication judgment task is a judgment that the implication does not arise from the linguistic expression. As with positive pieces of data, the connection between a negative piece of data and the hypothesis about meaning that the piece of data is taken to support is provided by linking hypotheses, like those in (27).

(27) a. **Linking hypothesis for a judgment of unacceptability:** If an expression is judged to be unacceptable in a context, the expression is syntactically ill-formed, it is infelicitous in that context, or its truth conditions are incompatible with that context, or a combination thereof.[22]

[22] Chomsky (1977) points to this linking hypothesis when he writes: "we may make an intuitive judgment that some linguistic expression is odd or deviant. But we cannot in general know, pretheoretically, whether this deviance is a matter of syntax, semantics,

b. **Linking hypothesis for a judgment of falsity of a syntactically well-formed, felicitous and pragmatically unmarked sentence:** Given a sentence that is hypothesized to be syntactically well-formed, felicitous and pragmatically unmarked in the context in which it is judged, if that sentence is judged to be false in that context, then its truth conditions are incompatible with that context.

c. **Linking hypothesis for a judgment that an implication does not arise:** If an expression is judged to not give rise to an implication in a context, the expression does not give rise to that implication in that context.

Consider Mucha's (2013: 384f.) hypothesis that a Hausa sentence cannot simultaneously have both past and present temporal reference. From this hypothesis, one can derive the prediction that the linguistic expression in (6), repeated below, is infelicitous in the context provided. Given the linking hypothesis in (27a), the negative piece of data in (6) provides empirical evidence that B's utterance is syntactically ill-formed, infelicitous or false, or a combination thereof. Thus, under the linking hypothesis in (27a), the negative piece of data in (6) does not yet provide empirical evidence for Mucha's hypothesis.

(6) Context: For lunch, Hàwwa cooked beans and ate them. Audu is cooking beans for dinner right now. Is it appropriate to say:

#Hàwwa dà Audu sun dafà wākē yâu.
Hàwwa and Audu 3PL.COMPL cook beans today

Intended: 'Hàwwa and Audu cook/cooked beans today.'
Comment: The reading is not suitable for Audu. (Mucha 2013:385)

Under a different linking hypothesis, namely the one in (28), the negative piece of data in (6) provides empirical evidence for Mucha's hypothesis. (A clue that (28) is the linking hypothesis Mucha assumes is that she refers to the judgment elicited for (6) as a "felicity" judgment (p.384).)

(28) **Linking hypothesis for a judgment of unacceptability of a syntactically well-formed and true sentence:** Given an expression that is hypothesized to be syntactically well-formed and whose truth conditions are hypothesized to be compatible with the context in which the expression is judged, if the expression is judged to be unacceptable in that context, then the expression is infelicitous in that context.

Given the linking hypothesis in (28), the negative piece of data in (6) provides empirical evidence that B's utterance is infelicitous in the context provided. Thus, it is under the linking hypothesis in (28) that the piece of data in (6) is transformed into a piece of evidence for Mucha's hypothesis about temporal reference.

Like positive pieces of data, negative pieces of data can provide empirical evidence for a wide range of hypotheses about meaning under linking hypotheses like those in (27) and (28), namely any hypothesis about the violation of a felicity condition of an utterance, as in example (6), about non-fulfillment of the truth conditions of an expression, as in example (4bii), or about an implication that the expression does

pragmatics, belief, memory limitation, style, etc." (p.4).

not give rise to, as in example (18b). However, negative pieces of data alone, just like positive pieces of data, cannot provide empirical evidence for hypotheses about which particular sub-part of the expression contributes a meaning, or about which facet of context an expression is sensitive to. For these hypotheses, we need minimal pairs.

5.3 Evidence from minimal pairs

In phonology, where minimal pairs play a crucial role in the identification of phonemes, minimal pairs are discussed front and center in textbooks (e.g., Hayes 2008:20, Zsiga 2013:203, Odden 2014:16). A piece of data in phonology consists of a linguistic expression and the meaning of that expression (typically provided by a translation for non-English expressions), e.g. Paraguayan Guaraní *pytã* 'red'. A minimal pair consists of two expressions attested in the language that "are differentiated exclusively by a choice between one of two segments" (Odden 2014:16) and that have different meanings. For example, the pair of Paraguayan Guaraní expressions *pytã* 'red' / *-pyta* 'stay' is a minimal pair. Under the assumption that expressions that differ in exactly one segment and in meaning show that the varying segments are allophones of different phonemes of the language, the Paraguayan Guaraní minimal pair shows that the (stressed) vowels /ã/ and /a/ are allophones of different phonemes of the language. For a discussion of minimal pairs in syntactic research see Beavers and Sells 2014:410.

A piece of data in research on meaning is more complex than a piece of data in phonology and, consequently, there are two types of minimal pairs rather than just one. Specifically, a minimal pair in research on meaning consists of two pieces of data that differ minimally in either the linguistic expression, as in (29a), or in the context in which the expression is uttered, as in (29b), and that receive distinct responses.[23]

(29) **Types of minimal pairs in research on meaning**

a. Linguistic variants: The two pieces of data have the same context but minimally different linguistic expressions that receive distinct responses by native speakers.

b. Context variants: The two pieces of data have the same linguistic expression but minimally different contexts, in which the linguistic expression receives distinct responses by native speakers.

These two types of minimal pairs provide evidence for different types of hypotheses about meaning, as we show in the remainder of this section. We focus primarily on minimal pairs of piece of data with binary acceptability judgments. Minimal pairs of pieces of data with other responses are briefly discussed in section 5.3.3.

5.3.1 Minimal pairs of pieces of data with linguistic variants

A minimal pair of type (29a), in which both pieces of data have the same context but minimally different linguistic expressions that receive distinct acceptability judgments in the context, provides evidence that

[23] Minimal pairs may also consist of pairs of pieces of data with distinct tasks or where the responses are given by different populations of speakers. Since such types of minimal pairs do not provide evidence for hypotheses central to research on meaning, but rather for hypotheses about e.g. dependent measures and speaker variation, we do not discuss them here.

what differs between the two linguistic expressions contributes a particular meaning or results in a change in meaning. Consider the two pieces of data in (30). These share the same context, and the Paraguayan Guaraní linguistic expression in (30a) differs from the one in (30b) in the presence of the exclusive clitic *=nte* ‘only’ on the name *Javier*. The linguistic expression in (30a) was judged to be unacceptable in the context provided by four native speakers of Paraguayan Guaraní, whereas the linguistic expression in (30b) was judged to be acceptable by the same four native speakers. Thus, the two pieces of data in (30) form a minimal pair of type (29a).

(30) Context: Javier has a cow and Maria has a cow, too.

a. #Javiér**=nte** o-guereko vaka.
Javier=only A3-have cow
‘Only Javier has a cow.’

b. Javier o-guereko vaka.
Javier A3-have cow
‘Javier has a cow.’

Consider the hypothesis that the clitic *=nte* ‘only’ contributes an exclusive meaning like English *only*. Under this hypothesis, the sentence in (30a) would mean that Javier has a cow and nobody other than Javier has a cow. Thus, under this hypothesis, the example in (30a) is predicted to be false in the context in (30) since the context specifies that Javier and Maria each have a cow. Under the linking hypothesis in (31), the fact that (30a) was judged to be unacceptable by the native speakers means that (30a) is false.

(31) **Linking hypothesis for a judgment of unacceptability of a syntactically well-formed and felicitous sentence:** Given an expression that is hypothesized to be syntactically well-formed and to be felicitous in the context in which the expression is judged, if the expression is judged to be unacceptable in that context, then the truth conditions of the expression are not fulfilled in that context.

Note, however, that the unacceptability of (30a), under the linking hypothesis in (31), only provides empirical evidence for the hypothesis that the entire linguistic expression in (30a) is incompatible with an exclusive interpretation. In order to provide empirical evidence for the hypothesis that *=nte* ‘only’ contributes this exclusive meaning, the negative piece of data in (30a) is combined with the minimally different positive piece of data in (30b). This example differs from (30a) only in the absence of *=nte* ‘only’. Since it is judged to be acceptable, i.e., is true under the linking hypothesis in (24a), it is the combination of (30a) and (30b) that provides empirical evidence for the hypothesis that *=nte* ‘only’ contributes the exclusive meaning.

The linguistic expressions in minimal pairs of type (29a) may also differ in the order of parts of the expressions. In the minimal pair in (32), for example, the two Paraguayan Guaraní linguistic expressions differ in whether the counterfactual suffix *–mo’ã* ‘CF’ is realized inside the negation circumfix *nd–....-i*, as in (32a), or outside it, as in (32b).

(32) Context: Javier told me that he is not going to Asuncion tomorrow. I tell my mother:

a. Javier nd-o-ho**-mo’ã**-i Paraguaý-pe ko’ẽro.
Javier NEG-A3-go-CF-NEG Asuncion-to tomorrow
‘Javier is not going to Asuncion tomorrow.’

b. #Javier nd-o-ho-i-**mo'ã** Paraguaý-pe ko'ẽro.
Javier NEG-A3-go-NEG-CF Asuncion-to tomorrow
'Javier almost didn't go to Asuncion tomorrow.'

Consider the hypothesis that the truth conditions of sentences with *–mo'ã* 'CF' differ depending on whether *–mo'ã* 'CF' occurs inside the negation circumfix, as in (32a), or outside of it, as in (32b). This hypothesis leads to the prediction that there are contexts in which one sentence is true whereas the other one is false. Given the linking hypotheses in (24a) and (31), the minimal pair in (32) provides support for this hypothesis: (32a) is judged to be acceptable in the context in (32) and so, under the linking hypothesis in (24a), it is true; (32b), on the other hand, is judged to be unacceptable in the context in (32), and so, under the linking hypothesis in (31), is false. Again, both the positive and the negative pieces of data in (32) are required to provide empirical evidence for the hypothesis that the truth conditions of sentences differ depending on the position of pieces of data *–mo'ã* 'CF' with respect to negation. (For discussion of *–mo'ã* 'CF' see Tonhauser 2009.)

In minimal pairs of type (29a), the linguistic expressions may also differ minimally in their constitutive parts. In the St'át'imcets (Lillooet Salish) minimal pair in (33), the two linguistic expressions differ in whether the inferential evidential *k'a* 'INFER' or the sensory-non-visual evidential *lákw7a* 'SNV' occurs after the sentence-initial focus marker. Let's assume that we have already established that *k'a* 'INFER' is an evidential that contributes the information that the speaker's evidence for their assertion relies on inference. Consider now the hypothesis that the evidential *lákw7a* 'SNV' has a different meaning and, specifically, is incompatible with inferential evidence. This hypothesis predicts that (33b) is infelicitous in the context of (33) since the context is designed such that inferential evidence obtains. We also expect (33a) to be felicitous in this context.

(33) Context (inferential): You are a teacher and you come into your classroom and find a nasty picture of you drawn on the blackboard. You know that Sylvia likes to draw that kind of picture.

a. nílh**=k'a** s=Sylvia ku=xílh-tal'i
FOC=INFER NMLZ=Sylvia DET=do(CAUS)-TOP
'It must have been Sylvia who did it.'

b. #nilh **lákw7a** s=Sylvia ku=xílh-tal'i
FOC SNV NMLZ=Sylvia DET=do(CAUS)-TOP
'It must have been Sylvia who did it.' (Matthewson 2011a:94)

Given the linking hypotheses in (24a) and (28), the minimal pair in (33) provides empirical support for the hypothesis that *lákw7a* 'SNV' is incompatible with inferential evidence: (33a) is judged to be acceptable in the context in (33) and so, under the linking hypothesis in (24a), it is felicitous; (33b), on the other hand, is judged to be unacceptable in the context in (33), and so, under the linking hypothesis in (28), is infelicitous. Again, both the positive and the negative pieces of data in (33) are required to provide empirical evidence for the hypothesis that the felicity conditions of the sentences differ depending on whether *k'a* 'INFER' or *lákw7a* 'SNV' occurs.

In all of the examples we have presented thus far, the contextual information that is kept constant across the two members of the minimal pair appears before the target linguistic expression. But, of course, this contextual information may also appear after the linguistic expression. Such minimal pairs are used to show that the two minimally different expressions contrast in whether they allow the same follow-up. An example from St'á't'imcets is given in (34). The hypothesis being tested here is that St'át'imcets noun phrases realized only with the plural (discontinuous) determiner *i...a* do not enforce reference to the maximal contextual salient set of individuals, in contrast to noun phrases that also contains *tákem* 'all'. From this hypothesis, one can derive the prediction that the first clause of A's first utterance in (34a), which only realizes the plural determiner on the noun 'children', can be continued with the claim that not all children are hungry, but not the first clause of A's first utterance in (34b), which also contains the quantifier *tákem* 'all'.

(34) Context: A and B are working in a day-care. They are looking after 14 children.

a. A: Wa7 q'7-áol'men **i=**sk'wemk'úk'wm'it**=a**; cuystwí malh áz'-cit ku=s-q'a7
IPFV eat-want DET.PL=child(PL)=EXIS ADHT buy-APPL DET=NMLZ-eat
'The/Some children are hungry. Let's buy some food.'

B goes to buy some food. When she returns, A says:

A: Cw7it-7úl! Cw7áy=t'u7 kw=s=tákem i=sk'wemk'úk'wm'it=a wa7 q'7-áol'men.
many-too NEG=just DET=NMLZ=all DET.PL=child(PL)=EXIS IPFV eat-want
'That's too much! Not all the children are hungry.'

b. A: Wa7 q'7-áol'men **tákem i=**sk'wemk'úk'wm'it**=a**; cuystwí malh áz'-cit ku=s-q'a7
prog eat-want all DET.PL=child(PL)=EXIS ADHT buy-APPL DET=NMLZ-eat
'All the children are hungry. Let's buy some food.'

B goes to buy some food. When she returns, A says:

A: #Cw7it-7úl! Cw7áy=t'u7 kw=s=tákem i=sk'wemk'úk'wm'it=a wa7 q'7-áol'men.
many-too NEG=just DET=NMLZ=all DET.PL=child(PL)=EXIS IPFV eat-want
'That's too much! Not all the children are hungry.'

A St'át'imcets speaker judged the discourse in (34a) to be acceptable, but not the discourse in (34b). Under the linking hypotheses in (24a) and (27a), these judgments supports the hypothesis that the universal quantifier *tákem* 'all' does, but the plain plural determiner *i...a* does not, enforce reference to the entire set of contextually salient individuals in the discourse context.

In sum, minimal pairs of type (29a) can provide evidence that what differs between the two expressions of the minimal pair contributes a particular meaning, as in (30) and (34), or results in a change in meaning, as in (32) and (33). As discussed, both the positive and the negative pieces of data are necessary to provide empirical evidence for such hypotheses.

5.3.2 Minimal pairs of pieces of data with context variants

A minimal pair of type (29b), in which the same linguistic expression receives distinct acceptability judgments in the two minimally different contexts, provides evidence that the meaning of the linguistic expression is sensitive to what differs between the contexts.

The two contexts of the two pieces of data are minimally different if they only differ with respect to the hypothesis that is being explored. To illustrate, take the hypothesis that Paraguayan Guaraní sentences with the clitic *=nte* 'only' entail an exclusive meaning, specifically that the linguistic expression in (30a), repeated in (35), entails that Javier is the only person who owns a cow. Given this hypothesis, the two contexts in (35) are minimally different: in the context in (35a), Javier is not the only person who owns a cow, and in the context in (35b) he is the only person who owns a cow. Since the linguistic expression in the two pieces of data is the same, but receives distinct acceptability judgments in the two contexts, the two pieces of data in (35) form a minimal pair of type (29b). From the hypothesis under investigation we derive the prediction that the linguistic expression is false in the context in (35a) and true in the context in (35b).

(35) a. Context: Javier has a cow and Maria has a cow, too.

#Javiér=nte o-guereko vaka.
Javier=only A3-have cow

'Only Javier has a cow.'

b. Context: Javier has a cow and nobody else has one.

Javiér=nte o-guereko vaka.
Javier=only A3-have cow

'Only Javier has a cow.'

Given the linking hypotheses in (24a) and (31), the minimal pair in (35) provides empirical support for the hypothesis that the linguistic expression in the examples in (35) entails an exclusive meaning: the linguistic expression of the minimal pair is judged to be unacceptable in the context in (35a), and so, under the linking hypothesis in (31), its truth conditions are not fulfilled in that context; the same linguistic expression is judged to be acceptable in the context in (35b), and so, under the linking hypothesis in (24a), its truth conditions are fulfilled in that context. Thus, the positive piece of data in (35b) provides empirical evidence that the truth conditions of the expression are compatible with an exclusive meaning, and the negative piece of data in (35a) provides empirical evidence that the truth conditions of the expression are incompatible with a context that denies the exclusive meaning. Crucially, both the positive and the negative pieces of data in (35) are required to provide empirical evidence for the hypothesis that sentences with *=nte* 'only' entail an exclusive interpretation.

As a second example, consider the hypothesis that implicit subject arguments in Paraguayan Guaraní require a familiar third person antecedent discourse referent. From this hypothesis, we derive the prediction that a sentence with an implicit subject argument, like that in the examples in (36), is felicitous in a context that establishes a familiar third person antecedent discourse referent and infelicitous in a context that does not establish such a discourse referent. The minimal pair in (36) realizes the same linguistic expression in two contexts that differ only with respect to this hypothesis: the context in (36a) establishes a familiar third person discourse referent, but not the one in (36b).

(36) a. Context: We're sitting on the sidewalk drinking terere. A stray dog walks up to us and lies down in the shade at our feet. I say:

Kuehe che-su'u.
yesterday B1sg-bite

'Yesterday, it bit me.'

b. Context: We're sitting on the sidewalk drinking terere. I say:

#Kuehe che-su'u.
yesterday B1sg-bite
(Yesterday, it bit me.) (Tonhauser under review)

Given the linking hypotheses in (24a) and (28), the minimal pair in (36) provides empirical support for the hypothesis that implicit subject arguments in Paraguayan Guaraní require a familiar antecedent discourse referent.

Finally, consider the hypothesis that the Gitksan clitic *=ist* 'QUDD' signals that the utterance addresses the question under discussion. The contexts in the Gitksan minimal pair in (37) differ minimally in whether a question under discussion is established: the context in (37a) does not establish one and the one in (37b) does. From this hypothesis, we derive the prediction that the Gitksan sentence with *=ist* 'QUDD' is infelicitous in (37a) and felicitous in (37b).

(37) a. Context: Adam and Betty are married. Betty is a traveling saleswoman and she works in a number of different towns in the surrounding area. The two are having dinner and nobody has said anything yet. Betty suddenly says

#G̲a'a=hl Abbotsford win ahle'lsd-'y**=ist**.
LOC=CN Abbotsford COMP work-1SG.II=QUDD

'I worked in Abbotsford today.'

b. Context: Adam and Betty are married. Betty is a traveling saleswoman and she works in a number of different towns in the surrounding area. The two are having dinner and nobody has said anything yet. Adam suddenly asks Betty "Where are you working now?". Betty says:

G̲a'a=hl Abbotsford win ahle'lsd-'y**=ist**.
LOC=CN Abbotsford COMP work-1SG.II=QUDD

'I worked in Abbotsford today.'

Given the linking hypotheses in (24a) and (28), the minimal pair in (37) provides empirical support for the aforementioned hypothesis.

In sum, minimal pairs of type (29b) can provide evidence that the meaning of the linguistic expression realized in both members of the pair is sensitive to what differs between the two contexts. As discussed, both the positive and the negative pieces of data are necessary to provide empirical evidence for the hypotheses. With minimal pairs of type (29b), just like with minimal pairs of type (29a), the contextual information may follow the target linguistic expression. We provide such an example in the next section.

5.3.3 Minimal pairs for other response tasks

The two types of minimal pair we have illustrated above can, of course, also be formed with pieces of data that involve response tasks other than a binary acceptability judgment task. The same types of minimal pairs can also be formed with forced choice truth value judgments, as illustrated with (4), or with implication judgments: minimal pair type (29a) is illustrated in (18). And these minimal pairs can also be formed with pieces of data with (non-)binary responses to tasks. In Amaral and Cummins (2015), for example, speakers were presented with Spanish dialogues like the ones in (38), and asked to judge the acceptability of the answer on a 5-point Likert scale. The minimal pairs in this task consist of dialogues: both dialogues realize the same question (which is the target linguistic expression) and minimally different answers (i.e., different continuations): the answers differ in whether the presupposition of the question (that Victoria was the director in the past) is denied, as in (38a), or not, as in (38b). This is thus a minimal pair of type (29b).

(38) (adapted from Amaral and Cummins 2015:165)

a. A: ¿Sigue siendo Victoria la directora del departamento?
'Does Victoria continue to be the director of the department?'

B1: Sí, aunque antes Victoria no era la directora.
'Yes, although Victoria was not the director before.'

b. A: ¿Sigue siendo Victoria la directora del departamento?
'Does Victoria continue to be the director of the department?'

B2: Sí, Victoria sigue siendo la directora del departamento.
'Yes, Victoria continues to be the director of the department.'

Amaral and Cummins (2015) found that dialogues like the one in (38b) received significantly higher acceptability ratings than dialogues like the one in (38a). Under a (presumed) linking hypothesis that one answer is preferred over another if the acceptability judgments if the answers are significantly different, this finding supports the hypothesis that answers that do not deny a presupposition are preferred over answers that deny a presupposition.

5.4 Summary

As illustrated in this section, positive pieces of data, negative pieces of data, and pieces of data in minimal pair form can be transformed into pieces of evidence for particular theoretical hypotheses about meaning. Statements about how the speakers' responses are interpreted, also called linking hypotheses, provide the connection between the speakers' responses to the pieces of data and the theoretical concepts that play a role in the theoretical hypotheses about meaning, such as truth and felicity conditions.

Positive pieces of data, negative pieces of data and the two types of minimal pairs each provide empirical evidence for different types of hypotheses. Positive pieces of data can provide empirical evidence that the felicity or truth conditions of a linguistic expression are satisfied or fulfilled, respectively, in a particular context, or that the expression gives rise to a particular implication. Negative pieces of data can provide empirical evidence that the felicity or truth conditions of a linguistic expression are not satisfied or fulfilled, respectively, in a particular context, or that the expression does not give rise to a particular implication. And the two types of minimal pairs can provide empirical evidence that a particular part of an expression contributes a particular meaning or results in a change in meaning, or that the meaning of an expression is sensitive to a particular feature of the context.

6 Conclusions

Empirical evidence is at the very heart of research on meaning. In this paper, we have made a three-part proposal about empirical evidence. We first proposed in section 3 that a piece of data in research on meaning has four components: a linguistic expression, a context, a response (task) and information about the speakers that responded. We argued that pieces of data that include these four components are more likely to be robust, replicable and transparent. Our second proposal, in section 4, was that acceptability, similarity and implication judgment tasks are more likely than others, including paraphrase and translation tasks, to lead to robust, replicable and transparent pieces of data. And, finally, we proposed in section 5 that empirical evidence in research on meaning consists of positive or negative pieces of data or pieces of data in minimal pair form, together with a linking hypothesis. The heterogeneity of current practices in research on meaning provided the impetus for our making of these proposals. We hope that this paper may lead to a fruitful, collaborative process of discussing the nature of empirical evidence in research on meaning.

References

Abrusán, Márta and Kriszta Szendrői. 2011. Experimenting with the king of France: Topics, verifiability and definite descriptions. *Linguistics & Philosophy* 34(6):491–535.

Allan, Keith. 2001. *Natural Language Semantics*. Oxford: Blackwell Publishers.

Alrenga, Peter and Christopher Kennedy. 2014. *No more* shall we part: Quantifiers in English comparatives. *Natural Language Semantics* 22:1–53.

Amaral, Patrícia and Chris Cummins. 2015. A cross-linguistic study on information backgrounding and presupposition projection. In F. Schwarz, ed., *Experimental Perspectives on Presuppositions*, pages 157–172. Heidelberg: Springer.

AnderBois, Scott and Robert Henderson. 2015. Linguistically established discourse context: Two case studies from Mayan languages. In R. Bochnak and L. Matthewson, eds., *Methodologies in Semantic Fieldwork*, pages 207–232. Oxford: Oxford University Press.

Beavers, John and Peter Sells. 2014. Constructing and supporting a linguistic analysis. In R. J. Podesva and D. Sharma, eds., *Research Methods in Linguistics*, pages 397–421. Cambridge: Cambridge University Press.

Bhatt, Rajesh and Shoichi Takahashi. 2011. Reduced and unreduced phrasal comparatives. *Natural Language and Linguistic Theory* 29:581–620.

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

Bohnemeyer, Jürgen. 2015. A practical epistemology for semantic elicitation in the field and elsewhere. In R. Bochnak and L. Matthewson, eds., *Methodologies in Semantic Fieldwork*, pages 13–46. Oxford: Oxford University Press.

Bott, Lewis and Ira A Noveck. 2004. Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of Memory and Language* 51:437–457.

Bowern, Claire. 2008. *Linguistic Fieldwork: A Practical Guide*. New York: Palgrave Macmillan.

Cable, Seth. 2014. Reflexives, reciprocals and contrast. *Journal of Semantics* 31:1–41.

Cann, Ronnie. 2007. *Formal Semantics: An Introduction*. Cambridge: Cambridge University Press.

Chelliah, Shobhana L. 2001. The role of text collection and elicitation in linguistic fieldwork. In P. Newman and M. Ratliff, eds., *Linguistic Fieldwork*, pages 152–165. Cambridge: Cambridge University Press.

Chelliah, Shobhana L. and Willem J. de Reuse. 2011. *Handbook of Descriptive Linguistic Fieldwork*. New York: Springer.

Chemla, Emmanuel and Benjamin Spector. 2011. Experimental evidence for embedded scalar implicatures. *Journal of Semantics* 28:359–400.

Chierchia, Gennaro and Sally McConnell-Ginet. 2000. *Meaning and Grammar*. Cambridge, MA: MIT Press.

Chomsky, Noam. 1977. *Essays on Form and Interpretation*. New York: North-Holland.

Clark, Herbert H. and Michael F. Schober. 1992. Asking questions and influencing answers. In J. M. Tanur, ed., *Questions about Questions: Inquiries into the Cognitive Bases of Surveys*, pages 15–48. New York: Russell Sage.

Coppock, Elizabeth and David Beaver. 2014. Principles of the exclusive muddle. *Journal of Semantics* 31:371–432.

Cover, Rebecca and Judith Tonhauser. 2015. Theories of meaning in the field: Temporal and aspectual reference. In R. Bochnak and L. Matthewson, eds., *Methodologies in Semantic Fieldwork*, pages 306–349. Oxford: OUP.

Crain, Stephen and Cecile McKee. 1985. The acquisition of structural restrictions on anaphora. In *Proceedings of North East Linguistic Society (NELS) 16*, pages 94–110.

Crain, Stephen and Mark Steeedman. 1985. On not being led up the garden path: The use of context by the psychological parser. In D. Dowty, L. Karttunen, and A. Zwicky, eds., *Natural Language Parsing: Psychological, Computational and Theoretical Perspectives*, pages 320–354. Cambridge: Cambridge University Press.

Crain, Stephen and Rosalind Thornton. 1998. *Investigations in Universal Grammar: A Guide to Experiments on the Acquisition of Syntax and Semantics*. Cambridge, MA: MIT Press.

Crnič, Luka. 2014. Non-monotonicity in NPI licensing. *Natural Language Semantics* 22:169–217.

Crowley, Terry. 1999. *Field Linguistics: A Beginner's Guide*. Oxford: Oxford University Press.

Cruse, Alan. 2011. *Meaning in Language: An Introduction to Semantics and Pragmatics*. Oxford: Oxford University Press.

Culicover, Peter and Ray Jackendoff. 2010. Quantitative methods alone are not good enough: Response to Gibson and Fedorenko 2010. *Trends in Cognitive Sciences* 14:234–235.

Davis, Henry, Carrie Gillon, and Lisa Matthewson. 2014. How to investigate linguistic diversity: Lessons from the Pacific Northwest. *Language* 90:180–226.

de Marneffe, Marie-Catherine and Christopher Potts. to appear. Developing linguistic theories using annotated corpora. In N. Ide and J. Pustejovsky, eds., *The Handbook of Linguistic Annotation*. Berlin: Springer.

de Marneffe, Marie-Catherine and Judith Tonhauser. 2015. Inferring meaning from indirect answers to polar questions: The contribution of the rise-fall-rise contour. Manuscript under review, The Ohio State

University.

de Swart, Henriëtte. 1998. *Introduction to Natural Language Semantics*. Stanford, CA: CSLI Publications.

Deal, Amy Rose. 2011. Modals without scales. *Language* 87:559–585.

Deal, Rose Amy. 2015. Reasoning about equivalence in semantic fieldwork. In R. Bochnak and L. Matthewson, eds., *Methodologies in Semantic Fieldwork*, pages 157–174. Oxford: Oxford University Press.

Degen, Judith. 2015. Investigating the distribution of *some* (but not *all*) implicatures using corpora and web-based methods. *Semantics & Pragmatics* 8:11:1–55.

Deo, Ashwini. 2012. The imperfective-perfective contrast in Middle Indo-Aryan. *Journal of South Asian Linguistics* 5:3–33.

Dowty, David R., Robert E. Wall, and Stanley Peters. 1981. *Introduction to Montague Semantics*. Dordrecht: Reidel.

Elbourne, Paul. 2011. *Meaning: A Slim Guide to Semantics*. Oxford: Oxford University Press.

von Fintel, Kai. 2004. Would you believe it? The king of France is back! (Presuppositions and truth-value intuitions). In A. Bezuidenhout and M. Reimer, eds., *Descriptions and Beyond*, pages 315–341. Oxford University Press.

von Fintel, Kai and Lisa Matthewson. 2008. Universals in semantics. *The Linguistic Review* 25:139–201.

Frawley, William. 1992. *Linguistic Semantics*. Hillsdale, New Jersey: Erlbaum.

Geurts, Bart and Nausicaa Pouscoulous. 2009. Embedded implicatures?!? *Semantics & Pragmatics* 2(4):1–34.

Gibson, Edward and Evelina Fedorenko. 2010. Weak quantitative standards in linguistic research. *Trends in Cognitive Sciences* 14:233–234.

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

Ginzburg, Jonathan. 1996. Dynamics and the semantics of dialogue. In J. Seligman and D. Westerstahl, eds., *Language, Logic and Computation*, pages 221–237. Stanford, CA: CSLI Press.

Gutzmann, Daniel and Elena Castroviejo Miró. 2011. The dimensions of verum. In O. Bonami and P. Cabredo Hofherr, eds., *Empirical Issues in Syntax and Semantics 8*, pages 143–165.

Hayes, Bruce. 2008. *Introductory Phonology*. Oxford: Blackwell.

Heim, Irene. 1982. *The Semantics of Definite and Indefinite Noun Phrases*. Ph.D. thesis, University of Massachusetts, Amherst.

Heim, Irene and Angelika Kratzer. 1998. *Semantics in Generative Grammar*. Oxford: Blackwell.

Hellwig, Birgit. 2006. Field semantics and grammar-writing: Stimuli-based techniques and the study of locative verbs. In F. Ameka, A. Dench, and N. Evans, eds., *Catching Language: The Standing Challenge of Grammar Writing*, pages 321–358. Berlin: Mouton de Gruyter.

Hellwig, Birgit. 2010. Meaning and translation in linguistic fieldwork. *Studies in Language* 34:802–831.

Henderson, Robert. 2014. Dependent indefinites and their post-suppositions. *Semantics & Pragmatics* 7:1–58.

Hurford, R. James, Brendan Heasley, and Michael B. Smith. 2007. *Semantics: A Coursebook*. Cambridge: Cambridge University Press.

Jacobson, Pauline. 2014. *Compositional Semantics: An Introduction to the Syntax/Semantics Interface*.

Oxford: Oxford University Press.

Jacobson, Pauline. ms. What is — or, for that matter, isn't — 'experimental' semantics? In D. Ball and B. Rabern, eds., *The Science of Meaning*. Oxford: Oxford University Press.

Jakobson, Roman. 1959. On linguistic aspects of translation. In R. Brower, ed., *On Translation*. Cambridge, MA: Harvard University Press.

Katz, Jerrold J. 1976. A hypothesis about the uniqueness of natura language. In S. R. Harnad, H. Steklis, and J. Lancaster, eds., *Origins and Evolution of Language and Speech*, pages 33–41. New York: Annals of the New York Academy of Science.

Kearns, Kate. 2011. *Semantics*. London: Palgrave Macmillan.

Kennedy, Christopher and Louise McNally. 2005. Scale structure, degree modification, and the semantics of gradable predicates. *Language* 81:345–381.

Kibrik, Aleksandr E. 1977. *The Methodology of Field Investigations in Linguistics: Setting up the Problem*. Berlin: Mouton.

Krifka, Manfred. 2011. Varieties of semantic evidence. In C. Maienborn, K. von Heusinger, and P. Portner, eds., *Semantics: An International Handbook of Natural Language Meaning*, vol. 1, pages 321–358. Berlin: Mouton de Gruyter.

Kripke, Saul A. 2009. Presupposition and anaphora: Remarks on the formulation of the projection problem. *Linguistic Inquiry* 40:367–386.

Larson, K. Richard and Gabriel Segal. 2005. *Knowledge Of Meaning: An Introduction To Semantic Theory*. Cambridge, MA: MIT Press.

Lyons, John. 1995. *Linguistic Semantics: An Introduction*. Cambridge: Cambridge University Press.

Matthewson, Lisa. 1999. On the interpretation of wide-scope indefinites. *Natural Language Semantics* 7:79–134.

Matthewson, Lisa. 2004. On the methodology of semantic fieldwork. *International Journal of American Linguistics* 70:369–415.

Matthewson, Lisa. 2011a. Evidence about evidentials: Where fieldwork meets theory. In B. Stolterfoht and S. Featherston, eds., *Empirical Approaches to Linguistic Theory: Studies of Meaning and Structure*, pages 85–114. Berlin: Mouton de Gruyter.

Matthewson, Lisa. 2011b. Methods in cross-linguistic semantics. In K. von Heusinger, C. Maienborn, and P. Portner, eds., *Semantics: An International Handbook of Natural Language Meaning*, pages 268–285. Berlin: Mouton de Gruyter.

Matthewson, Lisa. 2015. On 'emphatic' discourse particles in Gitksan. Keynote talk at the Annual Meeting of the *Deutsche Gesellschaft für Sprachwissenschaft*, Leipzig, March 2015.

Moltmann, Friederike. 2013. The semantics of existence. *Linguistics & Philosophy* 36:31–63.

Mucha, Anne. 2013. Temporal interpretation in Hausa. *Linguistics & Philosophy* 36:371–415.

Newman, Paul and Martha Ratliff. 1999. *Linguistic Fieldwork*. Cambridge: Cambridge University Press.

Nicolae, Andreea C. 2014. Questions with NPIs. *Natural Language Semantics* 23:21–76.

Odden, David. 2014. *Introducing Phonology*. Cambridge: Cambridge University Press.

Payne, Thomas E. 1997. *Describing Morphosyntax: A Guide for Field Linguists*. Cambridge: Cambridge University Press.

Podesva, Robert J. and Devyani Sharma. 2014. *Research Methods in Linguistics*. Cambridge: Cambridge University Press.

Portner, Paul. 2005. *What is Meaning: Fundamentals of Formal Semantics*. Oxford: Blackwell.

Riemer, Nick. 2010. *Introducing Semantics*. Cambridge: Cambridge University Press.

Roberts, Craige. 2012. Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics & Pragmatics* 5:1–69. Reprint of 1996 publication.

Rojas-Esponda, Tania. 2014. A discourse model for *überhaupt*. *Semantics & Pragmatics* 7(1):1–45.

Saeed, John I. 2009. *Semantics*. Oxford: Wiley-Blackwell.

Sakel, Jeanette and Daniel L. Everett. 2012. *Linguistic Fieldwork: A Student Guide*. Cambridge: Cambridge University Press.

Samarin, William. 1967. *Field Linguistics: A Guide to Linguistic Field Work*. New York: Holt, Rinehart and Winston.

Sawada, Osamu. 2014. An utterance situation-based comparison. *Journal of Semantics* 37:205–248.

Schütze, Carson. 1996. *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. Chicago: University of Chicago Press.

Schütze, Carson T. 2008. Thinking about what we are asking speakers to do. In S. Kepser and M. Reis, eds., *Linguistic Evidence: Empirical, Theoretical and Computational Perspectives*, pages 457–484. Berlin: Mouton De Gruyter.

Schütze, Carson T. and Jon Sprouse. 2014. Judgment data. In R. J. Podesva and D. Sharma, eds., *Research Methods in Linguistics*, pages 27–50. Cambridge: Cambridge University Press.

Smith, Carlota S., Ellavina Perkins, and Theodore Fernald. 2007. Time in Navajo: Direct and indirect interpretation. *International Journal of American Linguistics* 73:40–71.

Soames, Scott. 1976. *An Examination of Frege's Theory of Presupposition and Contemporary Alternatives*. Ph.D. thesis, MIT.

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.

Syrett, Kristen and Todor Koev. 2014. Experimental evidence for the truth conditional contribution and shifting information status of appositives. *Journal of Semantics* Online first, doi: 10.1093/jos/ffu007.

Szmrecsanyi, Benedikt. 2015. *Grammatical Variation in British English Dialects: A Study in Corpus-Based Dialectometry*. Cambridge: Cambridge University Press.

Tanenhaus, Michael K., James S. Magnuson, Delphine Dahan, and Craig Chambers. 2000. Eye movements and lexical access in spoken-language comprehension: Evaluating a linking hypothesis between fixations and linguistic processing. *Journal of Psycholinguistic Research* 29:557–580.

Thieberger, Nick. 2011. *The Oxford Handbook of Linguistic Fieldwork*. Oxford: Oxford University Press.

Thomas, Guillaume. 2014. Nominal tense and temporal implicatures: Evidence from Mbyá. *Natural Language Semantics* 22:357–412.

Tonhauser, Judith. 2009. Counterfactuality and future time reference: The case of Paraguayan Guaraní *–mo'ã*. In *Proceedings of Sinn und Bedeutung 13*, pages 527–541.

Tonhauser, Judith. 2011. The future marker *–ta* of Paraguayan Guaraní: Formal semantics and cross-linguistic comparison. In R. Musan and M. Rathert, eds., *Tense Across Languages*, pages 207–231.

Tübingen: Niemeyer.

Tonhauser, Judith. 2012. Diagnosing (not-)at-issue content. In *Proceedings of Semantics of Under-represented Languages in the Americas (SULA) 6*, pages 239–254. Amherst, MA: GLSA.

Tonhauser, Judith. 2015. Cross-linguistic temporal reference. *Annual Review of Linguistics* 1:129–154.

Tonhauser, Judith. under review. Implicit anaphoric arguments in Paraguayan Guaraní. Under review for Estigarribia, B. (ed.) *Guaraní Linguistics in the 21st Century*, Leiden: Brill Publishing.

Tonhauser, Judith, David Beaver, Craige Roberts, and Mandy Simons. 2013. Toward a taxonomy of projective content. *Language* 89:66–109.

van Tiel, Bob, Emiel van Miltenburg, and Natalia Zevakhina Bart Geurts. 2014. Scalar diversity. *Journal of Semantics* .

Vaux, Bert and Justin Cooper. 1999. *Introduction to Linguistic Field Methods*. Munich: Lincom Europa.

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

Zimmermann, E. Thomas and Wolfgang Sternefeld. 2013. *Introduction to Semantics: An Essential Guide to the Composition of Meaning*. Berlin/Boston: Mouton de Gruyter.

Zsiga, Elizabeth C. 2013. *The Sounds of Language*. Oxford: Wiley-Blackwell.