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*

July 14, 2015
DRAFT – Feedback very welcome!

Abstract

Empirical evidence is at the heart of research on natural language meaning. Surprisingly, however, discussions of what constitutes evidence in research on meaning 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 is based primarily on insights we and our colleagues have gained in research on under-studied languages and in quantitative research using offline measures, but we intend the proposal to cover research on natural language meaning more broadly, including research on well-studied languages that the researcher may even control natively.

Our proposal has three parts. First, we argue that a complete 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 that provided the responses. Incomplete pieces of data fail to satisfy our three proposed objectives that data be stable, replicable and transparent. Second, we argue that some response tasks, namely acceptability and implication judgment tasks, are better suited than others (e.g., paraphrase and translation tasks) for yielding stable, replicable and transparent pieces of data. Finally, we argue that empirical evidence for a hypothesis about meaning consists of one positive piece of data, or two pieces in minimal pair form, plus a linking hypothesis about how the piece(s) of data provide support for the meaning hypothesis. We show that different types of minimal pairs provide evidence for different types of meaning hypotheses.

*For helpful discussions and constructive comments, we thank Ryan Bochnak, Judith Degen, Ashwini Deo, Agata Renans, Craige Roberts, Kevin Scharp, Mandy Simons and Hubert Truckenbrodt. 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 All previously unpublished Gitksan data in this paper are from Barbara Sennott and Vincent Gogag; ha'miiyaa! We also thank Beste Kamali and Murat Yasavul for providing and judging the Turkish examples, and 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.

1 Introduction

Theoretical research on meaning has been thriving for over 40 years now, at least since Montague 1970. However, even though empirical evidence for or against particular hypotheses about meaning is at the very heart of this research, there is almost no discussion in the theoretical literature of what constitutes evidence for an empirical generalization about meaning. 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 theoretical research in semantics and pragmatics. The paper advances a three-part proposal about the nature of empirical evidence, specifically about what a complete 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 empirical generalizations about meaning.

Our goal in advancing this proposal is to contribute to improving the empirical basis of semantic and pragmatic theories. As we show in the course of 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. At the very least, this heterogeneity suggests that it is time for a discussion of what constitutes empirical evidence in research on meaning. More importantly, however, we argue that some of the current practices have severe shortcomings: they stand in the way of replicating results both within the same language and across languages, they limit readers' ability to fully understand what the empirical evidence consists of, and they impede cross-linguistic comparison. We hasten to add that our own practices have not always been ideal (and we provide some illustrative examples below), and that we are not writing this paper to wag our fingers at others.

The following illustration of two widely attested, but dramatically different, practices gives a sense for the need for discussion. We use the hypothesis in (1) as our example.

(1) **Hypothesis:** A Turkish sentence in which a content question is embedded under the verb root *bil* 'know' can have a strongly exhaustive reading, i.e., such a sentence may have an interpretation whereby the individual denoted by the subject noun phrase of the attitude verb knows the complete answer to the embedded question.

Author A adopts a first practice that is widely attested in the field. To provide empirical evidence for the hypothesis in (1), this author presents the Turkish example in (2), together with a gloss and a translation; in the example, a question (*parti-ye kim-ler-in gel-dig-in-i* 'who came to the party') is embedded under the verb root *bil* 'know'. Author A claims that the Turkish sentence has a strongly exhaustive interpretation, i.e., that it can mean that Ali knows who came to the party and also knows who did not come to the party.[1]

[1]We follow the Leipzig glossing conventions (*https://www.eva.mpg.de/lingua/resources/glossing-rules.php*) to gloss 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, ANA = anaphoric expression, ATTR = attributive, BDY = information structure boundary, CF = counterfactual, CIRC.POSS = circumstantial possibility modal, CL.CNJ = clausal conjunction, CONT = continuous, 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.

(2) Ali parti-ye kim-ler-in gel-dig-in-i bil-iyor.
Ali party-DAT who-PL-GEN come-NMLZ-3SG-ACC know-IPFV
'Ali knows who came to the party.'

The first practice is thus to present an example (with a translation, if needed), along with a claim that the example has a particular meaning.

Author B adopts a different practice. To provide empirical evidence for the hypothesis in (1), she provides the context in (3), which describes a situation in which Ali knows who came to the party his sister threw, and also knows who didn't come to that party. Author B also provides the example in (2) and asserts that the native speakers of Turkish she consulted judged (2) to be acceptable in the context in (3).

(3) Ayse's parents were out of town last weekend, and Ayse threw a party. Ayse's brother Ali knows who Ayse invited to the party, and since he attended the entire party, he also knows who came to the party and who, of the invited, didn't come. When Ayse's parents return, Ayse's mother notes that something was stolen from the house and so Ayse fesses up to the party. Ayse's mother wants to interrogate people who came to the party. Ayse tells her mother that she doesn't have a good overview of who came but...

Author B then comments that the Turkish sentence in (2) is compatible with a strongly exhaustive interpretation of the embedded question because the context in (3) describes a situation in which Ali knows who came and who didn't come to the party, and the sentence in (2) was judged to be acceptable in this context.

These practices are obviously different. For Author B, the empirical evidence for the hypothesis in (1) consists of i) a context that is compatible with a strongly exhaustive question interpretation, ii) the sentence in (2) that is uttered in this context, iii) a statement that the sentence in (2) is judged to be acceptable by a native speaker in this context, and iv) a comment about how the acceptability judgment of (2) in the context in (3) supports the claim that (2) is compatible with a strongly exhaustive interpretation (this is the linking hypothesis). Author A, on the other hand, merely provides the example in (2), together with a claim that the example has a strongly exhaustive interpretation. While it may be obvious that Author A's practice is less ideal than Author B's with respect to a claim about Turkish, a language under-represented in research on meaning, the former practice is extensively used in theoretical research on meaning on better-studied languages and those that are more widely spoken by semanticists, including English, German, Dutch, Italian, French, Greek, and Japanese. For such languages, there are numerous cases in published papers where empirical evidence for a hypothesis about meaning is made by following the practice of Author A.

In view of the differing practices in contemporary research, we hope to kick off a collaborative process of developing more rigorous and consistent standards for what counts as 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. Furthermore, some parts of our proposal have long been implemented in quantitative research on meaning. This paper thus synthesizes and builds on insights from these various strands of research in discussing and developing standards for what counts as empirical evidence in research on meaning. We also hope that this paper can serve as a resource for those wishing to undertake research on meaning.

Our three-part proposal focuses on issues that are fundamental to any research project involving empirical evidence about meaning comprehension, regardless of whether the evidence is collected through one-on-one elicitation with theoretically untrained 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.[2] We propose in section 3 that a complete piece of data has four parts: a linguistic expression, a context in which the expression was uttered,[3] a response to a task about that expression uttered in that context, and information about the native speakers. Incomplete pieces of data, which lack one or more of these components, fail to satisfy our three proposed objectives that data be stable, replicable and transparent. Second, we argue (section 4) that some response tasks, namely acceptability and implication judgment tasks, are better suited than others (e.g., paraphrase and translation tasks) for yielding stable, 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 of the linking hypothesis, i.e., how the pieces of data provide support for the meaning hypothesis. We show that different types of minimal pairs provide evidence for different types of hypotheses about meaning. The paper concludes in section 6.

2 Previous discussions of the nature of empirical evidence

Empirical evidence for a particular hypothesis about meaning cannot be read off of an uttered form. Rather, which meaning a speaker conveys when uttering an expression is only indirectly revealed, e.g., through responses to tasks about the utterance in context (for this point, see also Matthewson 2004, Bohnemeyer 2015, Deal 2015). Although pedagogical 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, not a single one of them contains any substantial discussion of the nature of empirical evidence, let alone a comprehensive guide to how hypotheses about meaning are empirically supported.[4] This lacuna is particularly surprising given that many textbooks point out the central importance that empirical evidence plays in 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

[2]Empirical evidence may also come from corpus studies. An expression attested in a corpus may, for instance, constitute a positive piece of data under the assumption that expressions in corpora are 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:sections 2.6 and 3. Since our focus in this paper is on empirical evidence collected through obtaining responses (including judgments) from native speakers, we largely 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.

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

[4]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.

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". Chierchia and McConnell-Ginet (2000:5f.) call speakers' judgments "the core of the empirical data against which semantic theories must be judged".

The most detailed discussion of the nature of empirical evidence is provided by Chierchia and McConnell-Ginet (2000). These authors acknowledge that there is a non-trivial process involved in working with speaker judgments: they observe that "[s]uch judgments are often very subtle, and learning how to tap semantic intuitions reliably and discriminate among the distinct phenomena that give rise to them is an important part of learning to do semantics" (p.6). Furthermore, they briefly illustrate that there is work to be done in getting from judgments to analysis (e.g., pp.42f.). But, overall, semantics/pragmatics textbooks merely acknowledge the central importance of empirical evidence in research on meaning, but provide no systematic discussion of the question of what constitutes a piece of data in research on meaning, which tasks can be performed reliably with native speakers, or how minimal pairs of pieces of data are used to provide empirical support for a generalization about meaning.

Volumes about research methods, including fieldwork methods, also do not discuss what constitutes empirical evidence in research on meaning.[5] 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.[6] However, these resources do not discuss in any detail how a speaker's response is combined with a context and a linguistic expression to constitute a piece of data. 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 a linguistic hypothesis 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 speaker(s) that provided the judgment.

Works specifically devoted to the methodology of research on meaning have only begun to appear within

[5]The works on which we base this claim are 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.

[6]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.

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 discuss in detail 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 theoretically untrained native speakers (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. However, quantitative research on meaning inherits and builds on the principles of experimental design, methodology, and quantitative analysis used in research in the cognitive and social sciences. As a consequence, some of the proposals we advance here are already established practice in quantitative research on meaning. For instance, quantitative research already considers a piece of data to include not just a linguistic expression, speakers' responses, and information about the speakers, but also often a context in which the linguistic expression is uttered (sometimes in the form of instructions to the speakers). Literature based on that research 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). Finally, quantitative research on meaning 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. The proposal about empirical evidence in research on meaning that we spell out in the next three sections can thus be regarded as an extension of established practices in quantitative research on meaning. It is important to note, however, that the focus of our proposal is on empirical evidence in research about meaning conducted through offline measures, e.g., to the exclusion of response time or eye movement measures, and to the exclusion also of quantitative production studies.[7]

3 Pieces of data in research on meaning

In this section, we propose that a complete piece of data in (offline) research on meaning consists of four components: a linguistic expression, a context in which that expression is uttered, a response to a task about

[7] An important question is whether some research methodologies provide more robust support for empirical generalizations than others, e.g. by virtue of relying on larger numbers of speakers, larger numbers of pieces of data, and quantitative analysis (for debate, primarily in the domain of syntax, see 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 discussed in this paper — namely, which components make up a piece of data and how combinations of pieces of data provide evidence for empirical generalizations — arise regardless of which methodology is used to collect the pieces of data.

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 *stable*, *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 lack a context, a response, information about the speakers, or a combination thereof have severe short-comings. Specifically, we argue that such incomplete data are not stable because they do not control for the context- and speaker-dependency of natural language interpretation; they are methodologically dispreferred because they impede replication; and they are not transparent because they do not make fully explicit how the piece of data supports the empirical generalization.

3.1 The four components of a complete piece of data

A piece of data in research on meaning is complex, with four components. Each of these components contributes in a different way to making the piece of data stable, replicable and transparent. The following four subsections characterize each of the four components of a complete piece of data.

3.1.1 The context of a piece of data

There are three types of context that play a role in the interpretation of natural language expressions. The first is the utterance context, which includes information about the speaker, the addressee(s), the time and the location of the utterance, and entities and eventualities that the interlocutors are currently attending to. The utterance context 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. A second type of context is the linguistic context, which consists of utterances previously made by the interlocutors. For instance, the referent of the English definite noun phrase *the cup* in the two-sentence discourse *Joan dropped a cup and a spoon. The cup broke* is taken to be the cup introduced in the first sentence, namely the cup that Joan dropped. The third type of context can be characterized as the structure of the discourse the linguistic expression is part of (e.g., Roberts 1996/2012): this includes information about the topic of conversation (also called the question under discussion) as well as the goals and intentions of the interlocutors. To illustrate the importance of the discourse structure in interpretation, consider an utterance of the English sentence *It's raining*. In uttering this sentence, a speaker intends to convey different information 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). All three of these types of context contribute to the common ground of the interlocutors engaged in conversation. When we refer to 'context' in this paper, we refer to the combination of these three types of context.

Given the complexity of the context that plays a role in natural language interpretation, it is clear that the context of a piece of data typically cannot be the entire context in which a linguistic expression 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 an expression is uttered and interpreted, because resources are limited (e.g., the cognitive capacities of the speakers who have to understand 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 in (4), 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 in (5), from Mbyá Guaraní, consists of a question inquiring about an individual, together with a description of the situation in which the question is uttered.

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

B: Su-nà̀ màganà̀.
3PL-CONT talk

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

(5) 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 (6) illustrates a context which establishes information about the prior discourse structure. The linguistic expressions in (6b) are uttered in the context of the discourse in (6a).

(6) 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 part of what is in the common ground, e.g., who slapped who in the examples in (7).

(7) 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 [(7ai,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 [(7aii)].

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. Rather, it is up to the researcher to decide which features of context to control for. 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. Under such circumstances, 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 (5), 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 and in the publication.

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 (4), (6) and (7), other possible linguistic expressions are sentences in the interrogative or imperative moods, multi-sentence discourses, as in (5), or sub-sentential expressions, as in (8B).

(8) 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). Even when minimal information about prosody is conveyed using capital letters, written representations of linguistic expressions abstract away from the prosodic realization of the utterance at the time at which it was judged by the native speaker. Thus, 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 prosody is indicated with capital letters.

From the above characterization of the linguistic expression of a piece of data in research on meaning it follows that certain expressions do not qualify. (We have observed instances of all of these in papers or research presentations.) First, an expression from one language that includes an expression from another language does not qualify as the linguistic expression of a piece of data since native speakers cannot judge the mixed-language expression — unless, of course, they are native speakers of both languages, and code-switching is plausible given the context of the piece of data. Second, an expression from one language that is followed by an expression from another language, e.g., to illustrate the (un)availability of a particular interpretation, is not a linguistic expression of a piece of data (again, unless the speakers asked to judge the two expressions are native speakers of both languages and code-switching is contextually supported). Finally, it is not appropriate to only present part of the sentence that speakers judged if it is the entire sentence that gives rise to the interpretation under investigation: for instance, it is not appropriate to only present the antecedent of a conditional with a presupposition trigger, representing the consequent with dots (...), to provide empirical evidence for the claim that the conditional has a particular presupposition.

3.1.3 The response task and response

Research on meaning is conducted using a plethora of research methods, including acceptability and truth value judgment tasks, production tasks, and reaction time and self-paced reading studies (see section 4, and Krifka 2011 and Bohnemeyer 2015 for overviews). Given the large variety of response tasks in research on meaning, the response component of a piece of data can take many forms, including a variety of verbal responses, such as judgments (e.g., 'Yes' or 'Jane didn't read all the books'), and a variety of non-verbal responses (e.g., mouse clicks, eye movements).[8] The response component of a piece of data is a native speaker's response to a task posed by the researcher with respect to a linguistic expression uttered in a context. 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). Particular response tasks, including acceptability judgment tasks, can be implemented with different response options, such as forced-choice binary responses, responses on a Likert scale, and magnitude estimations (see Schütze and Sprouse 2014 for an overview).

[8]In one-on-one elicitation, speakers may also indicate judgments of unacceptability through verbal responses other than the expected response options, including 'huh?', 'what?', 'that's funny', and even through non-verbal responses such as head-shaking, laughing, or frowning.

Since a particular response to a linguistic expression can only be interpreted in relation to the task that was used to elicit the response, the response task of a piece of data needs to be described:[9]

(9) **Description of the response task**

The description of a response task includes information about

a. the instructions given to the native speaker,

b. the specific question posed to the native speakers,[10]

c. how the linguistic expression, the context and the question were presented to the speakers, and

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

Works reporting results from quantitative research typically include the information in (9) in methods sections. In works that present pieces of data collected through introspection or one-on-one elicitation, such information — if it is included at all, see section 3.2 — is sometimes included as part of the data. For instance, the piece of data in (10), 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.).

(10) 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)

Other researchers opt to describe the type of judgment that was elicited and the speakers' responses in the text preceding the piece of data. Even in works that present pieces of data collected through introspection or one-on-one elicitation, it is often feasible to provide the information about the response tasks used in the research in a single location, akin to a methods section, early on in the work.

3.1.4 Information about the native speakers who provided judgments

Since different speakers may give different responses to a piece of data, pieces of data in research on meaning need to include information in the main body of the text about the speakers who responded to the tasks:

(11) **Information about the speakers to report**

a. The number of speakers that responded to the task,

[9] In quantitative research on meaning, especially research that relies on online measures, more detailed information about the response task is usually provided, including information about whether speakers had to give the response in a particular time frame.

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

b. the language background of these speakers, including whether they speak different dialects of the language under investigation (or come from different areas, in case dialects have not yet been established for the language), and

c. whether the speakers have had linguistic training,

d. any disagreements among the speakers that responded.

Works reporting results from quantitative research typically include information about (11a, b, c) in a methods section, whereas information about (11d) is reported in the results section. In works that present data collected through introspection or one-on-one elicitation, information about (11a, b, c) can be included early on in the work, in a section akin to a methods section, and information about (11d) can be included in connection with any pieces of data where such information is relevant.

From what we have said so far, it should be clear that we maintain that both theoretically trained and theoretically untrained native speakers can give responses. Pieces of data that involve a native speaker semanticist responding to a task about a linguistic expression in a context are an important source of evidence for empirical generalizations in contemporary research about meaning. This is true whether the researcher also serves as a speaker who responds to the response task (i.e., in introspection) or not (e.g., when a semanticist consults with a colleague).

However, we argue that while native speaker researchers can rely on their own responses to response tasks, they should not simply report their intuitions about meaning as support for an empirical generalization. The distinction between a response (including a judgment) and an intuition is important here. An intuition, as we use the term, is merely a speaker's impression or belief about meaning. Native speaker researchers rely on such intuitions all the time in the course of developing hypotheses about meaning. Importantly, however, these hypotheses should be empirically tested via response tasks. For example, suppose Authors A and B from section 1 above happen to be native speakers of Turkish. Both authors might start with the intuition that the example in (2) has a strongly exhaustive interpretation. Only Author A reports the intuition directly; Author B tests her hypothesis via a response task. In endorsing Author B's practice, we argue against the pervasive current practice of reporting an intuition about meaning, together with a linguistic expression, as support for an empirical generalization. As we propose in the next section, such a practice has the disadvantage of not yielding pieces of data that are stable, replicable and transparent.

3.2 Complete pieces of data are stable, replicable and transparent

In the preceding section, we characterized the four components that we argue make up a piece of data in research on meaning: a context, a linguistic expression, information about the response task and the responses, and information about the native speakers who responded. This proposal for a complete 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 (6) from Rojas-Esponda 2014, as well as in some research based on one-on-one elicitation, as illustrated with the example in (7) from Cable 2014. (Both of these authors provide information in their papers about who provided the relevant judgments: the author herself in the case of (6) and a French speaker in the case of (7).) However, the vast majority of contemporary research on meaning

relies on pieces of data that lack a context, information about the response task and the response, information about the responding native speakers, or a combination of the three. In fact, a survey of 40 recent journal articles we conducted[11] established that almost every paper surveyed included such incomplete data: almost half of the papers either exclusively or almost exclusively presented pieces of data consisting only of a linguistic expression (usually a sentence), and it was not rare to find pieces of data consisting of a linguistic expression and a context, but no information about the response (task), or 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 present 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., and is observed even when the authors of the paper are not native speakers of the language under investigation — which suggests that these non-native speakers provided the judgments.[12]

It is this heterogeneity of what is taken to be a piece of data in contemporary research on meaning that, in part, motivated us to write this paper. Our proposal that a piece of data should consist of the four components characterized in the previous section is guided by the objective that pieces of data that inform theories of meaning should be stable, replicable and transparent.

(12) **Objective:** A piece of data in research on meaning

a. is **stable**, i.e. includes information about factors that may lead to variation in speaker judgments,

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

c. is **transparent**, i.e. makes fully explicit how it supports the empirical generalization.

In the remainder of this section, we motivate that complete pieces of data, in contrast to incomplete ones, are stable, replicable and transparent.

[11] We surveyed 40 journal articles published between 2012 and early 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 from our survey that mostly or exclusively relied on secondary sources. These 40 articles cover a wide range of empirical phenomena and include data collected through introspection, one-on-one elicitation, quantitative research and corpus research. We examined each of the articles for what is considered a piece of data and the response tasks used.

[12] Some might suppose that the practice of non-native speakers giving responses, including acceptability judgments, is unproblematic, at least for English, since the journal review process would surely catch any errors in such a widely-spoken language. This is not the case, however. There is no guarantee that reviewers are native English speakers, reviewers often offer their own judgments even when they are not native speakers, and there are attested cases of spurious English judgments making it into print. For example, Moltmann (2013:36) argues that *Socrates is a man* "does not sound right" and marks it with '??'. However, in the judgment of our second author, this is an acceptable English sentence. We can only assume that the example was presented in Moltmann 2013 under the assumption of a context for the example and a referent for *Socrates* that render the example less than acceptable. But without providing such information, marking the example with '??' is not justified.

3.2.1 Stable pieces of data

The first objective in (12a) is for pieces of data to be stable, i.e., to include information about factors that may lead to variation in native speaker judgments. 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 of the language under investigation is affected by context. There is also much evidence that speakers may vary in their responses depending on, for instance, their dialect (e.g., Szmrecsanyi 2015) and whether they have had linguistic training (Schütze 1996:§4.4.1), but also depending on their handedness (Schütze 1996:§4.3.2) and their literacy and education (Schütze 1996:§4.4.2). We thus argue that the context and information about the responding native speakers are constitutive of a stable piece of data in research on meaning.

Some researchers might argue that a context may be omitted when the linguistic expression of the piece of data is claimed to be acceptable in *any* context. We reject this assumption since it confuses the piece of data with the hypothesis the piece of data is supposed to provide support for. If the hypothesis is that the linguistic expression under investigation is judged to be acceptable in any context, then a supporting piece of data is not a piece of data that lacks a context, but rather pieces of data with a variety of different contexts in which the linguistic expression is judged to be acceptable. (For reasons of space we cannot go into the question here of how to assemble a suitable set of contexts.)

Of course, if the linguistic expression was judged or responded to without having been presented in a context, then no context can be included as part of the data. However, for many of the offline response tasks frequently used in research on meaning, such as acceptability or truth value judgment tasks, omitting the context does not yield stable pieces of data. Since no utterance is ever made in a completely empty (null) context, devoid of any information about e.g., the speakers and the addressees, or about what information the interlocutors (do not) share, chances are high that speakers who are asked to respond to a task about linguistic expressions in null contexts imagine some context and respond relative to the context that they imagined. In this case, the researcher is not privy to that context, and hence does not know which features of the context may have led to the judgment. 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). (See Tonhauser 2015:144 for a critique of the use of null contexts in research on temporal and aspectual reference.)

It is sometimes assumed that omitting a context means that the linguistic expression was judged or responded to in a neutral, out-of-the-blue context. But, an out-of-the-blue context is not a null context. Rather, an out-of-the-blue context is one in which a speaker makes an utterance in a situation in which the interlocutors have very little or no common ground:

(13) **Out-of-the-blue context**

The context of a piece of data is an out-of-the-blue context when it describes a situation in which the interlocutors lack information about some aspect of the context, e.g., the interlocutors lack information about the utterance situation, the interlocutors have not engaged in prior discourse, or the interlocutors are unaware of each other's goals and intentions.

Thus, crucially, just like with other types of contexts, it is up to the researcher to control the information conveyed in an out-of-the-blue context and to clarify in which sense it is out-of-the-blue. Simply omitting all information about the context does not make for an out-of-the-blue context.

The following two examples illustrate different out-of-the-blue contexts. In the first example in (14), from Paraguayan Guaraní, the context provides information about the speaker and the addressee, as well as some information about the situation in which the utterance is made. 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 out-of-the-blue context of (14) 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. In other words, there is very little information in the common ground of the interlocutors.

(14) 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)

Since (14) was judged to be acceptable by four native speakers of Paraguayan Guaraní, this piece of data provides empirical evidence for the aforementioned hypothesis.

The second example of an out-of-the-blue context, in (15), comes from Gitksan. The hypothesis that was tested with this example was that the discourse particle *=is(t)* 'QUDD' indicates a downdate of the question under discussion (Gutzmann and Castroviejo Miró 2011), i.e., is unacceptable in a context in which its prejacent (here, the proposition that Charlie is sick) does not answer the current question under discussion (Ginzburg 1996, Roberts 2012). In contrast to the out-of-the-blue context in (14), where it was important to establish that the speaker and the addressee don't know each other and hence have a very limited common ground, the out-of-the-blue context in (15) establishes that the speaker and the addressee know each other (they are eating dinner together). The crucial feature of the context that is controlled in (15) is that there is no prior linguistic context: Adam and Betty have not yet raised a topic of conversation and, in particular, nothing about Charlie's state of health has been part of the conversation so far between the two.

(15) Context: Adam and Betty are eating dinner quietly. Nobody has said anything yet. Betty suddenly says:

#siipxw-t Charlie/Tsaalii=is(t)
sick-3SG.II Charlie=QUDD

'Charlie is sick.' (Matthewson 2015:(7), slide 28)

The fact that (15) was judged to be unacceptable by two native speakers of Gitksan supports the hypothesis about *=is(t)* 'QUDD'.[13]

3.2.2 Replicable pieces of data

The second objective in (12b) is for pieces of data to be replicable, i.e., to maximally facilitate replication of the piece of 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 (modulo the speakers or the language of the linguistic expression) is collected. If the researcher attempting replication has complete data available, this will be possible, but if the data is incomplete, replication is hampered. For example, if the context, the response or information about the response task is missing from the piece of data, the replicating researcher may use a different context or employ a different kind of response task. It is then impossible to ascertain whether any observed differences between the original data and the replication are due to linguistically interesting (inter-speaker or cross-linguistic) variation, or merely due to the replicating researcher having used slightly different pieces of data in their work.

Current practice in the field suggests that some researchers do not believe that it is important to specify the task to which native speakers responded, or the response given by the speakers. Possibly, what is assumed is that the diacritic that accompanies the linguistic expression (or the absence of such a diacritic) conveys which task was responded to (specifically, which judgment was elicited): e.g., if the example is marked with an asterisk (*), then the example was judged to be syntactically ill-formed (i.e., unacceptable for syntactic reasons). There are several problems with this argument. First, many writers fail to specify what the diacritics stand for (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). Others, e.g., Falaus (2014) and Coppock and Beaver (2014), use both the hash mark (#) and the asterisk (*) but do not comment on which judgment differences the two diacritics reflect. In fact, in at least a quarter of the 40 recent papers we surveyed, the meaning of the diacritics used was unclear.

However, even if the meanings of the diacritics were well-defined, diacritics do not identify the task the speaker responded to, or which response was given. Rather, the diacritics indicate the theoretical interpretation of a 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. Thus, the diacritic does not replace information about the response task or the response.

[13] Initially, the two native speakers judged (15) to be acceptable, apparently counter-exemplifying the hypothesis being tested. However, additional questioning revealed that the speakers had silently enriched the context in order to make the particle acceptable; one speaker had assumed that the addressee of (15) had previously asked about Charlie. When the speakers were explicitly reminded that Charlie had not previously been discussed in the context, they both rejected (15), as predicted by the hypothesis. The problem of speakers being willing to silently enrich contexts becomes more acute the more minimal the context they are given. This observation reinforces our claim that null contexts should not be utilized.

3.2.3 Transparent pieces of data

The third objective in (12c) 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 how the piece of data provides empirical support for the generalization, regardless of whether the readers are native speakers of the language of the piece of data. Take, for instance, a piece of data that is incomplete because the context and information about the response is missing, like (2), from section 1. Recall that our hypothetical Author A used this example to provide support for the hypothesis that questions embedded under *bil* 'know' are compatible with a strongly exhaustive interpretation.

(2) Ali parti-ye kim-ler-in gel-dig-in-i bil-iyor.
Ali party-DAT who-PL-GEN come-NMLZ-3SG-ACC know-IPFV
'Ali knows who came to the party.'

A reader who is a native speaker of the language of the piece of data (Turkish, in this case) may be able to reconstruct a context for the expression or to judge the acceptability of the expression in some suitable context in order to independently verify that (2) provides empirical evidence for the hypothesis. However, even with native speakers there is no guarantee that the reader would reconstruct the same context or give the same response. A reader who is not a native speaker of the language under discussion, be it Turkish, English or German, cannot reconstruct either context or response, and hence does not have access to the complete piece of data. Given that a complete piece of data is needed in order to assess how the piece of data provides empirical evidence for the hypothesis, it follows that incomplete pieces do not allow all readers of a work to fully understand how the piece of data provides empirical evidence for the hypothesis. In other words, providing incomplete pieces of data unfairly privileges native speakers of the language of the piece of data.

3.3 Summary

In this section, we proposed that a piece of data in research on meaning consists of four components, summarized in (16). We characterized these four components and argued that pieces of data that lack one or more of the components have severe shortcomings, namely that they are less stable than complete pieces of data, that they stand in the way of replication within the same language and across languages, and that they privilege native speakers' understanding of the piece of data and therefore are not transparent.

(16) **Complete pieces of data in research on meaning**
A complete 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,
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) that responded.

We suggested that information about the response task and about the speakers should be included in the main body of the text, e.g., a methods section, as is already common practice in quantitative research.

We suspect that some of our colleagues may find our arguments generally convincing, but may nevertheless be resistant to the idea of adopting (16) as standard information for pieces of data in research on meaning, for a variety of reasons. Wouldn't this information clutter up theoretical research papers? Can't we just trust the researchers that the empirical generalizations reported were established using complete pieces of data, even if only parts of those data are presented in the work? And isn't this information really only required for quantitative research or research on languages that the researcher does not speak natively?

It should be clear by now that we believe that the shortcomings of incomplete pieces of data by far outweigh any perceived advantages of keeping theoretical research papers focused on theory by including incomplete pieces of data, especially since information about response tasks and the native speakrs who responded can often be provided in a fairly short paragraph in papers that rely on introspection and one-on-one elicitation. Since theories of meaning are only as good as the empirical generalizations they capture, and empirical generalizations are only as good as the pieces of data that are provided to support them, research on meaning cannot afford to apply different standards for what counts as a piece of data depending on the methodology by which those data were collected.

4 Judgment tasks and other tasks in research on meaning

Having established the four components of a complete piece of data, we now characterize the main types of response tasks used in offline research on meaning (section 4.1).[14] We then argue that some response tasks, namely acceptability and implication judgment tasks, are better suited than others (e.g., paraphrase and translation tasks) for yielding stable, replicable and transparent pieces of data (section 4.2).

4.1 Characterization of tasks

In this paper, 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. In doing so, we 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.

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. Questions that might be posed to the speaker include 'Does this sound good to you?' or 'Would you say this?' (see (10) and Bohnemeyer 2015:36 for

[14] Krifka (2011) and Bohnemeyer (2015) mention a number of other offline tasks that we cannot go into detail about here for reasons of space, including the pointing task, the act-out task, and the association task.

further examples).[15] 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 will judge a linguistic expression uttered in a context to be acceptable if and only if the linguistic expression is syntactically well-formed, felicitous[16] and true in that context. Thus, an acceptability judgment should be 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.

Another assumption is that native speakers will judge the utterance to be unacceptable if the utterance is syntactically ill-formed, if it is infelicitous (even if it is syntactically well-formed and true), 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, true and felicitous, but a judgment of unacceptability does not by itself provide insight into why the utterance was judged so (see also Matthewson 2004:409). Consider, for example, the English example in (17), which is judged (by the second author) to be unacceptable in the context in which it is presented. (Here, we use the dollar symbol ($) to indicate that the linguistic expression was judged to be unacceptable in the context in which it was presented.)

(17) Context: John arrived in Hamburg yesterday.
$ He arrived tomorrow.

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 (see also Matthewson 2004:375). We show in section 5.2 how minimal pairs of data with acceptability judgments can be used to explore theoretical notions like syntactic well-formedness, felicity and truth.

In our survey of contemporary research on meaning in our field's four major theoretical journals, acceptability judgment tasks are very common, appearing in approximately three quarters of the papers. For example, in (10) above, native speakers of Hausa were asked to judge whether the given sentence is appropriate to say in the context provided. The diacritic used to mark the example (#) as well as the comment suggest that the example was judged to be unacceptable.[17]

[15] The instructions that precede the elicitation of 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 argue in section 3.

[16] We assume that an utterance is felicitous in a given context if and only if its felicity conditions are satisfied. 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., Kamp 1981, Heim 1982).

[17] An acceptability judgment is not always accompanied by a speaker's comment, but such a comment can be illuminating, as

In our survey, we find that that acceptability judgments are reported for linguistic expressions with and without a context. Another point of variation is the name by which the judgment task is identified (if it is identified at all): while some authors specify that speakers provided acceptability judgments, others say that speakers provided 'felicity judgments' or 'grammaticality judgments', even when speakers were asked to judge the acceptability of a linguistic expression: e.g., Matthewson 2004 writes about the elicitation of "felicity judgments" when acceptability judgments were elicited, and that utterances were judged to be "(in)felicitous" when they were actually judged to be (un)acceptable. Acceptability is also sometimes characterized as a judgment about whether an expression is '(in)coherent' or '(im)possible'. Furthermore, in our survey, it was sometimes difficult to determine what kind of judgment was elicited. For example, when a paper makes a claim about one sentence being able to report another sentence, this may (or may not) have been elicited with an acceptability judgment (about whether the first utterance is acceptable in a context in which it is used to report the second utterance). Finally, we note that in more than a quarter of the 40 papers in our survey, diacritics were presented with sentences without a comment about whether these diacritics represent an acceptability judgment, or some other judgment.

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.[18] 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 (5), repeated below:

(5) 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.)

discussed in Matthewson 2004. Furthermore, sometimes speakers' comments reveal their judgment, as discussed in Matthewson 2015. As far as we know, the practice of including relevant comments as part of the data originated with Matthewson 1999.

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

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 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.

At least one eighth of the papers in our survey appear to present implication (or: inference) judgments, where a speaker judges the implications of an expression.

Similarity judgment tasks A variant of the implication judgment task is the similarity judgment task, described in Degen 2015. This task requires speakers to judge the similarity of two sentences (e.g., a sentence with *some* and a sentence in which *some* is replaced by *some but not all*). Unlike in the direct implication judgment task, speakers are not asked to judge whether a particular utterance has a particular implication, but rather the similarity of the implications of two sentences.

Entailment judgment tasks An entailment judgment task is a sub-type 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 sub-type 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. The entailment judgment task occurred rarely in our survey but is mentioned in textbooks (e.g., Dowty et al. 1981:2). An example comes from Crnič (2014), who gives the examples in (19), and states that "that John read the book once is entailed by the proposition that John read the book twice" (p.176).

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

Paraphrase tasks A paraphrase task, like the entailment judgment task, is a sub-type of the implication judgment task. In fact, in its strictest interpretation, the paraphrase task is a type of entailment task. In the paraphrase 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. Paraphrase tasks are the second most common type of task (after the acceptability judgment task) found in our survey, with almost a quarter of the 40 papers relying on them. For example, Coppock and Beaver (2014) write about the examples in (20) that "a further fact to be explained is 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)

4.1.3 Truth value judgment and related tasks

The truth value judgment task was illustrated with the examples in (7) from Cable 2014 above (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.

In the literature on quantitative research, the truth value judgment task is one of the most widely used measures. In our survey, almost a fifth of the papers presented data which apparently relied on truth judgments. We write "apparently" because it was not always clear when an author reported that a particular utterance was true or false, whether this finding was based on a truth value judgment, or whether it was based on e.g., an acceptability judgment that supports the author's hypothesis about truth or falsity. A range of diacritics are used to mark pieces of data that were judged to be false, including the asterisk (*) and the hash mark (#). Authors sometimes appear to have collected truth value judgments when they describe utterances as 'odd', 'problematic' or 'inappropriate'. Likewise, statements about an uttered sentence being 'contradictory' may rely on truth value judgment tasks or may have been directly elicited as judgments about whether a sentence is contradictory (for the latter, see e.g., de Marneffe and Tonhauser 2015, and discussions in Matthewson 2004 and Bohnemeyer 2015). Sharvit (2014), for example, writes that "both answers in [(21a)] are felicitous, but the second answer in [(21b)] sounds contradictory (in particular, in contexts where John's leaving and Sally's arrival are completely independent of each other)" (p.281).

(21) Bill: When did John leave? Did Sally already arrive? (Sharvit 2014:281)

a. Fred: John left before Sally's arrival, which took place this morning / which is scheduled to take place tomorrow.

b. Fred: John left before Sally arrived, which she did this morning / #which she will tomorrow.

Finally, we also observed (in about a quarter of the papers in our survey) authors asserting generalizations about the kinds of context that a sentence is true in, without providing specific contexts in which the sentence was judged. Alrenga and Kennedy (2014), for example, point out that (22) is "true as long as it is possible to find a relevantly-sized group of students such that Sarah's height exceeds theirs" (p.31).

(22) Sarah is taller than some of my students are. (Alrenga and Kennedy 2014:31)

Ambiguity judgment task The ambiguity judgment task can be considered a sub-type 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. The ambiguity judgment task is mentioned in textbooks (e.g., Chierchia and McConnell-Ginet 2000:33-51, Larson and Segal 2005:9), but occurred very rarely in our survey. One example comes from Alrenga and Kennedy (2014), who state that the example in (23) "is...ambiguous" (p.4) and then describe the two readings:

(23) 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)

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).

In our survey, we found at least five cases where native speakers provide translations and these translations are then invoked as empirical evidence for the hypothesis about meaning. We also note reliance on translations in our own work. For example, Matthewson (2006) argues that "[s]uperficially tenseless sentences ... in St'át'imcets can be interpreted as either present or past" (p.676), giving as part of her support the St'át'imcets sentence in (24) with its English translation:

(24) táyt-kan
hungry-1.SG.SBJ
'I was hungry / I am hungry.' (adapted from Matthewson 2006:676)

And Tonhauser (2011) writes that, in Paraguayan Guaraní, "[i]n subordinate clauses, unmarked verbs are compatible with future time resference" (p.209), pointing to the Paraguayan Guaraní example in (25) with its translation for evidence.

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

4.1.5 Unclear task or no task

In the papers in our survey, many hypotheses about meaning were supported by the tasks discussed above. However, we also found many hypotheses that did not appear to be supported by a judgment of a piece of data, or by any other task: these hypotheses were merely asserted and typically accompanied by a de-contextualized linguistic expression. In other words, for these hypotheses, semanticists were reporting their intuitions (i.e., their impressions about meaning), without providing actual empirical evidence. This situation tends to occur 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 frequently find authors presenting de-contextualized sentences and simply asserting that they do or do not have some presupposition. We also found other hypotheses about meaning being merely asserted and accompanied by a de-contextualized expression, e.g., that some indefinite receives an existential rather than a universal interpretation, that the antecedent and consequent of some conditional have an epistemic rather than a causal connection, that some question has a negative bias, or that some piece of meaning is at-issue, just to name a few. We also found more than a dozen cases in which authors allude to a judgment, but it is not clear which judgment was elicited, e.g., to claim that a question can only be rhetorical, what the subject of a sentence was thinking, or that a pronoun has a bound de re interpretation.

4.1.6 Summary

A wide variety of types of tasks are used in contemporary research on meaning, including acceptability judgment, truth value judgment, implication judgment and translation tasks. The entailment judgment task and the paraphrase task were characterized as sub-types of implication judgment tasks, and the ambiguity judgment task as a sub-type of truth value judgment tasks. Which type of task is used is often not clearly identified. We also noted quite pervasive terminological inconsistencies in the naming of tasks. Specifically, researchers often refer to 'grammaticality judgment' and 'felicity judgment' tasks even when what was elicited were acceptability judgments.

4.2 Evaluation of tasks

In this section, we evaluate the response tasks we just characterized with respect to whether they lead to stable, replicable and transparent pieces of data.[19] We argue that the tasks are qualitatively different with respect to this objective:

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

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

Figure 1: Evaluation of response tasks in research on meaning

Translation task We start our evaluation with the translation task. 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 in Matthewson 2004, Bohnemeyer 2015 and Deal 2015, among others, since translations need not preserve truth or felicity conditions. One language could lack easy means to express truth conditions which are easily expressible in the other (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). 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 translations do not provide reliable evidence in research on meaning.

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

This includes that 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 pragmatic 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 about meaning.

Paraphrase, entailment judgment and ambiguity judgment tasks Tasks that can only be performed reliably by speakers with linguistic training fare poorly with respect to replicability. That the paraphrase task, the entailment judgment task and the ambiguity judgment task require linguistic training is already evident from their characterizations in the previous section: 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). 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 also Matthewson 2004 and Sprouse et al. 2013:§2.2 for the argument that tasks that require linguistically trained speakers are not ideal.)

Because these tasks require responding speakers to have linguistic training, these tasks also lead to pieces of data that are less than transparent. Once a task requires a speaker to perform linguistic analysis (e.g., to determine an entailment relation), the relation between the piece of data and the hypothesis about meaning it is supposed to support becomes less transparent. To illustrate, assume that a native speaker researcher judges that the expression in (26b) is a paraphrase of the expression in (26a) in the context provided:

(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.

Assume further that these two pieces of data are taken as evidence that *only* (at least on one of its meanings) is equivalent to the meaning of *just*. The linking hypothesis that is required to make this connection between the response and hypothesis about meaning is the following:[20]

(27) **Linking hypothesis for paraphrase judgment task**
Two expressions A and B can convey the same meaning if and only if two complex expressions that differ only in whether they realize A or B are judged to be paraphrases of one another.

[20]Recall that a linking hypothesis is an explicit statement of the link between some piece(s) of data and a hypothesis about meaning. 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, the empirical evidence for two expressions A and B being equivalent in meaning has only been provided insofar as there is empirical evidence that two complex expressions containing them are equivalent in meaning. But no such evidence has been provided in (26). What has been provided is evidence that (26a) and (26b) are both acceptable in the same context. This is not yet evidence that (26a) and (26b) convey the same meaning. The claim that they convey the same meaning is instead based on a speaker's intuition that the two have equivalent truth conditions. As such, pieces of data that involve the paraphrase task are not transparent.

A piece of data with an ambiguity judgment task is also not transparent. As characterized above, this task requires a speaker to develop hypotheses about potentially distinct meanings (sets of truth conditions) for the expression and to explore whether there are contexts that make one meaning true and another false. If the result of these considerations is simply reported as a statement that the expression is ambiguous, the piece of data fails the transparency criterion since the reader is not privy to the steps which underlie the judgment. To become transparent, the reader would need to be presented with contexts that make one of the meanings true and the other false, and judgments of the (un)acceptability of the expression in these contexts – in other words, a series of acceptability judgment tasks (ee e.g., Crain and McKee 1985:104 for discussion).

For the entailment and ambiguity judgment tasks, another reason they occur towards the lower end of the evaluation scale in Figure 1 is that these tasks are performed, by definition, on de-contextualized expressions. That is, a sentence entails another one if and only if the latter is true in all of the contexts in which the former is true, and the two sets of truth conditions for an ambiguous expression are determined without regard to context. Since pieces of data involving these tasks do not involve a context, they are less stable.

As with the translation task, we qualify our assessment of the paraphrase, entailment judgment and ambiguity judgment tasks by pointing out that it can be useful to ask speakers (whether theoretically trained or not) whether two expressions "mean the same thing" or "mean something different from each other" (for a recent application, see e.g., Matthewson 2015:slide 27). A negative response to the first question (or a positive response to the second) can provide an important clue that the speaker perceives the two expressions to differ in meaning. This clue can then serve to develop a more refined hypothesis about how the two expressions differ in meaning, which can be explored using e.g., acceptability or implication judgment tasks. Crucially, however, we argue that a speaker's assessment that two expressions differ in meaning does not yet constitute evidence about whether they differ in their truth conditions, in their felicity conditions, or in some other way. Likewise, a positive response to the question of whether two expressions mean the same thing at best constitutes a clue about meaning, but does not warrant the assumption that the two expressions have the same truth and felicity conditions. The burden of analysis should be on the researcher, not on the speaker providing the judgments. Empirical evidence that two expressions do not have the same meaning can be provided by showing, for instance, that there is a context in which the two expressions receive distinct acceptability judgments.

Truth value judgment tasks Truth is a theoretical concept. As such, the truth value judgment task cannot be reliably applied with native speakers who have not learned to distinguish the truth conditions of an utterance from other conditions on its felicitous and pragmatically unmarked use. Under this consideration, pieces of data that involve a truth value judgment task are less replicable than e.g., pieces of data that involve

an acceptability judgment task; cf. our evaluation in Figure 1.

We note, however, that this evaluation is limited to the truth value judgment task under the 'classical' linking hypothesis in (28a). Here, the 'no' response (to a question like 'Is this sentence true?') necessitates the assumption that the speaker has performed analysis and has determined that the sentence is indeed false, rather than infelicitous or pragmatically anomalous in some other way. (We assume that a speaker responds 'yes' only if the sentence is syntactically well-formed, true and felicitous, as in the acceptability judgment task.) Under the 'cautious' and 'very cautious' linking hypotheses, however, the truth value judgment task is applicable with theoretically untrained native speakers (including children; see Crain and McKee 1985 and Crain and Thornton 1998), since a 'no' response here does not require assuming that the speaker has performed this kind of analysis. Rather, under the 'cautious' linking hypothesis, the researcher first ascertains that the utterance to be judged is syntactically well-formed, felicitous and pragmatically unmarked, thereby excluding any other reasons for why a theoretically untrained speaker might reject the utterance. And, under the 'very cautious' linking hypothesis, a 'no' response is not taken as evidence for a particular theoretical notion. Thus, under the 'cautious' and 'very cautious' linking hypotheses, the truth value judgment task leads to replicable pieces of data.

(28) **Possible linking hypotheses for the truth value judgment task**

a. **Classical:** An expression that is judged to be true in a context is true. An expression that is judged to be false in a context is false.

b. **Cautious:** A syntactically well-formed, felicitous and pragmatically unmarked expression that is judged to be true in a context is true. A syntactically well-formed, felicitous and pragmatically unmarked expression that is judged to be false in a context is false.

c. **Very cautious:** A syntactically well-formed, felicitous and pragmatically unmarked expression that is judged to be true in a context is true. A syntactically well-formed expression that is judged to be false in a context is false, infelicitous or pragmatically anomalous in some way.[21]

In our research with theoretically untrained speakers of languages we do not speak natively, we have applied the truth value judgment task only under the 'cautious' and 'very cautious' linking hypotheses. That is, we had either confirmed independently that the expression to be judged was syntactically well-formed, felicitous and pragmatically unmarked, or we did not take a 'no/false' response as evidence that the expression was false. In research in which a 'no/false' response is taken as evidence that the expression is false, it is often apparent that the 'cautious' linking hypothesis was applied. Cable's 2014 example (7bii), for instance, is shown to be syntactically well-formed, and felicitous and pragmatically unmarked (in another context, namely (7aii), where it is judged to be true), and so the 'cautious' linking hypothesis was applied. 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.

The reason we have not placed the truth value judgment task to the far right in Figure 1, despite it leading to replicable pieces of data under at least two linking hypotheses, is because the task also appears to be ap-

[21] Under the 'very cautious' linking hypothesis, the truth value judgment task becomes very similar to an acceptability judgment task. See section 5 for further discussion.

plied under the 'classical' linking hypothesis. In Bott and Noveck 2004, for example, speakers were asked 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*). The truth value judgment task under the 'classical' linking hypothesis is widespread in quantitative research on scalar implicatures. In fact, however, speakers may just be unwilling to judge such pragmatically underinformative utterances as 'true' (cf. the 'cautious' and 'very cautious' linking hypotheses). Evidence that the truth value judgment task under the 'classical' linking hypothesis is not reliable with theoretically untrained speakers comes from Soames (1976:169), von Fintel (2004) and Abrusán and Szendrői (2011), who argue that adult speakers may judge utterances that are infelicitous (e.g., because the felicity condition of a definite noun phrase is not satisfied) to be false, even though they are assumed not to have a truth value.

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 meant 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: the pieces of data can be replicated in languages without theoretically trained speakers as well as through quantitative methods. By virtue of being collectible with linguistically untrained speakers, pieces of data collected through acceptability, implication or similarity judgment tasks can also satisfy the transparency objective, since they do not require the speakers to conduct analysis in giving their responses. Maximal transparency is achieved with these tasks by means of a clear linking hypothesis (see section 5 for further discussion of linking hypotheses). 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 stable.

4.3 Summary

Research on meaning employs a wide variety of response tasks. In this section, we characterized the ones most frequently used in contemporary research on meaning, and pointed out that they vary in the extent to which they lead to stable, replicable and transparent data. In particular, we argued that acceptability, implication and similarity judgment tasks are preferable to other tasks, including translation, paraphrase, entailment judgment and ambiguity judgment tasks, and truth value judgment tasks under a particular linking hypothesis.

We also noted that response tasks are not consistently identified in contemporary research on meaning. As we argued in section 3, including information about the response task is vital and can be easily accomplished through a short methods section. We further observed that response tasks are sometimes referred to

by names that do not reflect the question that was posed to the speaker. We encourage the convention of referring to a task that asks a native speaker to respond to a question about X as an 'X judgment task'.

5 Minimal pairs and linking hypotheses in research on meaning

In this section, we argue that empirical evidence in research on meaning is constituted by pieces of data, possibly in minimal pair form, plus linking hypotheses. A single, positive piece of data together with a linking hypothesis is one of the simplest forms that empirical evidence can take.[22] Take, for instance, the linking hypothesis that an expression that was judged to be acceptable in a context is syntactically well-formed, true and felicitous in that context. Under this linking hypothesis, the finding that (29) is judged to be acceptable by the second author of this paper provides empirical evidence that the expression of the piece of data is syntactically well-formed, and true and felicitous in the context.

(29) Context: Maggie calls her mother from Spain, and says:
Yesterday I visited the Alhambra in Granada.

A single, negative piece of data can also constitute empirical evidence. Consider the piece of data in (30), which was judged to be unacceptable by the second author of this paper.

(30) Context: Maggie calls her mother from Spain, and says:
#Yesterday I visit the Alhambra in Granada.

Under the linking hypothesis that an expression that is judged to be unacceptable is syntactically ill-formed, false, infelicitous, or a combination thereof, we can merely conclude that (30) is syntactically ill-formed, false, infelicitous or a combination thereof. To further pinpoint the reason behind an unacceptability judgment like for (30) requires linking hypotheses that make explicit reference to theoretical assumptions, as we discuss below. And, in fact, theories differ in whether the temporal mismatch between *yesterday* and the non-past tense verb *visit* is treated as an agreement mismatch (in which case (30) is syntactically ill-formed), as conflicting entailments (in which case (30) is false), or as conflicting constraints on temporal reference (in which case (30) is infelicitous).

Our survey of research articles showed that linking hypotheses are not consistently stated in research on meaning. In part, this may be due to an assumption that linking hypotheses like the ones used in connection with the examples in (29) and (30) are widespread enough that they need not be explicitly stated. However, given that the meaning of an utterance is only indirectly revealed through a speaker's response to the task for that utterance, it is important for linking hypotheses to be made explicit in research on meaning. At the very least, we argue, they should be stated once for each type of empirical evidence provided in a paper (e.g. together with the first example for each type of evidence).

In research on meaning, single, positive pieces of data provide evidence for a limited type of empirical generalization, namely empirical generalizations about which meanings complex expressions are compatible

[22] A positive piece of data is one where speakers gave a positive response to the task about the linguistic expression in the context in which the expression was presented, e.g. a judgment that the expression is judged to be acceptable or true. A negative piece of data is one where speakers gave a negative response, e.g. a judgment that the expression is judged to be unacceptable or false.

with, e.g., (29), or incompatible with, e.g., (30). However, more often than not, research on meaning is concerned with identifying which specific part of a complex expression contributes a particular meaning, or which specific feature of context some part of a complex expression is sensitive to. To provide evidence for such empirical generalizations, research on meaning relies on pieces of data in minimal pair form. In the next section, we define four types of minimal pairs and show how they provide empirical evidence for different types of generalizations about meaning. We then show in section 5.2 how minimal pairs of pieces of data with acceptability judgments can provide empirical evidence for hypotheses about syntactic well-formedness, truth and felicity.

5.1 Types of minimal pairs in research on meaning

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 linking hypothesis 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 is more than one type of minimal pair in research on meaning. Specifically, a minimal pair in research on meaning consists of two pieces of data that differ minimally in either the linguistic expression, as in (31a-i) and (31b-i), or in the context in which the expression is uttered, as in (31a-ii) and (31b-ii). The two pieces of data in a minimal pair may differ in the responses, as in (31a), or not, as in (31b).

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

a. Minimal pairs with two pieces of data that receive distinct responses

   i. Linguistic variants: The two pieces of data have the same context but minimally different linguistic expressions that result in different responses by native speakers.

   ii. Context variants: The two pieces of data have the same linguistic expression but minimally different contexts. The linguistic expression receives different responses by native speakers.

b. Minimal pairs with two pieces of data that both receive a positive response

   i. Linguistic variants: The two pieces of data have the same context but minimally different linguistic expressions that result in the same response by native speakers.

   ii. Context variants: The two pieces of data have the same linguistic expression but minimally different contexts. The linguistic expression receives the same response by native speakers.

These minimal pairs can be used for the all-important goal of identifying the specific part of a complex linguistic expression that contributes to the meaning of the complex expression, and of identifying the specific feature of the complex context that is responsible for the interpretation of the expression in the complex context.[23] Two pieces of data that are both negative (e.g., are both judged to be unacceptable or false) do not form a minimal pair that can be used to pursue this goal. This is because two pieces of negative data merely reveal that two expressions are judged to be unacceptable or false in their contexts, but they do not pinpoint which part of the expression or which feature of the context is responsible for the judgment. Hence, minimal pairs with two pieces of negative data are not included in (31).

We show in the next two sections (5.1.1 and 5.1.2) that the four types of minimal pairs in (31) provide evidence for different types of hypotheses about meaning. In these sections, we limit our attention to minimal pairs of piece of data with binary acceptability judgments. The linking hypothesis we assume throughout these sections is the following:

(32) **Linking hypothesis for binary acceptability judgments**
An expression that is judged to be acceptable in a context is syntactically well-formed, true and felicitous in that context. An expression that is judged to be unacceptable in a context is syntactically ill-formed, false or infelicitous, or a combination thereof.

As we discuss in detail in section 5.2, a judgment of unacceptability does not, under this linking hypothesis, support a conclusion about why the expression was judged to be unacceptable in the context, e.g. whether the expression was syntactically ill-formed, false or infelicitous; acceptability judgments can nevertheless be used to explore these three theoretical notions.

We briefly address other types of minimal pairs and other linking hypotheses in section 5.1.3.

5.1.1 Minimal pairs of pieces of data with distinct acceptability judgments

In this section, we discuss hypotheses that can be supported with minimal pairs of the form in (31a), i.e., where the two pieces of data receive distinct acceptability judgments.

Linguistic variants A minimal pair of type (31a-i), 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 what differs between the two linguistic expressions contributes or does not contribute a particular meaning. Consider the two pieces of data in (33). These share the same context, and the Paraguayan Guaraní linguistic expression in (33a) differs from the one in (33b) in the presence of the exclusive clitic *=nte* 'only' on the name *Javier*. The linguistic expression in (33a) was judged to be unacceptable in the context provided by four native speakers of Paraguayan Guaraní, whereas the linguistic expression in (33b) was judged to be acceptable by the same four native speakers. Thus, the two pieces of data in (33) form a minimal pair of type (31a-i).

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

(33) 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 the English adverb *only*. Under this hypothesis, the sentence in (33a) would mean that Javier has a cow and nobody other than Javier has a cow. If the hypothesis is correct, we expect (33a) to be judged to be unacceptable in the context in (33): The context specifies that both Javier and Maria own a cow and, hence, (33a) is expected to be false in this context under the given hypothesis. Since that is indeed what we find, the minimal pair in (33), under the linking hypothesis in (32), provides empirical evidence for the hypothesis about *=nte*.

It is important to realize that the fact that (33a) is judged to be unacceptable in the context in (33) only provides empirical support for the hypothesis that the entire sentence contributes an exclusive meaning, not that it is *=nte* 'only' that contributes this meaning. Furthermore, (33a) may be judged to be unacceptable for reasons other than the meaning that is hypothesized to be contributed by *=nte* 'only'. Therefore, it is the combination of the unacceptable example in (33a) with the acceptable minimal variant in (33b) that provides empirical support for the hypothesis that *=nte* 'only' contributes the exclusive meaning: (33b), which does not feature *=nte* 'only' and hence by hypothesis does not convey an exclusive meaning, is judged to be acceptable in the context.[24]

The linguistic expressions in minimal pairs of type (31a-i) may also differ in the order of parts of the expressions. In the minimal pair in (34), 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 (34a), or outside it, as in (34b).[25]

(34) 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.'

Let's assume that we are exploring the hypothesis that the interpretation of sentences with *–mo'ã* 'CF' depends on whether *–mo'ã* 'CF' occurs inside the negation circumfix, as in (34a), or outside of it, as in (34b).

[24] Of course, this argument relies on the researcher hypothesizing that the unacceptability of (33a) is not due to it being syntactically ill-formed or infelicitous in the given context. Support for the hypothesis that (33a) is syntactically well-formed is provided by the finding that (33a) is judged to be acceptable in a context in which Javier has a cow and nobody else has one, as in (36).

[25] We emphasize that the gloss and translation of examples like (34b) are not part of the piece of data but merely help the reader understand the important features of the (Paraguayan Guaraní, in this case) example. Crucially, (34b) does not provide empirical evidence that the Paraguayan Guaraní expression conveys the meaning of its English translation.

The minimal pair in (34) provides support for this hypothesis since (34a) but not (34b) is judged to be acceptable in the context in (34) by four speakers of Paraguayan Guaraní. Both pieces of data in (34) are required to support the hypothesis that the position of *–mo'ã* 'CF' with respect to negation influences its interpretation. Again, this argument relies on the researcher hypothesizing that (33a) wasn't judged to be unacceptable e.g., because it is syntactically ill-formed; see Tonhauser 2009 for discussion of *–mo'ã* 'CF'.

In minimal pairs of type (31a-i), the linguistic expressions may also differ minimally in their constitutive parts. In the minimal pair in (35), 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. Let's also assume that we are pursuing the hypothesis that *lákw7a* 'SNV' is an evidential that is incompatible with inferential evidence. The context in (35) is designed such that inferential evidence obtains. We therefore expect (35a) to be judged to be acceptable and (35b) to be judged to be unacceptable, and this is indeed what we observe:

(35) 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)

Again, both pieces of evidence are needed: the minimal variation in form between the sentences in (35a) and (35b) provides evidence that it is the *lákw7a* 'SNV' morpheme that results in (35b) being unacceptable in the context provided, i.e., that this morpheme has a meaning incompatible with inferential evidentiality.

In sum, minimal pairs of type (31a-i) can provide evidence that what differs between the two expressions of the minimal pair contributes a particular meaning, as in (33), or results in a change in meaning, as in (34) and (35). As discussed, both the positive and the negative pieces of data are necessary to provide empirical evidence for the hypotheses.

Context variants A minimal pair of type (31a-ii), 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.

To establish two pieces of data with minimally different contexts, we identify a single feature of context that matters for the purpose of our investigation and make two contexts that differ only with respect to that feature. Recall the hypothesis from example (33) that the Paraguayan Guaraní clitic *=nte* 'only' conveys an exclusive meaning. A feature of context that we hypothesized there to matter for the interpretation of *=nte* 'only' was whether Javier was the only person to have a cow, or whether somebody else also has a cow. The two contexts of the minimal pair in (36) differ in this regard, and the linguistic expression from example (33a) is realized in both (36a) and (36b). Given the aforementioned hypothesis that *=nte* 'only' contributes

an exclusive meaning, we expect (36a) to be judged to be unacceptable and (36b) to be judged acceptable, and this is indeed what we find.

(36) 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.'

Thus, this minimal pair provides empirical evidence in support of a hypothesis about the meaning of a linguistic expression realized in the two pieces of data in the minimal pair.

The minimal pair in (37) was established to explore the hypothesis that unrealized subject arguments require a salient discourse referent. The contexts in the minimal pair thus differ with respect to this feature: in (37a), the context establishes a contextually salient entity other than the interlocutors, namely a dog, whereas the context in (37b) does not establish such an entity. The linguistic expression in (37) consists of the temporal adverb *kuehe* 'yesterday' and a transitive verb stem, *-su'u* 'bite', that is marked for a first person singular direct object. The subject is not realized overtly in this sentence. Given the hypothesis, we expect the sentence to be judged to be acceptable when uttered in the context in (37a) but unacceptable when uttered in the context in (37b), and that is indeed what we find:

(37) (Tonhauser under review)

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
(Intended: Yesterday, something bit me.)

Thus, the minimal pair provides empirical evidence for the hypothesis that the implicit subject argument must be interpreted as a contextually salient entity: (37a) where such an entity is available is acceptable, whereas (37b) where such an entity is not available is judged to be unacceptable. Again, both members of the minimal pair are needed: (37a) by itself only provides evidence that the linguistic expression is acceptable when the context makes available such an entity but it does not show that such an entity is required; (37b) provides that evidence.

The contexts in the minimal pair in (38) differ minimally in whether a question under discussion is established: the context in (38a) does not establish one and the one in (38b) does. The motivation behind this minimal pair is the hypothesis, already mentioned in connection with the Gitksan example in (15) (repeated in (38a)), that *=ist* 'QQUD' signals that the utterance addresses the question under discussion. In the context of (38a), the issue of Charlie's sickness is not in the question under discussion at the time of utterance, so *=ist* 'QUDD' is predicted to be unacceptable. In the context of (38b), the speaker is answering the current question under discussion, so *=ist* 'QUDD' is expected to be judged to be acceptable.

(38) a. Context: Adam and Betty are eating dinner quietly. Nobody has said anything yet. Betty suddenly says:

#siipxw-t Charlie/Tsaalii=is(t)
sick-3SG.II Charlie=QUDD
'Charlie is sick.' (Matthewson 2015:(7), slide 28)

b. Context: Adam and Betty are eating dinner. Betty mentions that a lot of people at her work are sick at the moment, and Adam asks her 'Who is sick?'. Betty replies:

siipxw=t Charlie=ist
sick=DM Charlie=QUDD
'Charlie is sick.'

Both pieces of data are needed to support the hypothesis. Example (38a) alone would not reveal that the unacceptability is caused by the context of the example; it is the contrasting acceptability of (38b) which suggests that answering a question is the (or at least, one possible) mitigating factor. Conversely, the acceptability of (38b) would not provide conclusive evidence that a preceding question is required for *=ist* 'QUDD'; it would only show that a sentence with *=ist* 'QUDD' may occur in response to such a question.

In sum, minimal pairs of type (31a-ii) 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.

5.1.2 Minimal pairs with pieces of data that are both judged to be acceptable

In this section, we discuss hypotheses that can be supported with minimal pairs of the form in (31b), i.e., where the two pieces of data are both judged to be acceptable.

Linguistic variants A minimal pair of type (31b-i), in which both pieces of data have the same context but minimally different linguistic expressions that are both judged to be acceptable in the context, provides evidence that the linguistic expressions of the two members of the pair are both compatible with the contextually specified meaning. Recall the minimal pair in (26), repeated below for convenience:

(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.

The context of the two pieces of data establishes that, despite appearances, the fire alarm was sounded not 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 shows that both expressions are compatible with the so-called 'rank order' interpretation of exclusives (Coppock and Beaver 2014).

Of course, there is always the possibility that the context of a piece of data is not sufficiently controlled, and therefore does not actually force the desired interpretation. Minimal pairs like (26) can be strengthened by a negative piece of data that shows that the context of the piece of data is sufficiently controlled. Consider the example in (26c), a minimal variant of the linguistic expressions in (26), with the exclusive *alone*. Let's assume that we hypothesize that *alone* is not compatible with a rank order interpretation. We therefore expect (26c) to be judged to be unacceptable in the context of (26), and this is indeed what we find:

(26c) #This was a drill **alone**.

In sum, minimal pairs of type (31b-i) can provide evidence that the linguistic expressions of the two members of the pair are both compatible with the contextually specified meaning. For reasons of space we have only provided an example here in which the two linguistic expressions differ in whether one morpheme or another is realized, but of course the linguistic variants illustrated in section 5.1.1 are also possible.

Context variants A minimal pair of type (31b-ii), in which the two pieces of data have minimally different contexts but the same linguistic expression that is judged to be acceptable in both contexts, provides evidence that the linguistic expression is compatible with the two meanings controlled for by the two contexts.

One reason this type of minimal pair is useful in research on meaning is because it can be used to provide empirical evidence that a particular expression does *not* make a particular contribution to meaning. Let's assume, for instance, that we are exploring the hypothesis that non-plural-marked nouns in Paraguayan Guaraní are compatible with both singular and plural denotations, i.e., do not contribute a singular meaning. To show this, we need to provide empirical evidence that a noun not marked with the plural marker *–kuéra* is compatible with both singular and plural denotations. In the minimal pair in (39), the same linguistic expression with the unmarked noun *vaka* 'cow' is realized in a context that establishes a singular denotation for the noun in (39a) and in a context that establishes a plural denotation for the noun in (39b).

(39) a. Context: Maria owns a cow. She says:

A-guereko vaka.
A1sg-have cow

'I have a cow.'

b. Context: Maria owns two cows. She says:

A-guereko vaka.
A1sg-have cow

'I have cows.'

The fact that the linguistic expression is judged to be acceptable in both contexts provides empirical evidence that *vaka* 'cow' is compatible with a singular and a plural denotation.

A second example, from Gitksan, is given in (40). Here the hypothesis is that bare verb forms (like *ha'wits'am* 'crush') can denote habitual states in the actual world and also habitual states only found in possible, non-actual worlds. The hypothesis is supported by the minimal pair in (40): the sentence is acceptable in the context of (40a), which describes a situation in which the machine regularly crushes oranges, and also in that of (40b), which describes a situation in which the machine has not yet crushed an orange.

(40) 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.'

In sum, minimal pairs of type (31b-ii) can provide evidence that the meaning of the linguistic expression realized in the two contexts is compatible with both meanings controlled for by the contexts. Minimal pairs of this type thus can provide empirical evidence that a meaning distinction encoded in some language (e.g., number, tense, definiteness) is not encoded in another language.

5.1.3 Summary

This section has illustrated that there are four types of minimal pairs in research on meaning, and that each type of minimal pair is used to provide empirical support for different types of hypotheses about meaning. We illustrated the four types of minimal pair here for pieces of data that received binary acceptability judgments, under the linking hypothesis for binary acceptability judgments in (32).

That the majority of the examples we used in this section come from languages under-represented in research on meaning, namely St'át'imcets, Gitksan and Paraguayan Guaraní, is only due to the fact that we primarily conduct research on these languages. We want to emphasize here that we intend our proposal for what constitutes empirical evidence in research on meaning to apply to languages with and without native speaker semanticists alike. That is, the four types of minimal pairs provide evidence for different types of empirical generalizations under the linking hypothesis in (32) regardless of whether the language under investigation is a well-studied European one with native speaker semanticists or a less-studied language without native speaker semanticists.

Similarly, these four types of minimal pairs are not particular to pieces of data that received binary acceptability judgments. The same types of minimal pairs can also be formed with forced choice truth value judgments, as illustrated with (7), or with implication judgments: minimal pair type (31a-i) 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 (41), and asked to judge the acceptability of the answer on a 5-point Likert scale. The minimal pairs in this task consist of dialogues with the same question and answers that differ in whether the presupposition of the question (that Victoria was the director in the past) is denied, as in (41a), or not, as in (41b).

(41) (Amaral and Cummins 2015:165)

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

B: 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?'

B: 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 (41b) received significantly higher acceptability ratings than dialogues like the one in (41a). Under a (presumed) linking hypothesis that one answer is preferred over another if the acceptability judgments of the answers differ significantly, this finding supports the hypothesis that answers that do not deny a presupposition are preferred over answers that deny a presupposition.

In sum, minimal pairs are a powerful tool in research on meaning. As summarized in (42), different types of minimal pairs provide empirical evidence for different types of hypotheses about meaning: about the contributions to meaning of particular parts of a complex expression, about the contributions to meaning of particular features of context, and about which meanings an expression is compatible with.

(42) **Supporting hypotheses about meaning with minimal pairs**

a. Pairs of pieces of data that receive distinct responses

i. The two pieces of data have minimally different linguistic expressions but the same context. Such minimal pairs are used to provide evidence that the linguistic material that distinguishes the two expressions does (not) contribute a particular meaning.

ii. The two pieces of data have minimally different contexts but the same linguistic expression. Such minimal pairs are used to provide evidence that the meaning of the expression is sensitive to what differs between the two contexts.

b. Pairs of pieces of data that receive the same (positive) response

i. The two pieces of data have minimally different linguistic expressions but the same context. Such minimal pairs are used to provide evidence that both expressions are compatible with a particular meaning.

ii. The two pieces of data have minimally different contexts but the same linguistic expression. Such minimal pairs are used to provide evidence that the linguistic expression is compatible with two meanings.

5.2 The role of minimal pairs with linking hypotheses in providing empirical evidence

Given that minimal pairs are necessary to conclusively establish the contributions to meaning of parts of complex expressions, to establish which feature of context contributes to interpretation, and to identify which expressions are compatible with which meanings, one might expect them to be pervasive in research on meaning. However, although about three quarters of the papers in our survey made use of minimal pairs of pieces of data at least some of the time, only a tiny fraction of the 40 papers — namely Cable 2014 and Rojas-Esponda 2014 — consistently made use of them.

The minimal pair methodology is a powerful tool in research on meaning because minimal pairs with suitable linking hypotheses can provide empirical evidence for highly specific hypotheses about meaning concerning very subtle phenomena, including epistemic indefinites (Alonso-Ovalle and Menéndez-Benito 2013), temporal remoteness (Cable 2013), reciprocity and reflexivity (Cable 2014), discourse structure (Rojas-Esponda 2014), projective content (Tonhauser et al. 2013), and definiteness (Ionin 2006). In fact, we argue that the minimal pair methodology is so strong that minimal pairs with a suitable linking hypothesis can provide evidence for every empirical phenomenon in research on meaning. In other words, if minimal pairs cannot provide empirical evidence for some hypothesis about meaning, then perhaps that hypothesis is not tenable.

Our goal in this section is to illustrate the minimal pair methodology and the power of minimal pairs by showing how minimal pairs with suitable linking hypotheses can provide empirical evidence for hypotheses about syntactic well-formedness, truth and felicity. Our motivation behind choosing these three theoretical notions is that researchers often either seem to merely rely on their intuitions in motivating them, or rely on (what are dubbed) 'grammaticality judgments', truth value judgments, and (what are dubbed) 'felicity judgments'. Our goal is to show that minimal pairs of pieces of data with acceptability judgments and suitable linking hypotheses can be used to explore these theoretical notions, thus obviating the need for the aforementioned judgment types. (Along the same lines, see Bohnemeyer 2015:34f. for a discussion of how acceptability judgments can be used to explore entailment.)

To get started, recall from the linking hypothesis in (32) that a judgment of acceptability can be taken to mean that the linguistic expression that was judged is syntactically well-formed, true and felicitous in that context, as represented in line I of Table 1. A judgment of unacceptability, on the other hand, does not provide evidence for whether the linguistic expression of the piece of data is syntactically ill-formed, false or infelicitous in that context, or a combination of the three, as represented in lines II, III and IV of Table 1.

As mentioned above, we maintain that it is the researcher's task to identify the reason behind the unacceptability judgment, not that of the speakers delivering the judgment (except when the researcher delivers the judgments, of course). What is crucial to realize is that one can only explore one of the three theoretical concepts (syntactic well-formedness, truth, felicity) at any given time. That is, a judgment of unacceptability can be used to explore a hypothesis about syntactic well-formedness if the expression is hypothesized to be felicitous and true in the context provided (line II.a), to explore a hypothesis about felicity if the expression is hypothesized to be syntactically well-formed and true in the context provided (line II.b), and to explore a hypothesis about truth if the expression is hypothesized to be syntactically well-formed and felicitous in the

	Sentence S uttered in context C is hypothesized to be			
	syntactically well-formed	felicitous	true	Acceptability judgment
I.	✓	✓	✓	acceptable
II.a	no	✓	✓	unacceptable
II.b	✓	no	✓	unacceptable
II.c	✓	✓	no	unacceptable
III.a	✓	no	no	unacceptable
III.b	no	✓	no	unacceptable
III.c	no	no	✓	unacceptable
IV.	no	no	no	unacceptable

Table 1: Acceptability judgments in relation to syntactic well-formedness, felicity and truth

context provided (line II.c).[26] Other types of pieces of data that are judged to be unacceptable, represented in lines III and IV, are not useful to explore hypotheses about syntactic well-formedness, felicity and truth.

Of course, it is usually not possible to establish syntactic well-formedness, felicity and truth one after the other, in neatly separated packages. Rather, we often collect pieces of data based on some initial hypotheses, revise the hypotheses based on the data collected, collect more data, and repeat. The following three sections illustrate how minimal pairs of pieces of data with acceptability judgments can be used to provide empirical evidence for hypotheses about syntactic well-formedness, felicity and truth.

5.2.1 Empirical evidence for hypotheses about syntactic well-formedness

This section illustrates how minimal pairs of pieces of data with acceptability judgments can support a hypothesis about syntactic well-formedness. Our linking hypothesis here is one that amends the second part of the linking hypothesis in (32):

(43) **Linking hypothesis to explore syntactic well-formedness with binary acceptability judgments**
An expression that is judged to be acceptable in a context is syntactically well-formed, true and felicitous in that context. An expression that is judged to be unacceptable in a context in which it is true and felicitous is syntactically ill-formed.

Our first example is concerned with something that most theoreticians would agree is a syntactic phenomenon, namely the nominative case of a subject noun phrase in German. Let's say that (44a) was judged to be unacceptable and we want to argue that the judgment is due to the sentence being syntactically ill-formed (and hence we have marked (44a) with an asterisk, which we use to mark examples that we hypothesize to be judged to be unacceptable for syntactic reasons).

[26] As discussed in section 2, we hypothesize that it is because acceptability judgments can be used to probe syntactic well-formedness, felicity and truth that they are sometimes referred to as 'grammaticality judgments', 'felicity judgments' and 'truth judgments'. We have argued above against this use of the terminology.

(44) Context: We've been trying to reach a particular electrician for a while. Today, he finally called. I tell you:

a. *Des Elektrikers hat heute angerufen.
the.GEN.M electrician.GEN.M has today called
(Intended: The electrician called today.)

b. Der Elektriker hat heute angerufen.
the.NOM.M electrician.NOM.M has today called
'The electrician called today.'

(44a), together with the acceptable variant in (44b), is a minimal pair of type (31a-i): the two linguistic expressions are judged in the same context and receive distinct acceptability judgments (from the first author). The context is designed so that the intended/actual meaning of the two sentences is true in the context (the electrician called) and so that the intended/actual felicity conditions of the two sentences are satisfied (e.g., those of the definite article). Since we hypothesize that the intended/actual meaning of the two sentences is true and felicitous in the context, and since the sentence in (44b) is judged to be acceptable, we conclude that the unacceptability of (44a) is due to the genitive case marking of the definite noun phrase. If we further assume that the case of subjects in German is a syntactic phenomenon, we can now argue that (44a) is judged to be unacceptable because it is syntactically ill-formed.

The example in (44) also serves to make the point that syntactic well-formedness is a theoretical notion: if we thought that nominative case assignment was governed by semantics, for example, we would want to argue that (44a) is judged to be unacceptable for semantic reasons, and the argument made with the minimal pair in (44) would go through just as well.

The Gitksan example in (45) makes this point for a phenomenon that is less well-understood than the German case: (45a), which contains the prospective marker *dim* 'PROSP', is judged to be acceptable in the context provided; the minimal variant in (45b) without *dim* 'PROSP' is judged to be unacceptable.

(45) Context: I have been unable to work lately due to illness. Yesterday, I felt better, but I still didn't work.

a. da'a<u>k</u>hlxw-i-'y **dim** hahla'alsd-i'y ky'oots, ii ap nee=dii wil-i'y
CIRC.POSS-TR-1SG.II PROSP work-1SG.II yesterday CL.CNJ ¬PPS NEG=FOC do-1SG.II
'I was able to work yesterday, but I didn't.'

b. #da'a<u>k</u>hlxw-i-'y hahla'alsd-i'y ky'oots, ii ap nee=dii wil-i'y
CIRC.POSS-TR-1SG.II work-1SG.II yesterday CL.CNJ ¬PPS NEG=FOC do-1SG.II
(Intended: I was able to work yesterday, but I didn't.) (Matthewson 2013:372)

Given that (45a) is syntactically well-formed, true and felicitous in the context in (45), one could hypothesize that (45b) is judged to be unacceptable because it is syntactically ill-formed (perhaps there is a syntactic rule that the prospective marker *dim* 'PROSP' must be c-commanded by a circumstantial modal). But one could also hypothesize that (45b) is judged to be unacceptable because it is infelicitous (perhaps circumstantial modals must co-occur with a future marker like *dim* 'PROSP' if the prejacent eventuality follows the modal evaluation time). Thus, crucially, syntactic well-formedness, truth and felicity are theoretical notions that

the researcher must argue for; this information is not part of an acceptability judgment. See Matthewson 2013 for discussion of *dim* 'PROSP'.

Thus, we take a linguistic expression to be syntactically well-formed when it is judged to be acceptable in at least one context, and we take a linguistic expression to be syntactically ill-formed when it is consistently judged to be unacceptable in contexts that satisfy its hypothesized truth and felicity conditions. In practice, claims that an expression is syntactically ill-formed (rather than infelicitous or false) are often supported by native speakers' comments that the expression is not well-formed ('That's bad', 'That's not good English/Gitksan', 'That's how babies/foreigners talk', 'That's backwards' or 'We don't say it like that'), or by native speakers' suggestions of minimal variants that they judge to be acceptable.

5.2.2 Empirical evidence for hypotheses about felicity conditions

This section illustrates how minimal pairs of pieces of data with acceptability judgments can support a hypothesis about felicity conditions. The linking hypothesis is the following:

(46) **Linking hypothesis to explore felicity with binary acceptability judgments**
An expression that is judged to be acceptable in a context is syntactically well-formed, true and felicitous in that context. A syntactically well-formed expression that is judged to be unacceptable in a context in which it is true is infelicitous.

Our example is concerned with something that most (if not all) theoreticians would agree is a felicity condition, namely the fact that the English pronoun *he* requires a salient, male antecedent to be felicitous. Let's say we find that (47a) is judged to be unacceptable in the context of the piece of data, and that we want to argue that (47a) is infelicitous. We mark (47a) with a hash mark to indicate that we hypothesize that the unacceptability is due to a violated felicity condition.[27]

(47) a. Context: A woman walked in.
#He was wearing a hat.

b. Context: A man walked in.
He was wearing a hat.

The piece of data in (47a), together with the acceptable, minimally different piece of data in (47b), forms a minimal pair of type (31a-ii): utterances of the same sentence are judged in two minimally different contexts, and receive distinct judgments (by the second author of this paper); the context of the piece of data in (47b) satisfies the assumed felicity condition (the requirement for a salient, male antecedent), but the context of the piece of data in (47a) does not. Meanwhile, we hypothesize that the sentence is syntactically well-formed (supported by the acceptability judgment for (47b)), and that both contexts make the sentence true, modulo the violated felicity condition in (47b) (i.e., in both contexts, an individual is wearing a hat). Based on these hypotheses, we can conclude that the unacceptability of (47a) is due to the violation of the aforementioned felicity condition.

[27] In current practice, as noted in section 2, the symbol # is much more widely utilized than this.

5.2.3 Empirical evidence for hypotheses about truth conditions

This section illustrates how minimal pairs of pieces of data with acceptability judgments can support a hypothesis about truth conditions. The linking hypothesis is the following:

(48) **Linking hypothesis to explore truth with binary acceptability judgments**
An expression that is judged to be acceptable in a context is syntactically well-formed, true and felicitous in that context. A syntactically well-formed expression that is judged to be unacceptable in a context in which it is felicitous is false.

Our example is concerned with something that most (if not all) theoreticians would agree is a truth condition, namely the exhaustive inference of the English exclusive *only*. Let's say we find that (49a) is judged to be unacceptable given the facts established in the context of the piece of data, and that we want to argue that it is judged so because it is false. (We use the diacritic 'f' to mark that we hypothesize that the example is judged to be unacceptable since it is false; the diacritic % is also sometimes used, but it is also used to mark that there is speaker variation; # is also used.)

(49) a. Context: Jack went to karaoke with his friends; Jack and Mike sang.
f Only Jack sang.

b. Context: Jack went to karaoke with his friends; nobody other than Jack sang.
Only Jack sang.

(49a) and (49b) form a minimal pair of type (31a-ii): utterances of the same sentence are judged in two different contexts and receive distinct judgments (by the second author of this paper). If we hypothesize that both contexts satisfy the hypothesized felicity conditions of the uttered sentences (e.g., that there is a salient past reference time) and that the sentence is syntactically well-formed (which is supported by the acceptability of (49b)), then we can argue that piece of data in (49a) is judged to be unacceptable because it is false, i.e., because it is false that nobody other than Jack sang.

5.3 Summary: The power of minimal pairs

Empirical evidence in research on meaning consists of a single (positive or negative) piece of data, or pieces of data in minimal pair form, together with a linking hypothesis. This section characterized four different types of minimal pairs in research on meaning, and showed that they provide empirical evidence for different types of hypotheses. Specifically, minimal pairs establish the contributions to meaning of parts of complex expressions, they establish which feature of context contributes to interpretation, and they identify which expressions are compatible with which meanings. Minimal pairs are a powerful tool in research on meaning. As we showed, minimal pairs of pieces of data with acceptability judgments can provide empirical evidence for theoretical notions such as syntactic well-formedness, truth and felicity, given appropriate linking hypotheses. But they have also been successfully used to provide empirical evidence for much more subtle theoretical notions. We thus argue for a consistent use of minimal pairs in providing empirical evidence in research on meaning, in the interest of making such evidence as transparent as possible.

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 argued in section 3 that a complete piece of data has four components: a linguistic expression, a context, a response (task) and information about the speakers that responded. We argued that incomplete pieces of data are not stable, not replicable and not transparent, and therefore to be dispreferred. We furthermore argued in section 4 that acceptability and implication judgment tasks are preferred to other tasks, including paraphrase and translation tasks, because the former lead to stable, replicable and transparent pieces of data. And, finally, we showed in section 5 that empirical evidence in research on meaning consists of single pieces of data or pieces of data in minimal pair form, together with a linking hypothesis.

Our survey of research papers in the top four journals devoted to research on meaning revealed that there is a wide variety of views in contemporary research on meaning about e.g., whether a context is a constitutive part of a piece of data, which response tasks are used, which information about speakers to provide, and whether to use minimal pairs. The heterogeneity of current practices highlights the need for discussion and the establishment of consistent standards in our field. Our goal in this paper has been to kick off this discussion and the collaborative process of developing standards, and ultimately to contribute to improving the empirical basis of semantic and pragmatic theories.

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.

Alonso-Ovalle, Luis and Paula Menéndez-Benito. 2013. Modal determiners and alternatives: Quantity and ignorance effects. In *Semantics and Linguistic Theory (SALT) XXIII*, pages 570–586. eLanguage.

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 M. 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 of for semantic elicitation in the field and elsewhere. In M. 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. 2013. Beyond the past, present and future: Towards the semantics of 'graded tense' in Gĩkũyũ. *Natural Language Semantics* 21:219–276.

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*. Dordrecht, Heidelberg, London, 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.

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 M. 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.

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. On the role of context and prosody in the generation of scalar implicatures. Manuscript, The Ohio State University.

De Swart, Henriette. 1998. Aspect shift and coercion. *Natural Language and Linguistic Theory* 16:347–385.

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

Deal, Rose A. 2015. Reasoning about equivalence in semantic fieldwork. In M. 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.

Falaus, Anna-Maria. 2014. (partially) free choice of alternatives. *Linguistics & Philosophy* 37:121–173.

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* 2:1–34.

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, UMass, 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.

Ionin, Tania. 2006. This is definitely specific: Specificity and definiteness in article systems. *Natural Language Semantics* 14:175–234.

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

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

Kamp, Hans. 1981. A theory of truth and semantic representation. In J. Groenendijk, T. Janssen, and M. Stokhof, eds., *Truth, Representation and Information*. Dordrecht: Kluwer.

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.

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(4):369–415.

Matthewson, Lisa. 2006. Temporal semantics in a supposedly tenseless language. *Linguistics & Philosophy* 29:673–713.

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. 2013. Gitksan modals. *International Journal of American Linguistics* 79:349–394.

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.

Montague, Richard. 1970. Universal grammar. In *Montague 1974*. New Haven: Yale University Press.

Montague, Richard. 1974. *Formal Philosophy: Selected papers of Richard Montague*. New Haven: Yale University Press. Thomason, Richmond (ed.).

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. 1996. Information Structure in Discourse: Toward an Integrated Formal Theory of Pragmatics. In J. H. Yoon and A. Kathol, eds., *Ohio State University Working Papers in Linguistics*, vol. 49. The Ohio State University, Department of Linguistics.

Roberts, Craige. 1998. Information structure in discourse: Toward an integrated formal theory of pragmatics. Updated from 1996 publication in OSU Working Papers in Linguistics, vol. 49.

Roberts, Craige. 2012. Information structure in discourse: Toward an integrated formal theory of pragmatics. *Semantics & Pragmatics* 5:1–69. Reprint of Roberts (1998).

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.

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.

Sharvit, Yael. 2014. On the universal principles of tense embedding: The lesson from *before*. *Journal of Semantics* 31:263–313.

Soames, Scott. 1976. *An Examination of Freges 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, Natalia Zevakhina, and 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.