Empirical evidence in research on meaning

September 3, 2016

Abstract

Empirical evidence is at the heart of research on natural language meaning, but discussions of what constitutes such evidence are almost non-existent. Furthermore, the empirical evidence that is currently provided in research on meaning is heterogeneous and, we argue, not always as good as it could be. To help remedy this state of affairs, this paper advances a three-part proposal about the nature of empirical evidence in research on meaning. First, we argue that in order for empirical evidence to be robust, replicable and transparent, a piece of data in research on meaning consists of a linguistic expression, a context in which the expression is uttered, a response by a native speaker to a task involving the expression in that context, and information about the native speakers who provided the responses. Second, we argue that some response tasks, including acceptability and implication judgment tasks, are better suited than others (e.g., paraphrase and translation tasks) for yielding robust, replicable and transparent evidence. Finally, while some hypotheses are supported by positive or negative evidence alone, other types of hypotheses necessitate pieces of data in minimal pair form. Providing empirical evidence also means explicitly stating how the pieces of data support the hypothesis about meaning.

1 Introduction

Research on meaning has been thriving for many decades. However, even though empirical evidence for hypotheses about meaning is at the very heart of this research, there is almost no discussion in the literature, including textbooks and handbook articles, of what constitutes such evidence. This state of affairs would not be a problem if there was an implicit but generally agreed-upon understanding in the field of what constitutes empirical evidence about meaning. Unfortunately, this is not the case: the empirical evidence that is provided in research on meaning is heterogenous and, furthermore, not always as good as it could be. Specifically, there is heterogeneity in whether contexts are provided, in whether information about the response task and the responding speakers is included, in the quality of the response tasks that are used, in whether minimal pairs are provided and in whether it is made explicit how the pieces of data support the hypothesis about meaning (i.e., whether a linking hypothesis is made explicit).

As a concrete example of the heterogeneity of the empirical evidence that is currently provided in our field, consider the examples in (1) and (2), both of which are intended to support the hypothesis that a particular utterance is compatible with one interpretation but not another. The data in (1) are presented as evidence for the claim that the French sentence with *l'un l'autre* (the.one the.other) 'each other' can receive a reciprocal, but not a reflexive, interpretation.

(1) a. 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. 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 [(1a.ii)]. (Cable 2014:2)

The data in (2) are presented as evidence for the claim that Japanese present tense utterances with the comparative adverb *motto* give rise to a degree reading but not to a negative reading (under which the example would imply the the cake is not delicious).

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

The empirical evidence provided for the two theoretical hypotheses is strikingly different. In addition to the (glossed and translated) French sentence, Cable (2014) provides two disambiguating contexts, reports that the sentence was judged to be true in the reciprocal context and false in the reflexive context and provides information (in the acknowledgments) about the French speakers that judged the example. Sawada (2014), in contrast, merely provides a (glossed and translated) Japanese example marked with the diacritic '??' and a statement to the effect that the negative reading is not available.

In this paper, we argue that the empirical evidence provided in (1) is qualitatively better than the evidence in (2) because it is more **robust**, i.e., explicitly controls for factors that may lead to variation in speakers' responses, more **replicable**, i.e., facilitates attempts to reproduce the data in the same or another language, and more **transparent**, i.e., makes explicit how it supports the theoretical hypothesis.[1] In view of these three desiderata for empirical evidence in research on meaning, evidence like that in (2) is problematic because i) it does not control for factors that lead to variation in speakers' responses, ii) it is challenging to attempt to replicate even in the same language due to the lack of context and information about the response task and the speakers' responses, and iii) it does not allow readers to identify how the data support the hypothesis about meaning. In effect, what is given in (2) merely amounts to a claim that empirical evidence for the hypothesis about meaning exists, but it does not constitute empirical evidence.

Lest anybody be tempted to think that the heterogeneity of empirical evidence illustrated in (1) and (2) is exceptional and not representative of our field, that is not the case. We could have illustrated the heterogeneity with many examples from the literature and we will provide more representative examples as the paper proceeds. A survey we conducted of 40 journal articles published between 2012 and 2015 in the four leading journals in research on meaning (*Natural Language Semantics, Linguistics & Philosophy, Journal of Semantics, Semantics & Pragmatics*) also gives a sense of the heterogeneity of the nature of empirical evidence in research on meaning:[2]

[1] These three desiderata are fulfilled at different stages of research on meaning: while data is robust if it is collected in such a way that factors that lead to variation in responses are controlled for, data is replicable when it is reported in such a way that it facilitates attempts at reproduction. As discussed in sections 3 and 4, data transparency depends both on the data being collected with transparent response tasks and on reporting how the data supports the hypothesis about meaning. We thank an anonymous reviewer for this point.

[2] The 40 articles included in the survey cover a wide range of empirical phenomena and include data collected through introspection, one-on-one elicitation and quantitative research. Papers that primarily relied on secondary sources were not considered. We examined each article for the nature of the empirical evidence that was presented.

- Almost half of the papers presented pieces of data consisting only of a linguistic expression (usually a sentence), as in (2), i.e., without a context or information about the response task. In such cases, a hypothesis about meaning was stated and accompanied by the de-contextualized linguistic expression. This situation occurs often with particular linguistic phenomena, such as ambiguity, scope or presuppositions. For scope, 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, authors present a de-contextualized sentence as the sole evidence for the claim that the sentence is ambiguous, or has or does not have some presupposition.

- It was common to find pieces of data consisting of a linguistic expression and a context, but no information about the response task and the response. In other cases, authors only mention something akin to a response, e.g., that the expression was "(in)coherent", "(im)possible", "odd", "problematic" or "contradictory", without information about the response task or the response.

- There is no standard practice for including information about the speakers who provided the responses. Our survey revealed that only papers that presented results from quantitative research consistently include such information. In fact, the majority of papers in our survey did not include any information about the speakers whose responses were relied on. This practice is especially pervasive when the languages under investigation are widely spoken by linguists, such as English, German, Greek, Spanish or Korean (whether or not the authors are native speakers of the language under investigation).

This paper aims to kick off the collaborative process of developing consistent standards for empirical evidence about meaning. To this end, the paper advances a proposal about the nature of empirical evidence in research on meaning, regardless of whether the evidence is based on one-on-one elicitation of responses from native speakers (a.k.a. "fieldwork"), on the researcher's responses to response tasks about utterances in their language (a.k.a. "introspection") or through quantitative research (a.k.a. "experiments").[3] 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 quantitative research as well as research on languages we do not speak natively. The paper thus synthesizes and builds on insights from various strands of research on meaning in discussing and developing standards for what counts as empirical evidence. In view of the preference for empirical evidence that is robust, replicable and transparent, we define, in section 3, a piece of data to have four parts: a linguistic expression, a context in which the expression was uttered,[4] a response by a native speaker to a task about that expression uttered in that context, and information about the speakers who responded. Second, we argue (section 4) that some response tasks, including acceptability and implication judgment tasks, are better suited than others (e.g., paraphrase and translation tasks) for

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

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

yielding robust, replicable and transparent pieces of data.[5] Our third claim (section 5) is that, while some hypotheses about meaning can be supported by a positive or a negative piece of data alone, other types of hypotheses, such as the hypotheses discussed in connection with (1) and (2) above, necessitate pieces of data in minimal pair form. Different types of minimal pairs provide evidence for different types of hypotheses about meaning. Finally, providing empirical evidence also means explicitly stating the linking hypothesis, i.e., how the pieces of data provide support for the hypothesis about meaning. The paper concludes in section 6.

Before we get started, we note that there is another way in which research on meaning is heterogeneous, namely in whether the empirical evidence is collected through introspection, one-on-one elicitation or quantitative methods. Empirical evidence collected through these methods may vary along a number of dimensions, such as the number of responding speakers, the number of items considered and cognitive biases of the responding speakers, and may thereby lead to differences in how robust the empirical evidence is (for debate, see e.g., Wasow and Arnold 2005, Gibson and Fedorenko 2010, 2013, Culicover and Jackendoff 2010, Sprouse et al. 2013, Davis et al. 2014, Jacobson ms.). Which research method is appropriate depends, among other things, on the number of speakers available in the language, the extent to which the phenomenon has already been described in the language, how controversial the hypothesis is and how subtle it is. In this paper, we sidestep the question of the appropriateness of particular research methods and focus on the more fundamental question of what constitutes empirical evidence in research on meaning.

2 Previous discussions of the nature of empirical evidence

Semantics/pragmatics textbooks generally acknowledge the central importance of empirical evidence. Dowty et al. (1981:2), for example, write that “[i]n constructing the semantic component of a grammar, we are attempting to account [...] for [speakers’] judgements of synonymy, entailment, contradiction, and so on”. Larson and Segal (2005:9) assert that “[s]emantic facts...are verified by the judgments of native speakers” and Hurford et al. (2007:7) point out that “[n]ative speakers of languages are the primary source of information about meaning”. Cruse (2011:15) proposes that “native speakers’ intuitions are centre stage, in all their subtlety and nuances: they constitute the main source of primary data”. And Chierchia and McConnell-Ginet (2000:5f.) call speakers’ judgments “the core of the empirical data against which semantic theories must be judged”. However, none of the semantics/pragmatics textbooks contains a substantial discussion of the nature of empirical evidence.[6]

Volumes about research methods, including fieldwork methods, also fail to discuss what constitutes empirical evidence in research on meaning.[7] For one, beyond the lexicographic realm, semantic/pragmatic topics are rarely discussed. Several of these resources discuss tasks that native speakers can or should be

[5] This paper is limited to research on meaning conducted through offline measures, to the exclusion of, e.g., response time or eye movement measures, and to research on meaning comprehension, to the exclusion of research on production.

[6] 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.

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

asked to perform, and these discussions relate to what we argue is a component of a piece of data, namely a native speaker's response to a task.[8] Although there is frequent mention of the elicitation of minimal pairs in these resources, these are 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. Their presentation of how to develop and support hypotheses in phonology, morphology and syntax defines a piece of data as 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 responding speakers.

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

Like fieldwork-based research, quantitative research is also a comparatively recent development in research on semantics and pragmatics. Quantitative research on meaning builds on the principles of experimental design, methodology, and quantitative analysis used in research in the cognitive and social sciences, and parts of the proposals we advance here are already established practice in quantitative research on meaning (e.g., Johnson 2008). For instance, quantitative research already considers speakers' responses and information about the response task as components of a piece of data. Such research also typically involves minimal pairs simply by virtue of the fact that such research compares responses to one piece of data to responses to another, minimally different piece of data to make cause-effect inferences. Quantitative research also regularly engages in discussions about suitable experimental designs, including the tasks that speakers are asked to respond to (for an example, see Geurts and Pouscoulous 2009). With this paper, we hope to engage the wider community of researchers in a discussion about what counts as empirical evidence about meaning.

[8] 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 speakers to discuss whether a sentence can have particular meanings. However, asking speakers, whether they have linguistic training or not, what a sentence means does not yield robust, replicable and transparent evidence.

3 Pieces of data in research on meaning

We argue in section 5 that empirical evidence in research on meaning is provided by, depending on the hypothesis to be supported, positive pieces of data, negative pieces of data, or pieces of data in minimal pair form. To this end, the current section defines a piece of data in (offline, response-based) research on meaning as having four components:

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

a. a linguistic expression of language L,

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

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 responding speakers.

The next four subsections characterize the four components of a piece of data and argue that pieces of data that include these components are more likely to be robust, replicable and transparent than pieces of data that lack a context, information about the response (task) or information about the responding speakers.

3.1 The context of a piece of data

This section characterizes the context of a piece of data and then discusses what goes wrong when the context is omitted from a piece of data.

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

The context of a piece of data only includes those features of the context that the researcher hypothesizes to be relevant for the investigation. For example, the context of B's utterance 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 of B's utterance 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.[9]

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

B: Su-nā̀ màganā̀.
3PL-CONT talk

‘They were talking.’ (Mucha 2013:388)

(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.)

In (6), the context of the piece of data establishes information about the prior discourse structure. The two-person discourse in (6a) constitutes the context for the linguistic expressions in (6b).

(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 facts about the world, e.g., who slapped who in the two contexts in (7a), which were already presented in (1) above.

(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.

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

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. Of course, it may turn out later that some feature of the context was important for the investigation, but was not controlled for in the context of the piece of data, or that some other feature of context was not, ultimately, relevant, but was included in the context of the piece of data nevertheless. Including a context in a piece of data allows subsequent investigations to build on previous research by suitably 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, e.g., acted out, drawn or given in writing. In publications, the context of a piece of data may be written 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 given 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 given in the publication, under the hypothesis that essential features of the context remained the same across the speakers. Detailed information about contexts may be provided in an appendix, as is customary in quantitative research.

What goes wrong when there is no context? As discussed in section 1, many pieces of data in research on meaning do not include information about the context. The remainder of this section discusses what goes wrong when the context of a piece of data is omitted. First, it is well-known that a wide range of linguistic phenomena are context-dependent, including nominal, temporal, modal and aspectual reference, presuppositions, implicatures, discourse particles, and information structure. Furthermore, the context of the piece of data may influence 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 sentence of the language, i.e., a syntactically well-formed sentence, is influenced by context. Since contextual information matters for interpretation, contexts should be used in response tasks and when reporting pieces of data.

To illustrate, consider the de-contextualized piece of data in (8), which Moltmann (2013:36) argues “does not sound right” and marks with ‘??’.

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

The piece of data in (8) does not include a context (or information about the response task). Instead, we are left to infer that the expression was judged to be less than acceptable under the assumption that *Socrates* refers to the classical Greek philosopher and that the time at which (8) is uttered is a time after this philosopher's death. Under the assumption of a different context, e.g., one in which *Socrates* refers to a man called Socrates who is alive at the utterance time, or one in which *Socrates* refers to the philosopher but the example was uttered by a contemporary of his, the example is judged to be acceptable. Thus, speakers' responses to an acceptability judgment task about (8) vary depending on the context that is presented. Without such a context, the piece of data in (8) is not as robust as it could be.

Second, pieces of data presented without a context are less replicable, i.e., do not facilitate attempts to reproduce the data in the same language (e.g., from other speakers 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 important that maximally similar pieces of data are collected across different speakers or across different languages. If a researcher attempting replication of a piece of data does not have access to the context, the replicating researcher may unintentionally use a different context in their attempt at replication. (Likewise, if information about the response task is missing, they may unintentionally employ a different response task.) To illustrate, consider again the example in (8). The fact that this example does not include a context that fixes the referent of the name *Socrates* or the time at which (8) was uttered means that a researcher attempting to replicate this piece of data may use a different context, and hence obtain a different response. When the second author of this paper attempted to replicate (8) in Gitksan, she found that a speaker of the language judged the Gitksan variant of (8) given in (9) to be acceptable:

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

We now have an unfortunate situation on our hands since it is not clear whether the difference in acceptability of English (8) and Gitksan (9) is due to linguistically interesting variation (e.g., perhaps Gitksan does not have lifetime effects?) or merely due to the English and Gitksan speakers having given their responses relative to different contexts. For example, perhaps the Gitksan speaker in (9) silently imagined one of the contexts outlined above in which the English (8) would also be acceptable. In sum, omitting the context of a piece of data hampers replicability.

Pieces of data that omit information about the context are also often not transparent, i.e., do not make explicit how they support the hypothesis about meaning. The de-contextualized Japanese piece of data discussed in section 1, for instance, does not show how the author concluded that the sentence has one reading (the degree reading) and not another (the negative reading). That is, the piece of data does not show that the linguistic expression is judged to be acceptable in a context that supports the degree reading and judged to be unacceptable in a context that supports the negative reading. Therefore, the piece of data does not allow readers to identify it how supports the hypothesis about meaning it is claimed to support. Likewise, Frascarelli (2010) presents the Tagalog piece of data in (10) as evidence that the raising-like evidential construction is compatible with *Juan* being focused. Since, however, the example is presented without a context that would make clear the information structural properties of the linguistic expression in (10), the piece of data does not make clear how the author established empirical support for the hypothesis about meaning and, hence, does not provide evidence for the hypothesis.

(10) Si JUAN ang tila aalis bukas
DET.TRIG Juan DET.TRIG EVID ACT.FUT.leave tomorrow
'It's JUAN that seems to be leaving tomorrow.' (Frascarelli 2010:2134)

In sum, pieces of data that omit information about the context are less robust, less replicable and less transparent than pieces of data that provide such information.

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

> We begin with canonical examples of Navajo. [(11a)] . . . [has] Imperfective viewpoint [and is] taken as present in the absence of contextual information to the contrary. [(11b)] has the perfective viewpoint and is taken as past. The translations reflect the default temporal interpretations:
>
> (11) a. Jáan Tségh-áhoodzánídi naaghá
> John Window.Rock-in around-3subj-impf-go
> 'John is hanging out at Window Rock.'
>
> b. Shimá ch'iyáán ła' bá naháłii'
> 1-mother groceries some 3-for pref-1subj-perf-buy
> 'I bought some groceries for my mother.'

One issue with asking speakers to respond to expressions in null contexts is that the task is rather unnatural: utterances are not typically made in a completely empty context, devoid of any information about e.g., the interlocutors and the situation in which the utterance occurs.[10] It is possible that speakers who are asked to respond to expressions in a null context imagine a context in which the expression could be uttered (or could not be uttered) and that their response is influenced by that context. In this case, their response does not reflect the meaning of the expression in the null context provided by the researcher but rather in the context they imagine. The problem is that the researcher is not privy to this context and hence does not know which features of the context may have led to the response. As Crain and Steeedman (1985) put it: "The fact that the experimental situation in question makes a null contribution to the context does not mean that the context is null. It is merely not under the experimenter's control ... the so-called null context is in fact

[10] Another issue is that speakers are often asked to identify the context-independent meaning by providing a translation. See Matthewson 2004, Deal 2015, Bohnemeyer 2015 and section 4.1 for problems with relying on the translation task in research on meaning.

simply an *unknown* context" (p.338, italics in original). Consequently, getting a judgment in a null context does not necessarily reflect the context-independent meaning of the expression. See Tonhauser 2015:144 for a critique of using null contexts in research on temporal and aspectual reference.

Another example of the use of de-contextualized utterances to attempt to detect context-independent meaning comes from the literature on scalar implicatures. van Tiel et al. 2016 asked native speakers of English to judge whether de-contextualized sentences with weak scalar expressions (e.g., *She is intelligent*) give rise to a conversational implicature that denies a semantically stronger expression (e.g., 'intelligent, but not brilliant'). By presenting the sentences out of context, van Tiel and his colleagues claim to have established context-independent differences between different types of scalar expressions. However, as discussed in Geurts and Pouscoulous 2009:15, the very question whether a particular inference arises might itself change the context for the de-contextualized sentence with the weak scalar expression: for instance, a question about whether *She is intelligent* implies that she is intelligent but not brilliant might change the context in which *She is intelligent* is interpreted. Thus, again, presenting a de-contextualized sentence does not necessarily mean that the responding speakers do not consider a context.

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

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

Here, the same worry as mentioned above arises: When speakers respond to linguistic expressions presented without a context, the researcher has no control over whether they make up a context that the researcher is not privy to. One way to address this worry for the second type of research question is to present the expression in a context that makes clear that the expression is supposed to be uttered as the first or one of the first utterances of a discourse. We call such contexts 'discourse-initial':

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

One example of a discourse-initial context is in (13) from Gitksan (Tsimshianic). The hypothesis that was tested with this example was that the discourse particle =*ist* 'QUDD' indicates a downdate of the question under discussion (Gutzmann and Castroviejo Miró 2011), i.e., is infelicitous in a context in which the prejacent implication of =*ist* 'QUDD' (here, the proposition that Betty worked in Abbotsford) does not answer the current question under discussion (Ginzburg 1996, Roberts 2012). The context in (13) establishes that the speaker and the addressee know each other (they are married) but also, crucially, that there is no prior linguistic context: Adam and Betty have not yet raised a topic of conversation and, in particular, nothing about where Betty worked has been part of the conversation so far between the two. (A native speaker of Gitksan judged this example to be unacceptable. We return to this example in section 5.)

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

#G̲a'a=hl Abbotsford win ahle'lsd-'y**=ist**.
loc=cn Abbotsford comp work-1sg.ii=qudd

'I worked in Abbotsford today.'

It may, in some cases, not be plausible for the target utterance to be the very first utterance of a discourse. It is for this reason that the definition of a discourse-initial context in (12) allows for the relevant linguistic expression to be the first or one of the first utterances. For instance, in (14), an example from Paraguayan Guaraní (Tupí-Guaraní) that was judged to be acceptable by four native speakers, the relevant linguistic expression is uttered only after the mother has apologized on behalf of her daughter. The hypothesis that was explored with this piece of data was that sentences with the verb stem *–kuaa* 'know' are acceptable when the content of the complement clause is not something that both the speaker and the addressee know. Thus, a crucial feature of the discourse-initial context of (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.

(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)

For another example of a 'discourse initial' context, see the 'elevator' contexts used in Beaver and Zeevat's (2008) work on presupposition accommodation.

In sum, investigating meaning in null (or: out-of-the-blue) contexts is problematic because responding speakers may silently adopt their own imagined context. This situation can be addressed by establishing minimal, e.g., discourse-initial, context.

3.2 The linguistic expression of a piece of data

This section characterizes the linguistic expression of a piece of data and then discusses expressions attested in the literature that we argue are not suitable as linguistic expressions of pieces of data.

The linguistic expression of a piece of data in research on meaning can be a declarative sentence, as in the examples in (4), (6) and (7), but it can also be sentences in the interrogative or imperative moods, multi-sentence utterances as in (5), or sub-sentential expressions as in (15B).

(15) A: Who smokes?

B: Only John. (Coppock and Beaver 2014:401)

When the intonation of the linguistic expression is hypothesized to not be relevant to the hypothesis about meaning under investigation, the intonation of the utterance is typically not reported. If it is relevant for the hypothesis under investigation, the intonation is reported, e.g., using sophisticated transcription systems such as Tones and Break Indices (Beckman and Ayers Elam 1997).

Not all strings are linguistic expressions of a piece of data Not all strings are suitable as linguistic expressions of a piece of data. First, empirical evidence for a hypothesis about meaning in a particular language is not provided by a linguistic expression from another language. To illustrate, consider (16a), a Cheyenne sentence with a reportative evidential which contributes the proposition that Kathy sang and the proposition that the speaker has reportative evidence that Kathy sang. The English utterances in (16b) and (16c) are responses to the Cheyenne utterance in (16a). The acceptability of the English response in (16b) is taken to show that the proposition of (16a) that Kathy sang can be challenged, and the unacceptability of the English response in (16c) is taken to show that the proposition of (16a) that the speaker has reportative evidence cannot be challenged.

(16) a. *É-némene*-**sėste** *Kathy*.
3-sing-RPT.3SG Kathy
'Kathy sang, I hear.'

b.✓*No, she didn't (sing). She danced.*

c. #*No, you didn't (hear that).* (Murray 2014:4)

In (16), acceptability judgments about English responses, rather than about Cheyenne ones, are provided in support of a hypothesis about the meaning of the Cheyenne utterance in (16a). Since, in principle, English responses may be judged differently by Cheyenne speakers than Cheyenne responses, the empirical support for the hypothesis about Cheyenne comes from judgments about English responses, i.e., from judgments about a language other than the language under investigation.

Second, a piece of data in research on meaning should allow the reader to recover the meaning of the linguistic expression, both to allow for attempted replication of the piece of data and to allow the reader to identify how the piece of data supports the hypothesis about meaning. For an example that does not allow the reader to recover the meaning of the expression, consider (17), which is presented as evidence for the claim that a quantifier can bind a pronoun out of an adjunct (in the midst of other examples whose meanings can be recovered):

(17) . . . [after fetching each$_i$ pointer], but before dereferencing it$_i$. (Barker 2012:624)

Because (17) is presented in elided form (and without a context), it is difficult to recover the meaning of this example, which hampers both replicability and the readers ability to identify how the piece of data supports the hypothesis about meaning.

3.3 The response task and response

This section characterizes the response task and response of a piece of data and then discusses what goes wrong when the response task or information about the response is omitted from a piece of data.

A linguistic expression together with a context in which it is uttered does not yet make for a piece of data in research on meaning. What is missing — as evidenced also by the quotes from the textbooks given in section 2 — is a native speaker's response to a response task, e.g., a response to an acceptability judgment task or a truth value judgment task, or a translation of the expression into a different language. Even when we limit our attention to offline, response-based research, research on meaning is conducted using a wide variety of response tasks, including acceptability judgment tasks, implication judgment tasks, translation tasks and paraphrase judgment tasks.[11] (We review the main ones in section 4.) 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).

We refer to response tasks in the plural form since many of them can be implemented in several different ways. For instance, acceptability judgment tasks can differ in which specific question is asked (e.g., *Does this utterance sound good to you?* or *Is this utterance appropriate in this context?*) and in the response option provided to the native speaker (e.g., forced-choice binary responses, responses on a Likert scale, or magnitude estimations; see Schütze and Sprouse 2014 for an overview). Given the large variety of response tasks in research on meaning, the response component of a piece of data can take many forms, including 'yes', '3 out of 5', 'probably not' or 'Jane didn't read all the books', depending on the specific response task used. It follows that a speaker's response to a linguistic expression can only be understood in relation to the task that was used to elicit the response. For instance, it is only possible to understand whether a 'yes' response means that the speaker judges the example to be acceptable, unacceptable, or true if the particular response task that was used is identified. It thus follows that a piece of data includes information about speakers' responses as well as information about the task to which the speaker responded.[12]

Works reporting results from quantitative research typically include information about the response task in a methods section. In works that present pieces of data collected through introspection (i.e., researchers reporting their own responses to response tasks) or one-on-one elicitation, such information is sometimes included as part of the piece of data. For instance, the piece of data in (18), 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.).

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

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

[12] Speakers' comments can provide clues about the meaning of the expression as well as reveal what the actual judgment is, as discussed in Matthewson 2004 and Matthewson 2015. We thank an anonymous reviewer for reminding us of the fact that eliciting and reporting relevant comments has also been common practice in quantitative research with children. For example, Crain and Thornton (1998) discuss the value of eliciting explanations for 'no' responses in a truth value judgment task.

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

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

Other researchers opt to describe the type of judgment that was elicited and the speakers’ responses in the text preceding the piece of data. In general, descriptions of response tasks include the following information:

(19) **Description of the response task:**

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

b. the specific question posed to the native speakers,[13]

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

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

What goes wrong when information about the response task and the response are missing? As noted in section 1, including information about the response and the response task is currently not the standard practice in research on meaning, except in the quantitative literature. In some cases, authors merely mention something akin to a response, e.g., that the expression was “(in)coherent”, “(im)possible”, “odd”, “problematic” or “contradictory”, without information about the response task or the response. In other cases, authors refer to a specific task, e.g. “grammaticality judgment” or “felicity judgment”, but those task descriptions may not reflect what speakers were asked to judge. The term ‘felicity’, for instance, is sometimes used in research on meaning to mean ‘acceptability’ (as in Matthewson 2004:380) but can also refer to constraints imposed on prior context or to conditions that must be satisfied if the speech act is to be correctly performed (cf., ‘felicity constraints/conditions’). In section 4, we therefore refer to response tasks in a way that reflects the task that speakers were asked to perform.

Finally, it is common to find pieces of data reported in research on meaning without information about the response task and the response. Sudo (2014:283), for instance, reports that “[i]n [(20a)] the pronoun *her* can be construed as dependent on the quantifier *most other married men*, but this is not possible in [(20b)]”, without providing information about the response task or the responses. Likewise, Charlow and Sharvit (2014:4) claim that (21) has a “‘bound de re’ reading”, which “implies that the predicate *mother* is interpreted ‘de dicto”’ and they argue that “[f]or Keshet and two of our consultants, the interpretation of the pronoun *her* in *her mother* as a ‘de re’ pronoun biases *mother* towards a ‘de re’ interpretation”, again without providing information about the response task or the responses that substantiate this claim.

(20) a. John is sitting in front of his wife. Most other married men are sitting next to **her**.

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

b. (John and Mary are married.) John is sitting next to Mary. Most other married men are sitting next to **her**. (Sudo 2014:283)

(21) John believes that **every female student**$_i$ likes **her**$_i$ **mother**. (Charlow and Sharvit 2014:3)

When information about the response task and the response are omitted from a piece of data, the piece of data is less replicable since it is unclear what responding speakers are supposed to be asked and which responses indicate agreement versus disagreement between the speakers that gave the original judgments and the speakers involved in the replication. What were the native speakers of English asked to judge regarding the examples in (20) and (21) and how did they respond such that these pieces of data provide empirical evidence for the respective claims? Furthermore, since information about the response task or the responses is lacking, the empirical evidence is not as transparent as it could be: readers who are not native speakers of English might not be able to replicate the judgments for themselves and thus cannot reconstruct how the theoretical hypothesis about the interpretation of functional pronouns is supported by these pieces of data.

It is important to note that the diacritic that accompanies the linguistic expression (or the absence of such a diacritic) does not convey which task was responded to. Rather, diacritics indicate the researcher's interpretation of a speaker's response. If, for example, a judgment of acceptability is elicited for a linguistic expression and that expression is judged to be unacceptable by a native speaker, then the researcher may choose to mark the example with an asterisk (*) if she hypothesizes that the unacceptability is due to syntactic reasons, or with a hash mark (#) if she hypothesizes that the unacceptability is due to semantic/pragmatic reasons. Since diacritics are not consistently used in this way in the literature, one cannot reply on the diacritics to identify the response task and the response (for issues about the use of diacritics see also Schütze 1996:ch.2.3.3). For instance, even though the hashmark is often used to indicate that an expression is taken to be semantically or pragmatically anomalous, it is also used to indicate that an expression is not a paraphrase of another (Coppock and Beaver 2014). Likewise, the asterisk, though widely used to indicate syntactic ill-formedness, is also used to indicate unacceptability in particular contexts or under particular interpretations. Henderson (2014:49), for instance, marks (22) with an asterisk to indicate that the example does not mean 'Pairwise, the students ordered different drinks'. See also, e.g., Barker 2013 and Nicolae 2014 for uses of the asterisk to mean that an example does not have a particular interpretation.

(22) *The students ordered a different drink. (Henderson 2014:49)

In sum, the diacritic does not replace information about the response task or the response, and a piece of data is more replicable and transparent if it includes information about the response and the response task.

3.4 Information about the native speakers who responded

In this section, we characterize the information about the native speakers who responded and then discuss what goes wrong when this information is omitted from a piece of data.

It is generally acknowledged in linguistic research that different native speakers of a given language may differ in their responses to the same prompt. Native speakers may disagree, for instance, about whether a particular utterance is appropriate in a particular context. Different speakers may give different responses due to their dialect (e.g., Szmrecsanyi 2015), whether they have had linguistic training (Schütze 1996:§4.4.1) and their literacy and education (Schütze 1996:§4.4.2). Since different native speakers may give different responses, information about the speakers that provided the responses for a particular piece of data are an integral part of the piece of data. (It is also important to report when individual speakers' responses vary over time, if such information is collected.) In addition to information about the speakers' language background, age, linguistic training, etc., it is also important to report information about the number of speakers that provided judgments. In quantitative research, such information is typically provided in a methods section.

What goes wrong when information about speakers is missing Not providing information about the responding speakers is problematic for several reasons. First, when such information is missing, it is impossible to identify how generalizable the hypothesis about meaning that the data support is: did multiple speakers respond to the response task, or only one speaker, perhaps from a particular dialect? Hackl (2009), for example, makes claims about quantifiers in German but the piece of data in (23) with the expression *wer aller* 'who all' is judged to be unacceptable by the first author of this paper (also a native speaker of German).

(23) Wer aller hat den höchsten Berg beschneit?
Who-all has on the highest mountain snow made?
'Who all made snow on top of the highest mountain?' (Hackl 2009:72)

Since Hackl 2009 does not provide information about the speakers who provided the judgments for his paper, readers cannot identify the extent to which his claims about German generalize. Likewise, the second author of this paper judges B's utterance in (24) to be unacceptable, contrary to the claim in Toosarvandani 2014:24 that it is is acceptable.

(24) A: Why did we eat?
B: We were hungry. In fact, being hungry is not the reason we ate. (Toosarvandani 2014:24)

The robustness of the data in (24) would be heightened by the inclusion of information about the number of speakers who shared the judgment. In sum, omitting information about responding speakers makes pieces of data in research on meaning less robust.

Omitting information about responding speakers also means that it is impossible to identify what variation in judgments is due to. With respect to (23), for instance, it is possible that the variation is due to Hackl and the first author of this paper speaking different German dialects, given that *wer aller* 'who all' is widespread in Austrian German (Hackl's dialect), but not in High German or Swabian German (the first author's dialects). As a consequence, omitting information about responding speakers makes the pieces of data less replicable. Finally, omitting information about responding speakers is problematic because it makes it impossible to identify potential speaker bias. If, in research based on introspection, the researcher is the only one responding to the response tasks or if, in research based on one-on-one elicitation, the responding speakers are trained linguists with a vested interest in a particular hypothesis about meaning, then this information should be available to the readers.

In sum, providing information about the responding speakers makes for more robust and replicable data, serves to identify how generalizable the hypothesis is and also helps identify potential speaker bias.

3.5 Interim summary

In this section, we defined a piece of data in research on meaning as consisting of a context, a linguistic expression, a response by a native speaker to the expression in that context (plus information about the response task), and information about the responding speaker(s). We argued that pieces of data that include all four components are more likely to be robust, replicable and transparent than pieces of data that lack a context, a response, information about the response task or information about the speakers. We also argued that some strings are not suitable linguistic expressions for a piece of data.

What we put forward in this section about the constitutive parts of a piece of data is already established practice in parts of the contemporary literature on meaning. It is most consistently practiced in quantitative research, but it is also practiced in some research based on introspective judgments, as illustrated with the example in (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 authors provide information in their papers about who provided the relevant judgments: the author herself in the case of (6) and a French speaker for (7).) We argue that more consistent adoption of these practices will lead to improved quality in empirical evidence in research on meaning and will help avoid the pitfalls illustrated in the previous sections.

4 Response tasks in research on meaning

Having established the four components of a piece of data, we now characterize and critically assess the main types of response tasks used in offline research on meaning. Central to our assessment are the assumptions made for each task about how a native speaker's response to the task can be related to a theoretical concept of interest in research on meaning. As we show, it is these assumptions that reveal whether data that is collected using the task is robust, replicable and transparent. In the quantitative literature, the assumptions that underlie a particular task are referred to as the 'linking hypothesis' of the task: "[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" (Tanenhaus et al. 2000:564f.); linking hypotheses "are a necessary part of the inference chain that links theory to data" (p.565). Thus, linking hypotheses are hypotheses about how to interpret speakers' responses to response tasks. They are distinct from the hypotheses about meaning that semanticists explore, and from the predictions which derive from these hypotheses (as discussed in section 5).

In this section, we argue that some response tasks are better suited than others for yielding robust, replicable and transparent pieces of data. In particular, we argue that acceptability, implication and similarity judgment tasks yield more robust, replicable and transparent pieces of data than translation, paraphrase, entailment judgment and ambiguity judgment tasks, as indicated in Figure 1. Truth value judgment tasks can also yield robust, replicable and transparent pieces of data, depending on the linking hypothesis that is assumed (the hypothesis-dependent status of this task is captured by the '?' in the figure).[14]

In what follows, we refer to a response task that asks native speakers to perform X as an 'X task': for

[14] Our discussion is limited to the aforementioned tasks; for a discussion of other tasks with respect to whether theoretically untrained speakers can reliably perform them see Schütze 2008.

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

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

Figure 1: Evaluation of response tasks in research on meaning

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

4.1 Translation task

We begin our discussion of response tasks with the translation task, at the very left in Figure 1.

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. Tonhauser (2011:209), for example, offers up the de-contextualized Paraguayan Guaraní translation of the English example in (25) as evidence that, in Paraguayan Guaraní, "[i]n subordinate clauses, unmarked verbs are compatible with future time reference". Similarly, Matthewson (2006:676) provides the de-contextualized St'át'imcets example with the English translation in (26) in support of the claim that in St'át'imcets, "[s]uperficially tenseless sentences...can be interpreted as either present or past".

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

(26) táyt-kan
hungry-1SG.SUBJ
'I was hungry / I am hungry.' (Matthewson 2006:676)

The linking hypothesis that underlies the use of the translation task in (25) and (26), and many other examples presented in the literature, is that the "[t]he input to translation and the output of translation are equivalent in meaning" (Deal 2015:158). That is, the subordinate clause in the Guaraní example is taken to have future time reference because its English translation does. And the St'át'imcets sentence in (26) is taken to be interpretable with either present or past time reference because its English translations do.

Evaluation of the translation task If we could indeed assume the aforementioned linking hypothesis, then research on meaning would be much easier than it is: we would only have to identify what an utterance in one language means and could then ask for translations to identify how that exact meaning is conveyed in other languages. Unfortunately, this linking hypothesis cannot be assumed since the translation of a sentence may differ in its truth conditions, in its felicity conditions (presuppositions) and in its implicatures, as discussed extensively in the literature (see, e.g., Matthewson 2004, Krifka 2011, Bohnemeyer 2015 and Deal 2015, among others). The linking hypothesis that we can assume for translations, at best and even when the translations are provided by theoretically trained linguists, is that the input to translation and the output to translation have roughly the same meaning. Since having "roughly the same meaning" is rarely good enough for research on meaning, translations do not lead to robust or replicable data, and are at best a clue to meaning. 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, even if the meaning of the original language is fully understood, translations cannot provide empirical evidence for hypotheses about meaning because they only roughly convey the meaning of the original language.

4.2 Entailment, paraphrase and ambiguity judgment tasks

This section characterizes and discusses entailment, paraphrase and ambiguity judgment tasks, three types of tasks which by their very nature lead directly to information about theoretical concepts central to research on meaning, namely (mutual) entailment and ambiguity.

In the entailment judgment task, a native speaker is asked to judge whether an utterance of a sentence has a particular entailment. Crnič (2014), for example, states about (27): "that John read the book once is entailed by the proposition that John read the book twice" (p.176), thereby (presumably) illustrating an entailment judgment.

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

In the paraphrase judgment task, a native speaker is asked to judge whether one linguistic expression is paraphrased by another linguistic expression of their language. For example, Coppock and Beaver (2014) write about the examples in (28) 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).

(28) 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)

In the ambiguity judgment task, a native speaker is asked to judge whether an expression is ambiguous. To do so, the native speaker has to identify a context in which one of the two meanings of the expression is true and the other one is false, and vice versa. One example comes from Alrenga and Kennedy (2014), who state that the example in (29) "is ... ambiguous" (p.4) and then describe the two readings:

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

"Under one of its readings, [(29)] conveys that for each Tolkien novel *x* other than *Lord of the Rings*, the number of students who have read *Lord of the Rings* exceeds the number of students who have read *x*. [...] Under another reading, [(29)] instead conveys that the number of students who have read *Lord of the Rings* exceeds the number of students who have read all of the other Tolkien novels."

Evaluation of entailment, paraphrase and ambiguity judgment tasks Since the entailment, paraphrase and ambiguity judgment tasks ask speakers to provide judgments about theoretical concepts, the linking hypotheses that underlie these tasks are pretty trivial:

(30) Linking hypotheses for entailment, paraphrase and ambiguity judgment tasks

a. If a speaker judges that sentence S_1 entails (does not entail) sentence S_2, then S_1 entails (does not entail) S_2.

b. If a speaker judges that sentence S_1 is a paraphrase (is not a paraphrase) of sentence S_2, then S_1 has (does not have) the same truth conditions as S_2.

c. If a speaker judges that sentence S is (not) ambiguous, then S is (not) ambiguous.

As the task descriptions above and these linking hypotheses make clear, the entailment, paraphrase and ambiguity judgment tasks require responding speakers to have training in linguistics in order to perform the task: an understanding of truth conditions is required to assess whether one utterance has truth conditions that are at least as strong as those of another sentence (i.e., entails it), whether two utterances have the same truth conditions (i.e., are paraphrases of one another) or whether an expression has two distinct sets of truth conditions (i.e., is ambiguous). Anybody who has had the experience of teaching students the concepts of entailment, equivalence or ambiguity can attest to the fact that these concepts require training.

What makes these three tasks attractive, of course, is that they directly probe for theoretical concepts relevant to the analysis of natural language meaning, namely entailment and truth conditions. A disadvantage of these tasks, however, is that they cannot be performed reliably by speakers without linguistic training and consequently are less robust, less replicable and less transparent than pieces of data based on tasks that can be performed reliably by speakers without linguistic training. First, they are less robust because they may introduce confounds due to the linguistic training of the responding speakers. Second, they are less replicable because they can only be performed with speakers who have linguistic training but the vast majority of languages currently spoken do not have such speakers. And, third, they are less transparent because the linguistic analysis that a speaker is required to perform when responding to the task is not presented as part of the piece of data and, hence, that linguistic analysis is not accessible. (Consider, for instance, example (29) where the specific contexts that were used to assess that the example is ambiguous were not provided.) As a consequence, entailment, paraphrase and ambiguity judgment tasks are placed toward the lower end of our evaluation of response tasks in Figure 1. (See also Sprouse et al. 2013:§2.2 for the argument that tasks that require theoretically trained speakers are not ideal.)

We note that pieces of data that support hypotheses about entailment relations, truth conditional equivalence and ambiguity can, of course, be robust, replicable and transparent, as soon as other types of response tasks are used. For instance, Bohnemeyer (2015:34f.) points out that hypotheses about entailment can be empirically supported if the researcher constructs contexts and asks the speaker to judge whether the relevant expressions are true in the contexts: "the researcher is not asking for a direct judgment of entailment, but rather for a series of judgments about the truth of a pair of utterances in a series of scenarios" (p.35). Bohnemeyer also points to the possibility of supporting hypotheses about entailment using judgments of contradictions (*ibid*):

> Speakers appear to be able to tell relatively immediately whether two statements are logically consistent or not. Consequently, one method for testing whether an utterance has a given entailment is by combining it with a second utterance, which negates the hypothetical entailment. If in the speaker's judgment the conjunction of the two utterances may be true in the same scenario, this suggests that the proposition negated by the second utterance is not an entailment of the first. But if the speaker judges the utterances to be inconsistent, this supports the entailment analysis.

See also de Marneffe and Tonhauser accepted for the use of contradiction judgments in experimental research.

Pieces of data that support hypotheses about ambiguity can likewise be robust, replicable and transparent: as discussed in e.g., Crain and McKee 1985:104, to support a hypothesis about ambiguity the researcher can present the speaker with contexts that make one of the hypothesized meanings true and the other one false, and elicit judgments of acceptability of the relevant expression in these contexts from the speaker.

In conclusion, although pieces of data collected through entailment, paraphrase and entailment judgment tasks can provide direct insight into natural language meaning, a disadvantage of these tasks is that they yield less robust, replicable and transparent pieces of data than other tasks.

4.3 Truth value judgment tasks

A truth value judgment task was illustrated with example (7) from Cable 2014, repeated below for convenience. (Seth Cable confirmed in p.c. the use of a truth value judgment task in these examples.)

(7) Cable 2014:2

(7a) 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.

(7b) 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)].

In truth value judgment tasks, 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 (e.g., by handing a cookie to a puppet if the sentence is true). Speakers are typically asked to give a forced choice binary response (e.g., 'yes'/'no' or 'true'/'false'; e.g., Syrett and Koev 2014), though truth value judgment tasks with non-binary responses have also been used (see e.g., Chemla and Spector 2011, Abrusán and Szendrői 2013). For use of truth value judgment tasks with children see, e.g., Crain and McKee 1985 and Crain and Thornton 1998.

Truth value judgment tasks can only be applied to declarative sentences, to the exclusion of interrogative or imperative sentences. Furthermore, a theoretical assumption sometimes made 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.

Evaluation of truth value judgment tasks One linking hypothesis that is assumed for truth value judgment tasks is that a sentence that is judged by a responding speaker to be true (false) in a context is true (false) in that context. Thus, under this linking hypothesis, a responding speaker's judgment is taken to provide information about a theoretical concept, namely the truth conditions and truth value of a sentence in a context. Consequently, such tasks cannot be reliably applied with theoretically untrained speakers, who have not learned to distinguish the truth conditions of an utterance from other conditions on its felicitous and pragmatically unmarked use. For instance, anyone who has taught undergraduate semantics and has had to explain why a sentence like *John arrived or Mary arrived* is true in a situation in which both John and Mary arrived will appreciate that theoretically untrained speakers often do not differentiate between truth conditions and implicatures. Similarly, Soames (1976:169), von Fintel (2004) and Abrusán and Szendrői (2013) argue that speakers may judge utterances that are infelicitous to be false, even though they are assumed not to have a truth value. Thus, truth value judgment tasks under this linking hypothesis cannot be reliably applied with theoretically untrained speakers and, hence, pieces of data that involve a truth value judgment task are less than ideally robust, replicable and transparent.

The reason that truth value judgment tasks are nevertheless placed quite far to the right in Figure 1, and may even yield fully robust, replicable and transparent data, is that there are at least two ways in which such tasks can be reliably applied with theoretically untrained speakers. A first way is to ask speakers about the truth value of a sentence that the researcher hypothesizes to be syntactically well-formed, felicitous and pragmatically unmarked in the context in which it is presented. A speaker's 'false' response to such a sentence can then reasonably be taken to be due to unfulfilled truth conditions. For instance, Cable's (2014) example (7bii) was shown to be syntactically well-formed (since it was judged to be true in another context, namely (7aii)), and it was hypothesized to be felicitous and pragmatically unmarked in the context in which it was judged. Hence, a 'false' response to (7bii) can be taken as evidence that the truth conditions of the sentence in (7bii) are not fulfilled in the context in (7ai). The same holds for Syrett & Koev's (2014) experiment 4, where theoretically untrained speakers' 'no/false' responses were taken as evidence that the truth conditions of the utterances that were judged were not fulfilled.

A second way in which truth value judgment tasks can be reliably applied with theoretically untrained speakers is to assume a slightly weaker linking hypothesis: a sentence that is judged by a responding speaker to be true in a context is true in that context, but a sentence that is judged to be false may be false, or it may be infelicitous, or it may be true but pragmatically odd due to a conversational implicature, as in the disjunction example above. Thus, under this linking hypothesis, speakers 'false' responses are not taken to provide evidence that the truth conditions are not fulfilled, but merely as evidence that the speaker finds something about the sentence objectionable in the context in which it was presented. Thus, under this linking hypothesis, truth value judgment tasks can be reliably applied with theoretically untrained speakers, including children (see Crain and McKee 1985 and Crain and Thornton 1998), and pieces of data that involve this response task are robust, replicable and transparent.

In sum, whether pieces of data with responses to a truth value judgment task are robust, replicable and transparent depends on the linking hypothesis that is assumed and on whether the utterances were independently ascertained to be syntactically well-formed and felicitous.

4.4 Similarity of meanings judgment tasks

Similarity of meanings judgment tasks require speakers to judge the similarity of the meanings of utterances of two sentences (e.g., Schwarz 2007, Degen 2015, Matthewson 2015). For instance, to explore the scalar implicature *some but not all* that may arise from utterances with *some*, Degen (2015) asked speakers of English to judge how similar (on a scale from 1 to 7) the naturally occurring examples with *some* in (31a.i) and (31b.i) are to their constructed counterparts in (31a.ii) and (31b.ii), where *some* was replaced by *some, but not all.* (The examples were presented in contexts, which are omitted here.)

(31) Degen 2015:17

a. i. You sound like you've got **some** small ones in the background.

   ii. You sound like you've got **some, but not all,** small ones in the background.

b. i. I like **some** country music.

   ii. I like **some, but not all,** country music.

The linking hypothesis that underlies this task is that the more negative the response, the less similar the meanings of the two sentences are, and the more positive the response, the more similar the meanings are.

Evaluation of similarity of meanings judgment tasks Similarity of meanings judgment tasks are similar to paraphrase judgment task in that speakers are asked to compare the meanings of two utterances. However, unlike the paraphrase judgment task, similarity of meanings judgment tasks do not require speakers to perform linguistic analysis, i.e., to judge whether the two utterances have the same truth or felicity conditions. Thus, a positive response does not warrant the assumption that the two expressions have the same truth or felicity conditions and a negative response does not identify *how* the two utterances differ in meaning. As a consequence of the fact that similarity of meanings judgment tasks can be performed with theoretically untrained speakers, pieces of data based on such tasks do not suffer the same drawbacks as entailment, paraphrase and ambiguity judgment tasks and are consequently placed on the high end of our evaluation of response tasks in Figure 1.

Given that speakers only judge how similar in meaning two sentences are, similarity of meanings judgment tasks obviously do not provide insight into entailments or truth conditions. But, as Degen (2015) shows, they can provide insight into the scalar implicature that can arise from utterances with *some*. The scalar implicature *some but not all* can be assumed to be more clearly part of the speaker's originally intended meaning when two sentences are judged to be similar, as was the case for (31b), than when the two sentences are judged to not be similar, as was the case for (31a).

4.5 Implication judgment tasks

In an implication judgment task (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.[15] 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 here:

(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

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

'Who is this man?'

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

Thomas (2014) writes about this example that "[a]fter reading this discourse, consultants were asked whether they think that the man A is asking about is still the leader of the village" (p.394). For other uses of the direct implication judgment task, see e.g., van Tiel et al. 2016, Tonhauser et al. 2015.

In an indirect implication judgment task, the responding 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 of interest. This task was used in Tonhauser et al.'s (2013) investigation of projective content in Paraguayan Guaraní. With respect to the examples in (32), the implication of interest was that Marko used to smoke in the past. Rather than asking Paraguayan Guaraní speakers whether they would infer from (32a) or (32b) that Marko used to smoke in the past – a direct implication judgment – speakers were asked to judge whether Maria would give the medicine to Marko.

(32) 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 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.

The linking hypothesis of a direct implication task is trivial, again: if a speaker assesses that an utterance implies (does not imply) a particular proposition, then the utterance implies (does not imply) that proposition. For indirect implication judgment tasks, the statement of the linking hypothesis depends on the question the speakers were asked: if a speaker responds to that question (e.g., the question about whether Maria would give the medicine to Mario) in one way, then the utterance implies the relevant implication (e.g., that Marko smoked in the past), but not if the speaker responds in a different way.

Evaluation of direct and indirect implication judgment tasks Neither direct nor indirect implication judgment tasks require the responding speaker to have linguistic training, and thus data obtained with these tasks is replicable and transparent. If the pieces of data include a context (and information about the responding speakers), the tasks also lead to robust pieces of data. Neither task directly reveals the theoretical status of the implications that are diagnosed (e.g., whether it is a conversational implicature or an entailment), but implication judgment tasks can be used to this effect (see, e.g., Tonhauser et al. 2013).

4.6 Acceptability judgment tasks

In an acceptability judgment task, a native speaker judges the acceptability of an utterance of a linguistic expression in a context or chooses the most acceptable among two or more expressions (cf., e.g., Syrett and Koev 2014). In (18), repeated below, native speakers of Hausa were asked to judge whether the given sentence is appropriate to say in the context provided.

(18) 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 questions that might be posed to the speaker include 'Does this sound good to you?' or 'Would you say this?' (see Bohnemeyer 2015:36 for further examples).[16] 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 (see Tonhauser et al. 2013:fn.13 for a brief discussion).

One linking hypothesis for such tasks is the following (other, stronger linking hypotheses are discussed in section 5): If an expression is judged to be acceptable in a context, then that expression is syntactically well-formed, its felicity conditions are fulfilled in that context and, if the expression denotes a proposition, its truth conditions are fulfilled in that context. If an expression is judged to be unacceptable in a context, the expression is syntactically ill-formed, infelicitous, false or any combination thereof. Thus, under this linking hypothesis, a judgment of unacceptability does not by itself provide insight into why the utterance was judged so (see also Chomsky 1977:4 and Matthewson 2004:409).

Consider, for example, the English example in (33), which the second author judged to be unacceptable in the context in which it is presented. We use the diacritic '×' here to indicate that the sentence was judged to be unacceptable in the context in which it was uttered and to remain neutral about whether this judgment is due to the sentence being syntactically ill-formed, infelicitous or false.

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

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

From a judgment of unacceptability, we do not know whether (33) is syntactically ill-formed, infelicitous, false, or a combination of the three; it is up to the researcher to determine the reasons for the unacceptability judgment, in conjunction with the hypothesis under which (33) was elicited (see also Matthewson 2004:375). We discuss in section 5 how minimal pairs of data with acceptability judgments can be used to tease apart these different sources of unacceptability.

Evaluation of acceptability judgment tasks Acceptability judgment tasks appear at the right end of the spectrum in Figure 1, together with similarity and implication judgment tasks. Acceptability judgment tasks tap into properties of utterances that do not require training in linguistics to be reliably detected. We assume that speakers have conscious access to the concept of whether an utterance sounds good (see also Sprouse et al. 2013:220). As a consequence of being applicable with theoretically untrained speakers, the acceptability judgment task leads to pieces of data that are replicable, transparent and – provided a context and information about the responding speakers is given – robust.

4.7 Summary

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

The statements of linking hypotheses of these tasks make clear that with any of these tasks there is a trade-off between whether the task can be performed reliably by theoretically untrained speakers and whether the task provides direct insight into theoretical concepts, like entailment and truth conditions. Tasks that provide direct insight into such theoretical concepts cannot be performed by theoretically untrained speakers. And tasks that can be performed reliably by theoretically untrained speakers do not provide direct insight into theoretical concepts. With these latter tasks, the researcher has to do the work of connecting the judgments obtained to the theoretical concepts. In this section, we have argued in favor of tasks that can be reliably performed by theoretically untrained speakers. The following section addresses how pieces of data based on these tasks are turned into evidence for theoretical hypotheses about meaning.

5 Turning pieces of data into empirical evidence

A piece of data by itself is just that: a piece of data. It becomes empirical evidence for or against a hypothesis about meaning once a researcher states how the piece of data provides empirical support for or against that hypothesis. Our goal in this section is to illustrate how hypotheses about meaning are empirically supported, i.e., which types of pieces of data provide empirical evidence for which types of hypotheses. There are four types of pieces of data that can support hypotheses: positive pieces of data, negative pieces of data, and two types of minimal pairs of pieces of data.[17]

5.1 Evidence from positive pieces of data

A positive piece of data is one in which a speaker's response to the response task was positive. What counts as 'positive' is determined relative to the response task: e.g., a positive response to a binary acceptability judgment task is a judgment of acceptability, to a binary truth value judgment task is a judgment of truth, and to a direct implication judgment task is a judgment that the implication arises from the linguistic expression. The linking hypotheses in (34) relate speakers' responses to the theoretical assumptions about positive pieces of data; they are partial because negative responses are not considered here (but see section 5.2).

(34) a. **Partial linking hypothesis for an acceptability judgment task:** If an expression is judged to be acceptable in a context, the expression is syntactically well-formed, it is felicitous in that context and its truth conditions are fulfilled in that context.

b. **Partial linking hypothesis for truth value judgment task:** If an expression is judged to be true in a context, the truth conditions of the expression are fulfilled in that context.

c. **Partial linking hypothesis for a direct implication judgment task:** If an expression is judged to give rise to an implication in a context, the expression gives rise to that implication in that context.

To illustrate how positive pieces of data provide evidence for hypotheses about meaning, consider the hypothesis (from Mucha 2013) that temporally unmarked Hausa sentences are felicitous with past temporal reference. From this hypothesis, one can derive the prediction that B's utterance in (4), repeated below, is felicitous in the context of A's question. Given the linking hypothesis in (34a) for a judgment of acceptability, the positive piece of data in (4) provides empirical evidence that B's utterance is felicitous in the context of A's question (and also that B's utterance is syntactically well-formed and that its truth conditions are compatible with the context). Thus, it is under the linking hypothesis in (38a) that the piece of data in (4) constitutes empirical evidence for Mucha's hypothesis about temporal reference.

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

B: Su-nā̀ màganā̀.
3PL-CONT talk

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

In research on meaning, it can be useful to group together two or more minimally different, positive pieces of data. Take, for example, the hypothesis that Gitksan bare verb forms, like *ha'wits'am* 'crush' in (35), can denote habitual states in the actual world as well as habitual states only found in possible, non-actual worlds. From this hypothesis we can derive the prediction that the linguistic expression in the

[17] We limit our discussion to pieces of data based on response tasks that were identified in section 4 as leading to robust, replicable and transparent pieces of data.

examples in (35) is true in the context in (35a), which describes a situation in which the machine regularly crushes oranges in the actual world, and in the context in (35b), which describes a situation in which the machine has not yet crushed an orange.

(35) a. Context: This machine regularly crushes oranges.
Ha-'wits'-am olents tun=sa.
INS-squeeze-ATTR orange DEM=PROX
'This machine crushes oranges.'

b. Context: This machine was built to crush oranges, but has not crushed any yet.
Ha-'wits'-am olents tun=sa.
INS-squeeze-ATTR orange DEM=PROX
'This machine crushes oranges.'

Given the linking hypothesis in (34a), the fact that the linguistic expression in (35) was judged to be acceptable by a native speaker of Gitksan in the context in (35a) and in the context in (35b) supports the conclusion that the linguistic expression is true in both contexts, thereby supporting the aforementioned hypothesis. Thus, two or more positive pieces of data that differ only in the context in which the expression is judged can provide empirical evidence that the meaning of the expression is compatible with the different meanings conveyed by the contexts.

Conversely, grouping together positive pieces of data that differ minimally in the linguistic expressions can provide empirical evidence that the linguistic expressions are all compatible with a particular meaning. Consider the hypothesis that the exclusives *only* and *just* are both compatible with rank-order interpretations (Coppock and Beaver 2014). The context in (36) establishes that the fire alarm was not sounded because there was an actual fire emergency. Native speakers of English are taken to know that an actual fire emergency outranks a fire drill on a scale of danger. The fact that both (36a) and (36b) are judged to be acceptable in this context, and hence by the linking hypothesis in (34a) are true in this context, shows that both exclusives are compatible with the so-called 'rank order' interpretation of exclusives.

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

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

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

In sum, under the linking hypotheses in (34), positive pieces of data can provide empirical evidence for a wide range of hypotheses about meaning, namely any hypothesis about a felicity condition of an expression, as in the examples in (4) or (6), about the truth conditions of an expression, as in examples (7bi), (35) and (36), or about the implications that an expressions gives rise to, as in example (5). Under other linking hypotheses, positive pieces of data may provide empirical evidence for other types of hypotheses. Crucially, however, positive pieces of data alone cannot provide empirical evidence for hypotheses about which part of an expression contributes a meaning, or about which feature of context an expression is sensitive to. For these hypotheses, we need minimal pairs, which are discussed in section 5.3.

5.2 Evidence from negative pieces of data

A negative piece of data is one in which a speaker's response to the task about the linguistic expression was negative. What counts as 'negative' is again determined relative to the response task: e.g., a negative response to a binary acceptability judgment task is a judgment of unacceptability, to a binary truth value judgment task is a judgment of falsity, and to a direct implication judgment task is a judgment that the implication does not arise from the linguistic expression. The linking hypotheses in (37) relate speakers' responses to the theoretical assumptions about negative pieces of data.

(37) a. **Partial linking hypothesis for an acceptability judgment task:** If an expression is judged to be unacceptable in a context, the expression is syntactically ill-formed, it is infelicitous in that context, or its truth conditions are incompatible with that context, or a combination thereof.[18]

b. **Partial linking hypothesis for a truth value judgment task:** Given a sentence that is hypothesized to be syntactically well-formed, felicitous and pragmatically unmarked in the context in which it is judged, if that sentence is judged to be false in that context, then its truth conditions are not fulfilled in that context.

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

Consider Mucha's (2013: 384f.) hypothesis that a Hausa sentence cannot simultaneously have both past and present temporal reference (unlike in St'át'imcets, as discussed in Matthewson 2006). From this hypothesis, one can derive the prediction that the linguistic expression in (18), repeated below, is infelicitous in the context provided. Given the linking hypothesis in (37a), the negative piece of data in (18) provides empirical evidence that B's utterance is syntactically ill-formed, infelicitous or false, or a combination thereof. Thus, under the linking hypothesis in (37a), the negative piece of data in (18) does not yet provide empirical evidence for Mucha's hypothesis.

(18) 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)

[18] Chomsky (1977) points to this linking hypothesis when he writes: "we may make an intuitive judgment that some linguistic expression is odd or deviant. But we cannot in general know, pretheoretically, whether this deviance is a matter of syntax, semantics, pragmatics, belief, memory limitation, style, etc." (p.4).

Under a different linking hypothesis, namely the one in (38), the negative piece of data in (18) provides empirical evidence for Mucha's hypothesis. (A clue that (38) is the linking hypothesis Mucha assumes is that she refers to the judgment elicited for (18) as a "felicity" judgment (p.384).)

(38) **Partial linking hypothesis for an acceptability judgment task for sentences hypothesized to be syntactically well-formed and true in the context:** Given an expression that is hypothesized to be syntactically well-formed and whose truth conditions are hypothesized to be fulfilled in the context in which the expression is judged, if the expression is judged to be unacceptable in that context, then the expression is infelicitous in that context.

Given the linking hypothesis in (38), the negative piece of data in (18) provides empirical evidence that B's utterance is infelicitous in the context provided. Thus, it is under the linking hypothesis in (38) that the piece of data in (18) is transformed into a piece of evidence for Mucha's hypothesis about temporal reference.

Like positive pieces of data, negative pieces of data can provide empirical evidence for a wide range of hypotheses about meaning under linking hypotheses like those in (37) and (38), namely any hypothesis about the violation of a felicity condition of an utterance, as in example (18), about non-fulfillment of the truth conditions of an expression, as in example (7bii), or about an implication that the expression does not give rise to, as in example (32b). However, negative pieces of data alone, just like positive pieces of data, cannot provide empirical evidence for hypotheses about which particular sub-part of the expression contributes a meaning, or about which facet of context an expression is sensitive to. For these hypotheses, we need minimal pairs.

5.3 Evidence from minimal pairs

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

A piece of data in research on meaning is more complex than a piece of data in phonology and, consequently, there are two types of minimal pairs rather than just one. Specifically, a minimal pair in research on meaning consists of two pieces of data that differ minimally in either the linguistic expression, as in (39a), or in the context in which the expression is uttered, as in (39b), and that receive distinct responses.[19]

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

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

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

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

These two types of minimal pairs provide evidence for different types of hypotheses about meaning, as we show in this section. For reasons of space, we focus on minimal pairs of piece of data with binary acceptability judgments, which are perhaps more frequently used in research based on introspection and one-on-one elicitation than quantitative work. Minimal pairs of pieces of data with other responses are briefly discussed in section 5.3.3.

5.3.1 Minimal pairs of pieces of data with linguistic variants

A minimal pair of type (39a), in which both pieces of data have the same context but minimally different linguistic expressions, provides evidence that what differs between the two linguistic expressions contributes a particular meaning or results in a change in meaning. Consider the two pieces of data in (40). These share the same context, and the Paraguayan Guaraní linguistic expression in (40a) differs from the one in (40b) in the presence of the exclusive clitic *=nte* 'only' on the name *Javier*. The linguistic expression in (40a) was judged to be unacceptable in the context provided by four native speakers of Paraguayan Guaraní, whereas the linguistic expression in (40b) was judged to be acceptable by the same four native speakers. Thus, the two pieces of data in (40) form a minimal pair of type (39a).

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

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

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

Consider the hypothesis that the clitic *=nte* 'only' contributes an exclusive meaning like English *only*. Under this hypothesis, the sentence in (40a) would mean that Javier has a cow and nobody other than Javier has a cow. Thus, under this hypothesis, the example in (40a) is predicted to be false in the context in (40) since the context specifies that Javier and Maria each have a cow. Under the linking hypothesis in (41), the fact that (40a) was judged to be unacceptable by the native speakers means that (40a) is false.

(41) **Partial linking hypothesis for an acceptability judgment task for sentences hypothesized to be syntactically well-formed and felicitous in the context:** Given an expression that is hypothesized to be syntactically well-formed and to be felicitous in the context in which the expression is judged, if the expression is judged to be unacceptable in that context, then the truth conditions of the expression are not fulfilled in that context, i.e., the expression is false in that context.

Note that the unacceptability of (40a), under the linking hypothesis in (41), merely provides empirical evidence for the hypothesis that the entire linguistic expression in (40a) is false. In order to provide empirical evidence for the hypothesis that *=nte* 'only' contributes an exclusive meaning which renders the expression false, the negative piece of data in (40a) is combined with the minimally different positive piece of data in (40b). This example differs from (40a) only in the absence of *=nte* 'only'. Since it is judged to be acceptable, i.e., is true under the linking hypothesis in (34a), it is the combination of (40a) and (40b) that provides empirical evidence for the hypothesis that *=nte* 'only' contributes the exclusive meaning.

The linguistic expressions in minimal pairs of type (39a) may also differ in the order of parts of the expressions. In the minimal pair in (42), 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 (42a), or outside it, as in (42b).

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

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

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

Consider the hypothesis that the truth conditions of sentences with *–mo'ã* 'CF' differ depending on whether *–mo'ã* 'CF' occurs inside the negation circumfix, as in (42a), or outside of it, as in (42b); cf. Tonhauser 2009. This hypothesis leads to the prediction that there are contexts in which one sentence is true whereas the other one is false. Given the linking hypotheses in (34a) and (41), the minimal pair in (42) provides support for this hypothesis: (42a) is judged to be acceptable in the context in (42) and so, under the linking hypothesis in (34a), it is true; (42b), on the other hand, is judged to be unacceptable in the context in (42), and so, under the linking hypothesis in (41), is false. Again, both the positive and the negative pieces of data in (42) are required to provide empirical evidence for the hypothesis that the truth conditions of sentences differ depending on the position of pieces of data *–mo'ã* 'CF' with respect to negation.

In minimal pairs of type (39a), the linguistic expressions may also differ minimally in their constitutive parts. In the St'át'imcets (Lillooet Salish) minimal pair in (43), the two linguistic expressions differ in whether the inferential evidential *k'a* 'INFER' or the sensory-non-visual evidential *lákw7a* 'SNV' occurs after the sentence-initial focus marker. Let's assume that we have already established that *k'a* 'INFER' is an evidential that contributes the information that the speaker's evidence for their assertion relies on inference. Consider now the hypothesis that the evidential *lákw7a* 'SNV' has a different meaning and, specifically, is incompatible with inferential evidence. This hypothesis leads to the prediction that (43b) is infelicitous in the context of (43) since the context is designed such that inferential evidence obtains. We also expect (43a) to be felicitous in this context.

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

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

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

Given the linking hypotheses in (34a) and (38), the minimal pair in (43) provides empirical support for the hypothesis that *lákw7a* ‘SNV’ is incompatible with inferential evidence: (43a) is judged to be acceptable in the context in (43) and so, under the linking hypothesis in (34a), it is felicitous; (43b), on the other hand, is judged to be unacceptable in the context in (43), and so, under the linking hypothesis in (38), is infelicitous. Again, both the positive and the negative pieces of data in (43) are required to provide empirical evidence for the hypothesis that the felicity conditions of the sentences differ depending on whether *k’a* ‘INFER’ or *lákw7a* ‘SNV’ occurs.

In all of the examples we have presented thus far, the contextual information that is kept constant across the two members of the minimal pair appears before the target linguistic expression. But, of course, this contextual information may also appear after the linguistic expression. An example from St’á’t’imcets is given in (44). The hypothesis being tested here is that St’át’imcets noun phrases realized only with the plural (discontinuous) determiner *i...a* do not enforce reference to the maximal contextual salient set of individuals, in contrast to noun phrases that also contain *tákem* ‘all’. From this hypothesis, one can derive the prediction that the first clause of A’s first utterance in (44a), which only realizes the plural determiner on the noun ‘children’, can be continued with the claim that not all children are hungry, but the first clause of A’s first utterance in (44b), which also contains the quantifier *tákem* ‘all’, cannot.

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

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

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

A: Cw7it-7úl! Cw7áoy=t’u7 kw=s=tákem i=sk’wemk’úk’wm’it=a wa7 q’7-áol’men.
many-too NEG=just DET=NMLZ=all DET.PL=child(PL)=EXIS IPFV eat-want
‘That’s too much! Not all the children are hungry.’

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

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

A: #Cw7it-7úl! Cw7áoy=t’u7 kw=s=tákem i=sk’wemk’úk’wm’it=a wa7 q’7-áol’men.
many-too NEG=just DET=NMLZ=all DET.PL=child(PL)=EXIS IPFV eat-want
‘That’s too much! Not all the children are hungry.’

A St’át’imcets speaker judged the discourse in (44a) to be acceptable, and the discourse in (44b) to be unacceptable. Under the linking hypotheses in (34a) and (37a), these judgments support the hypothesis that the universal quantifier *tákem* ‘all’ does, but the plain plural determiner *i...a* does not, enforce reference to the entire set of contextually salient individuals in the discourse context.

In sum, minimal pairs of type (39a) can provide evidence that what differs between the two expressions of the minimal pair contributes a particular meaning, as in (40) and (44), or results in a change in meaning, as in (42) and (43). As discussed, both the positive and the negative pieces of data are necessary to provide empirical evidence for such hypotheses.

5.3.2 Minimal pairs of pieces of data with context variants

A minimal pair of type (39b), in which the same linguistic expression receives distinct acceptability judgments in two minimally different contexts, provides evidence that the meaning of the linguistic expression is sensitive to what differs between the contexts.

The contexts of two pieces of data are minimally different if they only differ with respect to the hypothesis that is being explored. To illustrate, take the hypothesis that Paraguayan Guaraní sentences with the clitic *=nte* ‘only’ entail an exclusive meaning; specifically, that the linguistic expression in (40a), repeated in (45), entails that Javier is the only person who owns a cow. Given this hypothesis, the two contexts in (45) are minimally different: in the context in (45a), Javier is not the only person who owns a cow, and in the context in (45b) he is the only person who owns a cow. Since the linguistic expression in the two pieces of data is the same, but receives distinct acceptability judgments in the two contexts, the two pieces of data in (45) form a minimal pair of type (39b). From the hypothesis under investigation we derive the prediction that the linguistic expression is false in the context in (45a) and true in the context in (45b).

(45) 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.’

The linguistic expression of the minimal pair is judged to be unacceptable in the context in (45a), and so, under the linking hypothesis in (41), its truth conditions are not fulfilled in that context. The same linguistic expression is judged to be acceptable in the context in (45b), and so, under the linking hypothesis in (34a), its truth conditions are fulfilled in that context. Thus, the positive piece of data in (45b) provides empirical evidence that the truth conditions of the expression are compatible with an exclusive meaning, and the negative piece of data in (45a) provides empirical evidence that the truth conditions of the expression are incompatible with a context that denies the exclusive meaning. Crucially, both the positive and the negative pieces of data in (45) are required to provide empirical evidence for the hypothesis that sentences with *=nte* ‘only’ entail an exclusive interpretation.

As a second example, consider the hypothesis that implicit subject arguments in Paraguayan Guaraní require a familiar third person antecedent discourse referent. From this hypothesis, we derive the prediction that a sentence with an implicit subject argument, like that in the examples in (46), is felicitous in a context that establishes a familiar third person antecedent discourse referent and infelicitous in a context that does not establish such a discourse referent. The minimal pair in (46) realizes the same linguistic expression in two contexts that differ only with respect to this hypothesis: the context in (46a) establishes a familiar third person discourse referent, but the one in (46b) does not.

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

Kuehe   che-su'u.
yesterday B1sg-bite

'Yesterday, it bit me.'

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

#Kuehe   che-su'u.
yesterday B1sg-bite

(Yesterday, it bit me.) (Tonhauser 2016)

Given the linking hypotheses in (34a) and (38), the minimal pair in (46) provides empirical support for the hypothesis that implicit subject arguments in Paraguayan Guaraní require a familiar antecedent discourse referent.

Finally, consider the hypothesis that the Gitksan clitic *=ist* 'QUDD' signals that the utterance addresses the question under discussion. The contexts in the Gitksan minimal pair in (47) differ minimally in whether a question under discussion is established: the context in (47a) (which is repeated from (13)) does not establish one and the one in (47b) does. From this hypothesis, we derive the prediction that the Gitksan sentence with *=ist* 'QUDD' is infelicitous in (47a) and felicitous in (47b).

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

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

'I worked in Abbotsford today.'

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

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

'I worked in Abbotsford today.'

Given the linking hypotheses in (34a) and (38), the minimal pair in (47) provides empirical support for the aforementioned hypothesis.

In sum, minimal pairs of type (39b) can provide evidence that the meaning of the linguistic expression realized in both members of the pair is sensitive to what differs between the two contexts. As discussed, both the positive and the negative pieces of data are necessary to provide empirical evidence for the hypotheses. With minimal pairs of type (39b), just like with minimal pairs of type (39a), the contextual information may follow the target linguistic expression. We provide such an example in the next section.

5.3.3 Minimal pairs for other response tasks

The two types of minimal pair we have illustrated above can, of course, also be formed with pieces of data that involve response tasks other than binary acceptability judgment tasks. The same types of minimal pairs can also be formed with forced choice truth value judgments, as illustrated with (7) and in Syrett and Koev's (2014) work, or with implication judgments: minimal pair type (39a) is illustrated in (32). And these minimal pairs can also be formed with pieces of data with (non-)binary responses to tasks, as are frequently used in quantitative research. In Amaral and Cummins (2015), for example, speakers were presented with Spanish dialogues like the ones in (48), and asked to judge the acceptability of the answer on a 5-point Likert scale. The minimal pairs in this task consist of dialogues: both dialogues realize the same question (which is the target linguistic expression) and minimally different answers (i.e., different continuations): the answers differ in whether the presupposition of the question (that Victoria was the director in the past) is denied, as in (48a), or not, as in (48b). This is thus a minimal pair of type (39b).

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

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

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

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

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

Amaral and Cummins (2015) found that dialogues like the one in (48b) received significantly higher acceptability ratings than dialogues like the one in (48a). Under a (presumed) linking hypothesis that one answer is preferred over another if the acceptability judgments of the first answer are significantly higher than the acceptability judgments of the second, this finding supports the hypothesis that answers that do not deny a presupposition are preferred over answers that deny a presupposition. For other illustrations of minimal pairs in quantitative research on meaning based on pieces of data with non-binary response tasks see, e.g., Chemla and Spector 2011, Degen 2015 and de Marneffe and Tonhauser accepted.

5.4 Summary

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

6 Conclusions

Even though empirical evidence is at the very heart of research on meaning, there has been little discussion in the literature of what constitutes such evidence. In particular, there has been little discussion of how speakers' responses to response tasks can be linked to and hence lead to empirical evidence for theoretical hypotheses about meaning. This state of affairs has led, we argue, to a situation in which empirical evidence presented in research on meaning is heterogeneous and, unfortunately, not always as good as it could be, i.e., less than ideally robust, less than ideally replicable and less than ideally transparent.

In this paper, we discussed the ways in which empirical evidence in research on meaning can be robust, replicable and transparent:

(49) **Desiderata:** Empirical evidence in research on meaning

a. is **robust** if it controls for factors that may lead to variation in speakers' responses,

b. is **replicable** if it facilitates attempts to reproduce the data in the same or another language, and

c. is **transparent** if it is explicit about how it supports the hypothesis about meaning.

Starting with pieces of data in section 3, we defined a piece of data as having four components: a linguistic expression, a context, a response (task) and information about the responding speakers. We argued that pieces of data that include these four components are more likely to be robust, replicable and transparent. Response tasks were the topic of section 4, where we characterized the main response tasks used in research on meaning, including linking hypotheses that underly these tasks. We argued that acceptability, similarity, implication and (under certain linking hypotheses) truth value judgment tasks are more likely than others, including paraphrase, entailment and ambiguity judgment tasks, as well as translation tasks, to lead to robust, replicable and transparent pieces of data. Finally, in section 5, we argued that empirical evidence in research on meaning consists of positive or negative pieces of data or pieces of data in minimal pair form, together with linking hypotheses that make explicit how the pieces of data provide empirical support for the hypotheses about meaning.

References

Abrusán, Márta and Kriszta Szendrői. 2013. Experimenting with the king of France: Topics, verifiability and definite descriptions. *Semantics & Pragmatics* 6(10):1–43.

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

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

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

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

Barker, Chris. 2012. Quantificational binding does not require c-command. *Linguistic Inquiry* 43:614–633.

Barker, Chris. 2013. Scopability and sluicing. *Linguistics & Philosophy* 36:187–223.

Beaver, David and Henk Zeevat. 2008. Towards a general theory of presupposition and accommodation. Ms.

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.

Beckman, Mary E. and Gayle Ayers Elam. 1997. *Guidelines for ToBI labelling, version 3.0.* The Ohio State University.

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

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

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

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

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

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

Charlow, Simon and Yael Sharvit. 2014. Bound 'de re' pronouns and the LFs of attitude reports. *Semantics & Pragmatics* 3(7):1–43.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

de Marneffe, Marie-Catherine and Judith Tonhauser. accepted. Inferring meaning from indirect answers to polar questions: The contribution of the rise-fall-rise contour. In E. Onea, M. Zimmermann, and K. von Heusinger, eds., *Questions in Discourse*. Leiden: Brill.

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

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

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

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

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

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

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

Frascarelli, Mara. 2010. Narrow focus, clefting and predicate inversion. *Lingua* 120:2121–2147.

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

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

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

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

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

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

Hackl, Martin. 2009. On the grammar and processing of proportional quantifiers: *most* versus *more than half*. *Natural Language Semantics* 17:63–98.

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

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

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

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

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

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

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

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

Johnson, Keith. 2008. *Quantitative Methods in Linguistics*. Malden/Oxford: Blackwell Publishing.

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

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

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

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

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

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

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

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

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

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

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

Murray, Sarah. 2014. Varieties of update. *Semantics & Pragmatics* 7(2):1–53.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Schwarz, Florian. 2007. Processing presupposed content. *Journal of Semantics* 24:373–416.

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

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

Sprouse, Jon, Carson T. Schütze, and Diogo Almeida. 2013. A comparison of informal and formal acceptability judgments using a random sample from Linguistic Inquiry 2001-2010. *Lingua* 134:219–248.

Sudo, Yasutada. 2014. Dependent plural pronouns with Skolemized choice functions. *Natural Language Semantics* 22:265–297.

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. 2016. The distribution of implicit arguments in Paraguayan Guaraní. In B. Estigarribia, ed., *Guaraní Linguistics in the 21st Century*. Leiden: Brill Publishing.

Tonhauser, Judith, David Beaver, Judith Degen, Marie-Catherine de Marneffe, Craige Roberts, and Mandy Simons. 2015. Negated evaluative adjective sentences: What projects, and why? Talk presented at Experimental Pragmatics conference, Chicago, July.

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

Toosarvandani, Maziar. 2014. Contrast and the structure of discourse. *Semantics & Pragmatics* 4(7):1–57.

van Tiel, Bob, Emiel van Miltenburg, Natalia Zevakhina, and Bart Geurts. 2016. Scalar diversity. *Journal of Semantics* 33:137–175.

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.