CHAPTER

2 A Guide to Stimulus-Based Elicitation for Semantic Categories 🔓

Asifa Majid

Abstract

Fieldwork is the collection of primary data outside of the controlled environments of the laboratory or library, and is the province of many scientists: biologists, geologists, anthropologists, as well as linguists. Traditional linguistic fieldwork has relied heavily on elicitation and observation, with a view to producing a grammar, dictionary, and texts. This is often accompanied by mining texts, i.e., narratives by speakers, for naturalistic examples. This sort of data can fruitfully elucidate lexical and constructional resources within a language, their formal properties, the kinds of expressions that occur, and so on. The article sets out a guide to the various stages of constructing a non-linguistic stimulus set in order to investigate semantic categories within a language. This should furnish a novice to this field with some of the key concepts and issues so that they can construct their own study. The focus in this article is how to use non-linguistic stimuli for a more thorough investigation of local semantic categories. Semantics is at the heart of linguistic description. The field linguist attempts to identify the sound units that convey distinctions in meaning — the lexical and grammatical classes that can be grouped together and distinguished for function, and so on. The bulk of this article sets out a guide to the various stages of constructing a non-linguistic stimulus set in order to investigate semantic categories within a language. This should furnish a novice to this field with some of the key concepts and issues so that they can construct their own study.

Keywords: data collection, semantics, non-linguistic stimulus, linguistic fieldwork, lexical resource

Subject: Linguistic Anthropology, Semantics, Linguistics

Series: Oxford Handbooks

Collection: Oxford Handbooks Online

2.1 Introduction[1]

Fieldwork is the collection of primary data outside of the controlled environments of the laboratory or library, and is the province of many scientists: biologists, geologists, anthropologists, as well as linguists (see also Senft 2009). Traditional linguistic fieldwork has relied heavily on elicitation and observation, with a view to producing a grammar, dictionary, and texts. The linguist might use a word list or questionnaire and ask a consultant, 'How do you say X?', probe for grammaticality judgements, or solicit translations. This is often accompanied by mining texts, i.e. narratives by speakers, for naturalistic examples. This sort of

p. 55 data can fruitfully ↳ elucidate lexical and constructional resources within a language, their formal properties, the kinds of expressions that occur, and so on. In recent years, with the widespread availability of cheap and portable recording technology, more and more field workers rely on audio or video recordings. This has made it easier for linguists to include a wider variety of linguistic materials in their repertoire, most notably to encompass everyday conversational data. Recordings make it possible to listen to speakers' utterances again and again, thus improving the quality of final transcripts and making it possible to update and refine analyses. Nevertheless there are limitations to these techniques, especially when it comes to understanding semantics, the topic of this chapter. As a result, there is a move—which has gained new momentum in recent years—towards using non-linguistic stimuli for elicitation as a means of exploring semantics.

The bulk of this chapter sets out a guide to the various stages of constructing a non-linguistic stimulus set in order to investigate semantic categories within a language. This should furnish a novice to this field with some of the key concepts and issues so that they can construct their own study. There are, however, many existing resources already available—off-the-shelf-materials, as it were (see §2.8 below). Stimulus sets have been developed for spatial and event categories, the language of perception and emotion, and more. These materials can be invaluable tools towards fulfilling traditional linguistic fieldwork goals, as well as serving as worked out examples of this approach.

As previously stated, the focus in this chapter is how to use non-linguistic stimuli for a more thorough investigation of local semantic categories. Semantics is at the heart of linguistic description. The field linguist attempts to identify the sound units that convey distinctions in meaning—the lexical and grammatical classes that can be grouped together and distinguished for function, and so on. Individual forms will be provided with glosses in translations. Some will receive fuller descriptions in dictionaries, and ideally will also be contextualized with respect to local cultural practice and knowledge (see Hellwig 2006 and Evans and Sasse 2007 on rich semantic description in linguistic fieldwork). This chapter aims to provide some basic tools and methodology to inform semantic analyses. But the methods discussed are not limited to the exploration of semantics by any means. The 'pear story' (Du Bois 1980) and 'frog story' (Berman and Slobin 1994) studies, which utilized picture and video-based stimuli, have led to key insights into morphosyntactic packaging and discourse construction. Creative use of non-linguistic stimuli could, without doubt, benefit linguistic analysis of most language phenomena.

This chapter also assumes a qualitative orientation to data analysis (§2.6), but the use of stimulus materials does not require such an approach. Stimulus-based elicitation can be used in conjunction with traditional data collection methods to increase the amount of primary linguistic material available and thus provide further information for qualitative description.

p. 56 Before embarking on the how-to guide, it is worth considering what the benefits of a stimulus-based approach are, and how it overcomes some limitations of conventional methods. Consider traditional field linguistic techniques, for example translation: find a consultant who speaks a contact language and ask them to translate word lists, simple sentences, etc. This can be an effective way to get into the linguistic system, but there are limitations. Using a specific formulation of a statement or word can 'prime' speakers

to produce a similar formulation (Pickering and Ferreira 2008), even though it may not be the default, or most natural, within the language. More importantly, much of the juicy semantic detail of the language cannot be elucidated this way. Another method is to mine 'naturalistic' data, i.e. elicited narratives, myths, etc. This is an important source of evidence, but when consulting these sources a term of interest may occur infrequently, or not at all, making it difficult to generalize or extrapolate further about meaning or grammar (Hellwig 2006). Elicitation using questions and acceptability judgements go a step further, but there are notorious difficulties in obtaining and interpreting such judgements. If asked to make grammaticality judgements, for example, speakers may reject sentences that are actually grammatical because they violate some prescriptive norm (Greenbaum and Quirk 1970), or because they are difficult to process (Bever 1970). Repeated questioning can lead to an increase in acceptability judgements for sentences previously thought to be ungrammatical (Dale 1978) as can embedding a sentence in the 'right context' (Bolinger 1968). These same issues plague semantic judgements too. Instead of asking for grammaticality judgements, the field linguist may want to establish whether an interpretation is semantically entailed or only pragmatically implied, whether two forms are synonymous, or whether they are taxonomically or partonomically related, and so on. To disentangle these issues, semantic judgements can be elicited (e.g. Cruse 1986; Lyons 1968). As with grammaticality judgements, these tests rely on native-speaker intuition and are therefore subject to the same sorts of problems. Does the speaker accept a sentence because it is semantically acceptable or because of context and repeated questioning, for example?

Non-linguistic stimuli can avoid these pitfalls. We need not presume meaning equivalence where there is translation equivalence, we no longer wait for a form simply to turn-up in a text, and we can avoid conflating linguistic data with metalinguistic judgements. And there are other advantages. Speaker descriptions of non-linguistic stimuli do not require special training of consultants or specialized knowledge, so data can be collected easily. Knowing precisely what the speaker saw when they produced the description minimizes erroneous interpretations on the part of the analyst, as the exact stimulus can be referred to later in the analysis stage, long after the utterance was produced. And the constant platform of the stimulus enables cross-speaker and cross-language comparison (cf. Berlin and Kay 1969; Levinson, Meira, and The Language and Cognition Group 2003; Majid, Boster, and Bowerman 2008). In the classical

study of colour by Berlin ↳ and Kay (1969), a palette of Munsell colour-chips was used to establish the boundaries and foci of colour categories across languages. The standardized colour-space provides an objective space where indigenous categories can be mapped, and cross-linguistic equivalence measured (e.g. Kay and Regier 2003; Regier, Kay, and Khetarpal 2007). Of course, non-linguistic stimuli are not without flaws (see §2.7), but used in concert with traditional methods, the linguist has a much richer, more nuanced and firmly grounded dataset to inform his or her analyses.

A stimulus-based approach, as laid out in this chapter, will enable the researcher to access one aspect of meaning, namely the extension. What things in the world are denoted by which forms in the language? This can feed into an analysis of intension or sense, the abstract linguistic meaning. Intension is often equated with the sets of relations which hold between linguistic forms (Cruse 1986: 15–20; Lyons 1968: 427–8), such as semantic relations of synonymy, antonymy, taxonomic inclusion, contrast, etc. A word's extension is a function of its intension, thus a study of extension informs our analyses of intension. The extensional array in a stimulus set serves as an etic metalanguage sometimes referred to as an etic grid (Levinson and Wilkins 2006:8; Levinson, Meira, and The Language and Cognition Group 2003: 487; Pike 1967). This is an objective array that makes a criterial number of discriminations, so that language-specific groupings—the emic concepts—can be identified. The rest of this chapter provides a guide to creating such a stimulus-set.

2.2 Formulating Your Research Question: Some Theoretical Considerations

The first step towards constructing your stimulus set is to identify your specific research question. It could be a modest endeavor: perhaps you are investigating the number system in your field language and you have a singular/dual/paucal/plural distinction. You wish to know the semantics of the paucal term. Is it, in fact, a quadral (meaning precisely four)? Does it exclude two? What is the maximum quantity it includes? Where does the boundary between the paucal and plural lie? In order to tease apart these questions you could construct a stimulus set of pictures, depicting differing amounts of various objects and ask participants to describe them. You could then establish the boundaries of the categories, and establish whether there was agreement across speakers.

The research question could be grander in scale. Many classical investigations of how semantic domains are categorized by lexical resources—in studies of colour, emotion, and folk biology, for example—attempt to quantify the extent to which ↳ there are universally shared or culturally relative categories. If you were to investigate number from a cross-linguistic perspective, then a stimulus set might be built differently than if you were only to investigate the boundaries of a 'paucal' category within one language. For cross-linguistic comparison, you would want to first consult existing resources on what the possible distinctions across multiple languages from different linguistic stock might be, and then be sure to build those into your stimulus set (more on this later). In any case, the first step is to articulate the scope and aim of the investigation.

2.2.1 Do I start with form or function?

There are alternate emphases in formulating a research question. Investigations differ in whether their primary focus is on formal class (e.g. adpositions) or conceptual domain (e.g. space). One approach is to begin with a form-class, or rather a specific form, and track the possible functions it can have. An alternative approach, common to most anthropological and psychological investigations, begins with a conceptual domain and then tracks how that is 'carved up' in different languages. Both approaches can lead to valuable insights about semantic categorization. Nevertheless, it is important to be aware of the underlying assumptions and problems of both the 'bottom-up' approach of beginning with form and the 'top-down' approach of beginning with function.

2.2.2 Beginning 'bottom-up'

The advantage of beginning with form is that the phenomenon of interest is relatively easily delineated and identified—at least within a single language, or related languages. To take an example from Haspelmath (1997), one could conduct a study on the semantics of the Instrumental case in Slavic languages, where the modern day reflexes of the Proto-Slavic forms *-mĭ,*-mi, and so on, were studied. Or we could take a lexical domain and compare cognate terms, for example, the English *break*, German *brechen*, Dutch *breken*, Swedish *bräcka*, etc., all of which spring from the single Proto-Indo-European root *bhreg-* (Majid, Gulberg, et al. 2007).

The disadvantages of this approach become clear as soon as we try to scale up to include other languages. What should we count as equivalent forms (Croft 2001; Haspelmath 2007)? Obviously, as we move away from Slavic languages, we are not going to find suffixes that have the precise instantiation of *-mĭ*, *-mi*, etc. We could, in response, expand the scope and criteria of our investigation. We could simply study Instrumental case across languages. The problem with this is that the definition used to identify 'case' differs from researcher to researcher (see Haspelmath 2007). Even if we could agree on a way to identify the

phenomenon of interest, the ↳ criteria used to identify Instrumental case in Language A may be difficult to apply or not relevant at all to Language B. Moreover, because a form stands in paradigmatic relation to other forms within a system, comparable forms across languages will not be equivalent since they will stand in different oppositional arrangement to each other (Boas 1911; Saussure 1966[1916]). Finally, this approach fails to capture the affinities between the job done by case in one language and that done by a prepositional adjunct phrase or serial verb construction in another. That is, the same function can be expressed in different forms across languages, and a purely bottom-up approach will miss these interesting points of comparison.

2.2.3 Beginning 'top-down'

In the top-down approach, we begin with a conceptual domain, such as 'colour' or 'body'. This approach has been criticized on various grounds. An often-voiced objection against studying domains such as 'colour' or the 'body' stems from questioning whether the domain actually forms a coherent construct for speakers. Critics point out that there are languages which lack superordinate terms for such domains and suggest therefore there is no universal *concept* for colour or body or whatever (Wierzbicka 2005; 2007). The assumption is that subordinate terms are not deemed to be a cohesive set by speakers, if the superordinate concept is not lexicalized. But this inference relies on two faulty premises. First, it supposes an isomorphism between words and concepts. Lack of a word does not imply lack of a concept. In fact, there is a whole research agenda devoted to uncovering how linguistic semantics and conceptual structure are related —and if our concepts are indeed limited to those which find lexical expression in language. This consequence is accepted grudgingly by some. Wierzbicka (2005: 220) states: 'It is true that the absence of a word does not prove the absence of a concept; but how does one prove the *presence* of a concept for which there is no word?' Non-linguistic behavioural responses, such as sorting pictures or videos, can provide such evidence (see e.g. Boster and Johnson 1989; Khetarpal et al. 2010; Malt et al. 2008). The second faulty premise is that terms do not form a semantic domain without a lexicalized superordinate. Words form a semantic domain if they have related meanings, are deemed similar to one another by speakers (synonymy), or opposites in meaning (antonymy). None of this requires the presence of a lexicalized superordinate. People access information from memory based on the semantic closeness of terms. If I say to you 'cat, sheep, horse...what other things are like this?', you are likely going to respond with 'dog, cow, goat, etc.'. There is a set of related terms here, whether the language has a word for 'animal' or not.

Another line of critique against a top-down approach worries about the neglect of the emic perspective, and lack of attention to language-internal structural considerations. By ignoring structural encoding, it has been argued, non-equivalent ↳ objects are being studied: in essence apples are being compared with oranges (see Lucy 1994 for a critique of the work on space, and Lucy 1997 on colour). Lucy argues that it is essential to begin with a structural analysis of the language, that is, first establish that the domain of study forms a coherent category on formal grounds. This is because a crucial component of the meaning of a word is determined by its combinatorial properties: the meaning of *A*, is determined by what construction *A* can enter into, and what other words, *B, C, D...*, enter into that construction, since these provide information about how these terms contrast. In the weak interpretation of Lucy's critique, paying attention to structural facts can reveal meaning components that cannot be discovered otherwise. In the strong interpretation, ignoring structural facts means the analyst is imposing categories that may not exist in the language in the first place: 'Lexical items are grouped together and analysed as a coherent set not because speakers of those languages group them together in a set as revealed, for example, by common grammatical treatment, but because the analyst so groups them' (Lucy 1994: 624).

Lucy (1997) argues that differences in meaning components exist with each difference in formal encoding. While English uses adjectives to express colour, in Kilivila nouns are used,[2] in Chinook particles, and in Samoan verbs (Dixon 1982). Examination of the distributional properties of English colour terms shows that

although they are all adjectives they do not form a homogeneous class. For example, we can *blacken (the chimney with soot)* and *whiten (your shirt with detergent)*, but we cannot **yellowen, *greenen*, or **blueen*. Why should the difference in grammatical encoding of colour categories matter? Because differences in grammatical encoding are associated with differences in meaning components. Nouns, for example, refer to more stable entities than verbs (e.g. Gentner 1982; Hopper and Thompson 1984; Sapir 1921), leading to the proposal that colour terms encoded in verbs encode the notion of change, or 'becoming', whereas colour terms encoded in nouns and particles refer to an intrinsic, unchanging property (Hickerson 1975). On the other hand, although there are associations between form class and meaning entailments, not every form difference results in a meaning difference[3] (Kay 2006). Thus, restrictions in morphosyntactic distributions can be due to grammatical arbitrariness, rather than meaning difference. The larger point made by Lucy nevertheless remains: by ignoring differences in grammatical encoding the analyst can mistakenly conclude that there is a unified construct (of colour, space, parts of the body, etc.), where in fact there is no such coherent category for the native speaker.

p. 61   While it is clear that one piece of evidence for a unified construct would be a unified encoding in the grammar of a language, other sorts of evidence can demonstrate that a set of terms together form a coherent construct. In Arrernte (a Pama-Nyungan language spoken in Australia), colour terms do not constitute a separate form class, but are part of a much larger set of terms to do with surface properties of objects (e.g. reflectance). Yet in a free word association test (where the consultant is given a word and they have to produce the first word that comes to mind in response), colour terms overwhelmingly elicit other colour terms, and are rarely elicited in response to non-colour terms (data from Wilkins, reported in Kay 2006). This data suggests that this 'close-knit semantic set' forms a coherent construct for speakers of Arrernte, even though they are not formally a class. Similarly, the body can be considered a coherent construct to the extent that speakers conceptually group terminology for the body together, regardless of the formal category the term falls into.

But how small or big can the semantic domain be? Wierzbicka and colleagues have argued that rather than looking for universals of colour, we should look for universals of the higher order concept of 'seeing' (because all languages have a word for seeing). But why stop there? If there is a named higher-order superordinate to 'seeing' (such as 'perceiving'), should we take perception as our domain instead? And if there is another collapsing (e.g. between perception and emotion), then a bigger domain again? On this, I advocate a pragmatic approach of 'fractal domain'. Investigations of subsets of semantic domains, if that is what they are, can nevertheless lead to real insights. For example, some languages include in their basic colour words information about texture or succulence (Conklin 1955; Lucy 1997), thus going beyond the psychophysically defined colour space. But these terms still display comparable restrictions in range over hue and lightness as dedicated colour terms (Kay and Regier 2003). That is, the extension of colour terms in psychophysically defined colour space is similar regardless of whether the term also has a meaning in terms of texture or succulence. Thus constraints on semantic range within a narrowly defined domain (i.e., colour as hue and lightness) have an identifiable structure not predictable from the wider uses and broader senses beyond it.

2.2.4 Summary

There are two different potential starting points for constructing your stimulus set. In one approach, you begin with form and trace possible functions. In the other, you delineate your conceptual domain and examine how it is populated by different forms. Both approaches have problematic aspects. No one study will be able to address all weaknesses. But care can be taken in minimizing these as much as possible. The

p. 62   point is to be conscious of these weaknesses and consider how ↳ they may affect your stimulus design and, more importantly, the interpretation of your findings.

2.3 Creating an Etic Grid

In some domains—such as colour—there is a well-worked-out psychophysical space that can help determine the selection of a stimulus set. But even with an informed array, there are still problems in how exactly to select stimuli. Take colour, for example—the actual number of colours discriminable to the human eye is in the millions, but obviously no study has used all of these in a naming task. Instead, a sub-selection is made of the colour space. In the World Colour Survey (Kay et al. 2009; Berlin and Kay 1969), 330 colour chips were selected that were equally spaced for hue and brightness, while holding saturation constant. Further work has separately examined the role of saturation variation (e.g. Boynton 1997). And, as we saw above, a stimulus space that also included variations in luminance, texture, and reflectance might be better to study how colour is categorized in language (Lucy 1997). Nevertheless, it is impossible to explore all possible dimensions or contrasts at once in a systematic fashion. Thus, it is important to set the priorities and scope of the investigation before constructing your stimuli.

2.3.1 Aren't the domains you can study restricted to concrete ones?

In principle there is no restriction on what domains can be handled using a stimulus-based approach. Recently, within the Language and Cognition group at the Max Planck Institute for Psycholinguistics, we have begun to investigate smells, tastes, and tactile texture using a non-linguistic stimulus-based approach (Majid 2007). There is also no restriction on lexical categories. The same logic and motivation could be used to investigate constructional resources as exemplified by the work of Bohnemeyer and Caelen (1999), Bohnemeyer et al. (2007) on event complexity, and by Evans et al. (2004), and Evans et al. (forthcoming) on reciprocals.

Wierzbicka (2009: 165) critiques the stimulus methodology on the ground that 'the most important things are invisible'. She argues that video clips and other such depictions cannot capture 'human values, moral categories, emotions, intentions, relationships or understandings', that what really matters for some sorts of concepts are motivations and projected outcomes rather than the physical acts ↳ themselves. No doubt there are serious challenges in depicting complex psychosocial states, but nevertheless there is reason to be optimistic about using a stimulus-based approach. When it comes to the interpretation of even the simplest of depictions, people go beyond the physical and interpret intentions, motivations and projected outcomes. Heider and Simmel (1944), for example, showed American undergraduates very simple cartoons featuring geometrical shapes, such as triangles and circles (see Fig. 2.1). These shapes were depicted as moving using a trick-film method, where the shapes are actually paper cutouts and are placed on a transparent plate. For the illusion of movement the shapes are moved small distances and then snapped at a location. The resulting snapshots are then played as one movie. When participants are shown these simple movies, they describe the movements depicted, not in terms of physical motion, but instead ascribing psychological intentions to the shapes. Most participants describe the video as a love triangle with two of the geometric shapes in an antagonistic relationship to the third, and provide elaborate stories accompanying each of the movement shifts. People also attribute personality traits to the individuals: aggressive, villainous, heroic, defiant, etc.

This type of inferencing is not limited to the American undergraduate. Social psychologists have been using this type of material to study cross-cultural differences in the attribution of causes (e.g. Morris and Peng 1994). Although these studies typically resort to urbanized populations in East Asia who may have familiarity with these sorts of materials anyway, we know from work in our own group that simple animations can also yield rich data from peoples not familiar with video technologies. The Heider and Simmel studies demonstrate that even from the simplest of cues people infer complex social and

psychological states, *contra* Wierzbicka. Moving from 2D black-and-white animations with geometric figures to naturalistic video would only increase the possible scenarios that could be depicted.

Figure 2.1.
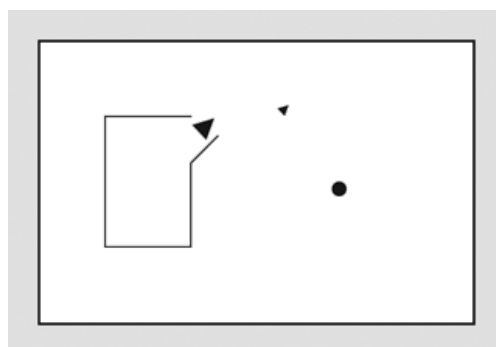


Adaptation of a still taken from Heider and Simmel (1944).

2.3.2 What contrasts to build into your stimulus set?

The stimulus space selected will depend on the question being addressed. Is the purpose a typological study? If so, then a first step could be to explore the available literature and identify the relevant cross-linguistic parameters. Another useful arena for potential parameters of relevance can come from language acquisition data. In our work on cutting and breaking events, we looked to developmental data to identify common errors children make in learning cutting and breaking verbs as a way to sketch out possible bordering domains (such as 'opening' and 'peeling'). We hypothesized that confusion errors attested in child verb usage may be the result of perceived similarity between breaking-type events and opening-type events, and that this similarity might also be reflected in cross-linguistic verb semantics (Majid, Bowerman et al. 2007; Majid, Boster, and Bowerman 2008) (see Fig. 2.2). We took a different approach when studying reciprocal constructions. In that investigation, we were led to test whether certain logical parameters (such as temporal organization, simultaneous vs. sequential and cardinality, or dual vs. plural) were of universal

relevance (Evans et al. 2004; Majid et al. 2011). ↳ The stimulus parameters in that study were thus derived from a theoretical framework, and only indirectly based on attested typological variation.

Figure 2.2.
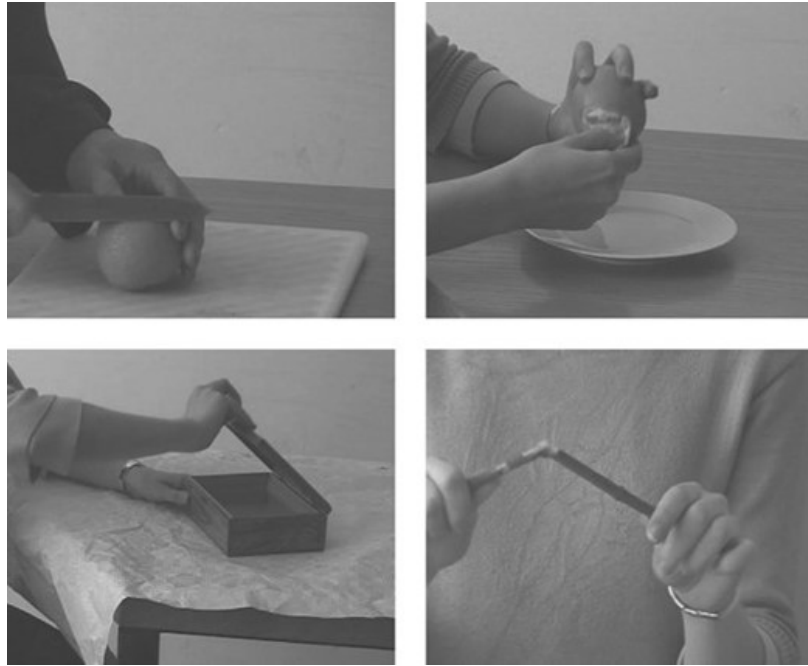
Stills taken from the 'cut and break' videoclips designed by Bohnemeyer, Bowerman, and Brown (2001), see also Majid, Bowerman, et al. (2007), and Majid et al. (2008).

Is the purpose to investigate a domain within a single language? Here the typological perspective can also be useful, but in addition, consulting your language corpora for potential hypotheses about the domain is invaluable (Hellwig 2006). In addition, you can ask speakers to 'free-list' exemplars from a domain ('Can you tell me all the types of X?'). This can provide possibilities for further investigation. Additional exploratory interviewing and observation beforehand can be productive for identifying relevant contrasts for constructing a stimulus set.

2.3.3 How many stimulus materials should I use?

There is a practical limitation to how many materials you should use. If you are doing a naming task, where consultants describe the materials, then a session with a 60-stimulus set can taken anywhere from 10 minutes to an hour to conduct (depending on how many ancillary questions you ask, and how free you invite your participants to be in their talk—a straight run through without additional questions should take around 15 minutes per participant on average). It is possible to use far larger stimulus sets (cf. the World Colour Survey: Kay et al. 2009), but participants can get bored and distracted, so the quality of data that you elicit may suffer. Also the more time it takes to run a single participant, and the less they enjoy it, the harder it will be to find additional volunteers for that task.

2.3.4 But there will be lots of things you miss!

No single stimulus set will address all of the possible factors of potential relevance to a semantic domain. Human fatigue is only one limiting factor. The larger the stimulus set, the more worrisome the possibility of participants becoming stuck in a response set (e.g. only producing a limited number of responses as a result of priming). So the selection of a stimulus set has to be clearly motivated. And, of course, the results will prompt you to pursue further hypotheses and avenues.

2.4 Collecting data

2.4.1 I have my stimuli, now what?

Construction of a stimulus set is the first phase, but thinking about how you are going to use your stimuli is just as important. When presented with a pictorial or ↳ video stimulus, a person may focus on many different aspects of the materials, and not necessarily the one that is critical for the researcher (Du Bois 1980). Unfamiliar objects and settings may call for commentary, joking, and curious speculations. Field researchers have reported more than once that consultants viewing videos of a woman with short hair were drawn to comment on her gender. Perhaps even more extreme, Segall and colleagues working in the 1960s on the cross-cultural perception of visual illusions reported various indigenous groups commenting more on the white band around the edge of a photograph than the picture depicted therein (Segall, Campbell, and Herskovits 1966). Such descriptions may be interesting in their own right, but more often they are tangential to the aspects of interest. Build into your elicitation task a period of familiarization. Cross-cultural research on the perception of pictorial materials suggests that a brief period of familiarization is sufficient to achieve a coherent interpretation of pictures (Miller 1973).

It is important to ensure that participants are familiarized with pictorial materials so that they can identify objects, especially as objects of a particular type. For example, in the Topological Relations Picture Series (also known as the 'bowped' pictures: Bowerman and Pederson 1992), which has been used to investigate spatial notions such as 'on' and 'in', spatial descriptions can vary either because the indigenous semantic categories are indeed different, or because the consultants misinterpreted the depicted objects and had a different construal of the scene. For example, in Fig. 2.3, if the ground is correctly interpreted as a ring, then a consultant might describe the apple as being 'in', but if they did not recognize the ground and instead interpreted it as a plate, then they might describe it as an 'on' relation instead. A casual inspection of the distribution of responses may lead a researcher to conclude that spatial relations are encoded differently in language X, because Fig. 2.3 was described as 'in' in all other languages but as 'on' within language X. But this would be a wrong conclusion. We simply do not know what speakers in language X would have said had they the appropriate construal in mind. So it is of utmost importance that the researcher reflect before beginning data collection: what am I trying to establish with this data session and how can I ensure that the speakers are oriented towards the relevant contrasts, i.e. those the stimuli are intended to depict?

Figure 2.3.
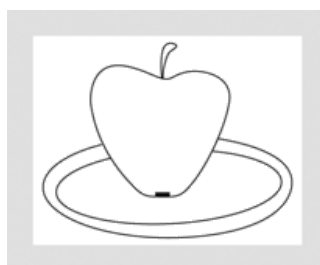
Taken from Bowerman and Pederson (1992).

The specific protocol being used in the elicitation also requires a phase of familiarization. If you are focusing on free spontaneous descriptions, then do not begin with the target stimulus materials straight away. Always include in your set two or three stimulus items that are not designed to test anything in particular, but are there to familiarize the participant with the task and the nature of the materials. Ask consultants to describe these first. Items should share general properties with the task materials (i.e. they are also videos

p. 66

p. 67

or pictures of the same sort). This helps orient the participant to what it is that you are hoping to get from them. It is an important step, because the kind of description you get can be qualitatively different when the consultant does not know what is expected of them.

2.4.2 What procedures can I use to elicit descriptions?

The simplest method is free description. Show participants the stimuli and ask them to describe the relevant stimulus. You can use the stimulus materials to elicit metalinguistic judgements too. As well as knowing what the most salient description is, you may wish to establish whether another descriptor could be used with that stimulus. This can be useful data but, as discussed in the Introduction, there are limitations to this method of data gathering. To illustrate, Labov (1978) showed speakers line drawings of containers and then asked them either to describe the dishes or to judge whether a drawing could be labelled *cup* or *mug*. He found that the data from free descriptions could be scaled to produce a perfect set of implication hierarchies. But explicit judgements did not show this regularity and were much messier overall (although see Malt et al. 2008 for a case where naming data and judgements do converge to reveal category structure).

It is important, however, not to mix procedures. If you are collecting free descriptions, do not ask explicit metalinguistic questions in the middle of the procedure. By asking for a reflective answer, you change the task from a purely linguistic task to a metalinguistic one, which can encourage strategic responding (see e.g. the discussion in Hellwig 2006). If you want to do both, complete the free description task, and then go back and ask for any metalinguistic judgements later. Always run free description tasks before any metalinguistic tasks, for the same reasons.

Another method to collect language data is to use a 'matching game' (Clark and Wilkes-Gibbs 1986). In this procedure, you have two sets of the non-linguistic stimuli and two participants. One set of materials is given to a 'director', the other set to a 'matcher'. They are screened off from one another so that they cannot see each other's stimuli. The director has to describe the materials one at a time such that the matcher can find the corresponding stimulus in their own set. Since participants are screened from each other, they have to rely purely on the verbal descriptions to complete the task. They can discuss each stimulus as long as they

p. 68

like, until they are both satisfied that they have a match, at which point they set ↳ aside that match and move on to the next one. When all stimuli are 'matched', the barrier between the participants is lifted and the consultants can compare their cards to see how many matched. This method will produce rich linguistic responses. The procedure eliminates the researcher from the linguistic interaction, and so speakers are interacting with each other in a free manner, and without having to tailor their language for a non-native speaker. Under this procedure, the speech event is no longer an interview but rather a collaborative endeavor between native speakers. Because the game is the focus, rather than the linguistic response, it is highly unlikely that consultants will be engaging in any kind of metalinguistic responding. So there are advantages to this method over a standard free description paradigm. On the other hand, there are limitations. The responses will not be as standard and therefore it can make interpretation trickier. The practicalities involved may also be difficult: two consultants need to be coordinated, the screen and recording need to be in place, you must monitor the matches/mismatches, etc.

2.4.3 How should I record the responses?

The ideal way to record responses is with an audio or video recording. Trying to note what people are saying while they are saying them is slow, clumsy, and error-prone. It slows down the elicitation process, and in the worst-case scenario participants will produce responses tailored to your activity. They produce simpler formulations because they see that you are struggling to write down what they are saying. Moreover, you have to make on-the-fly decisions about what the relevant response token is. By recording responses, you can later, with reflection, parcel out the bits of the response that are of interest.

2.5 Coding the data

The aim of coding your data is to be able to identify the referential range of the expressions that were used within the task. To do this, you must first transcribe the full responses of consultants, assuming you have recorded responses. You can then identify the constructional resources employed, how the semantic information you are interested in is distributed over the clause, etc. This can be a period to consider whether additional forms or constructions should also be part of the analysis. It is also a useful phase for weeding out responses which were not 'on target', by which I mean that the participant clearly had a different construal

in mind than what the stimulus item was designed to elicit. For example, if when working with spatial ↳ stimuli the participant reverses the figure and ground relations ('the table is under the cup' rather than 'the cup is on the table'), then this is not a target response. Or if, in describing a colour chip in a colour-naming task, the participant describes the shape, this is not on target. These can be interesting responses in themselves but do not speak to the intended depicted event or object.

In the process of collecting the data, consultants may have volunteered multiple responses, perhaps a first 'spontaneous' response, and then secondary responses. You may also have additional judgements about which other forms could be used, and whether some forms are not applicable. In the process of coding your data, it is important and worthwhile to tag your responses for whether or not they are the first response; whether they are free descriptions or reflective comments; and so on. In the analysis stage, you may want to consider separately each type of response.

2.6 Analysing the data

In order to explore the extensional semantics of a form (word, morpheme, construction), you will need to identify all of the stimuli that were described by that particular form. One simple way to do this is to code your data in a spreadsheet. You can create a column which has an ID for each of your stimuli, and alongside this insert your linguistic responses in the next column. A separate column for each speaker is helpful. You can then use the 'sort' function which will enable you to identify all the stimuli that were grouped together with the same term.

Standardized coding and eyeballing your data will get you quite far in figuring out the sorts of situations that call for descriptions of particular types but there are tools to go beyond this. There are a number of multivariate statistical techniques, such as multidimensional scaling, correspondence analysis, and cluster analysis, to name a few, which can be implemented in ordinary statistical packages, such as SPSS, SAS/STAT, or the open source freeware package R. These techniques are becoming increasingly important in the field of cross-linguistic semantic studies (e.g. Croft and Poole 2008; Majid et al. 2008). A comprehensive overview of these techniques is beyond the scope of this chapter, but for a gentle introduction see Grimm and Yarnold (1995).

2.6.1 Why do I need to use statistics?

There are numerous benefits to a statistical approach, but perhaps it is worth briefly outlining the perils of *not* using it. Let's say you have coded your data and ↳ have looked over the results, in a procedure akin to that described above. Now you are to examine the sorts of situations that receive a particular description and speculate on what feature of the stimulus may be calling for that form. The danger of relying only on intuition, here, is that people are notoriously prone to seeing correlations where there are none and missing them where they do exist (see Piattelli-Palmarini 1994 for an engaging illustration of all the ways we misjudge data). That is, you may be lured into thinking a feature of the stimulus is important when, in fact, it is not, or overlook something that is vital because you are not oriented towards it. This becomes all the more likely when you are dealing with multiple speaker data, or even multiple languages, and there is variation in precise extensional patterns. A statistical analysis can help in identifying which patterns are robust and reliable, and offer novel perspectives on your data.

There are multiple advantages of a statistical approach. Multivariate techniques can extrapolate more complex (multidimensional) solutions from your data, whereas working by hand will quite likely limit you to much simpler (and possibly only unidimensional) solutions. Most multivariate statistical tools also allow you to graphically represent your findings—a 'semantic map', as it were. These maps visually represent data in a manner that is much easier to absorb than reams of numbers. Finally, these techniques allow for quantification of how accurate a model is to the data, which means that we can assess our degree of confidence in any specific outcome.

2.7 Watch out! Pitfalls and dangers

A standardized stimulus set has numerous advantages. You do not need to wait for the phenomena of interest to turn up in spontaneous speech. You can efficiently map out the range of situations a term or construction applies to. And, in cross-linguistic (or cross-speaker) comparison, you can identify exceptional cases among common patterns for further exploration. However, there are potential pitfalls. The data obtained from a stimulus-based approach is only as good as the methods applied. It is crucial that you are clear and consistent in application of your procedures, coding, and most importantly in your reporting of your findings. This is essential for appropriate interpretation of your data.

More critically, the stimulus-based approach—like any other method of obtaining data—is not flawless. As discussed earlier, you have to be selective in the stimuli you use in any one sitting; the way you ask the question can lead to differences in responses; etc. But more egregious, perhaps, is the potential to overlook the emic perspective entirely. Even restricting ourselves to referential range, it is important ↳ to supplement the stimulus-based approach—like the one described here—with an examination of typical referents outside of the confines of the specific stimulus-set. Where else are these forms used? What are the typical foci? (See e.g. Conklin's 1955 classical study on colour.) It is important to explore how these forms are used *outside* the confines of the task to be able to interpret appropriately what they mean *within* the task. The stimulus-based approach does not replace corpus and verbal elicitation methods but supplements them and is supplemented by them. By combining methods, the limits of any one technique are overcome and converging evidence provides more confidence in findings (see e.g. Evans et al. forthcoming; Levinson and Wilkins 2006; Majid and Bowerman 2007; Majid and Levinson 2011).

2.8 Additional resources

For ready-to-go stimulus materials of the type described above, see the Language and Cognition Field Manuals and Stimulus Materials website, hosted by the Language and Cognition group, Max Planck Institute for Psycholinguistics: http://fieldmanuals.mpi.nl/

The Pear Story, a short film made at the University of California at Berkeley, is available on:

http://www.linguistics.ucsb.edu/faculty/chafe/pearfilm.htm; *Frog, Where Are You?* from the original Mayer (1969) book, as published in the Appendix of Berman and Slobin (1994):

http://childes.psy.cmu.edu/manuals/frog.pdf

Notes

1   Thanks to Loretta O'Connor, Niclas Burenhult, Nick Enfield, Gunter Senft, Stephen Levinson, Nick Thieberger, and two anonymous reviewers for thoughtful and critical comments on an earlier version of this chapter. The writing was funded by the Max Planck Gesellschaft.

2   According to Senft (1986), Kilivila colour words are adjectives.

3   This is a point that Lucy himself recognizes: 'It has long been taken as a truism in linguistics that meaning and form class are not usually in perfect one-to-one relationship. There will always be exceptional cases because of the historical and situationally contingent nature of linguistic structures' (1994: 649).