10 Studying Psycholinguistics out of the Lab

Laura J. Speed, Ewelina Wnuk, and Asifa Majid

Abstract

Traditional psycholinguistic studies take place in controlled experimental labs and typically involve testing undergraduate psychology or linguistics students. Investigating psycholinguistics in this manner calls into question the external validity of findings, that is, the extent to which research findings generalize across languages and cultures, as well as ecologically valid settings. Here we consider three ways in which psycholinguistics can be taken out of the lab. First, researchers can conduct cross-cultural fieldwork in diverse languages and cultures. Second, they can conduct online experiments or experiments in institutionalized public spaces (e.g., museums) to obtain large, diverse participant samples. And, third, researchers can perform studies in more ecologically valid settings, to increase the real-world generalizability of findings. By moving away from the traditional lab setting, psycholinguists can enrich their understanding of language use in all its rich and diverse contexts.

Research Methods in Psycholinguistics and the Neurobiology of Language: A Practical Guide, First Edition. Edited by Annette M. B. de Groot and Peter Hagoort.

Introduction

Taking part in a psycholinguistic study typically involves going to a university, meeting a researcher and completing a computer task in a quiet laboratory cubicle under the researcher's instruction. This chapter takes psycholinguistic research out of this traditional laboratory setting, and moves it to the outside world—both real and online. Such forms of research continue to use standard psycholinguistic *methods* for the most part, but change the research *location* in order to include a more diverse sample of observations into the consideration of psycholinguistic theories.

The use of diverse samples is a pillar of modern science. In order for generalizations to accurately portray a population, research samples must be representative, that is, selected so as to best reflect the population's diversity. Hence, psycholinguistics—a discipline whose major goal is to understand the mental representations and processes underlying human language use—must strive to be representative of the whole of humanity. It must not leave aside neglected populations (e.g., sign language users, bilinguals, aphasia patients, etc.), or culturally diverse groups. Given the challenges of reaching some of these populations, researchers need to venture outside the lab setting and take a more active role approaching people in their homes, schools, community centres, clinics, and so forth.

For studies conducted outside of the lab, we perceive there to be two general issues, both of which address concerns about external validity of research findings, that is, do current theories of psycholinguistics hold for the everyday language use of ordinary people conversing in the 7,000 or so diverse languages spoken today? In order to answer this question we need to know, first, whether established psycholinguistic phenomena generalize to other populations across the globe and out of the university setting. Second, we must establish whether observations made inside the lab can be replicated in ecologically valid settings outside of the lab.

In service to this broader goal, there are three motivations for being out of the lab: (1) reaching neglected populations, including speakers of diverse languages through cross-cultural studies in the field, (2) collecting large, demographically diverse samples within specific languages, which can be achieved through online experiments (i.e., crowdsourcing) or experiments in institutionalized public spaces (e.g., museums), and (3) increasing ecological validity of research findings by conducting pseudo-experiments in real-world settings. Both (1) and (2) typically employ traditional psycholinguistic experimental paradigms but reach a wider pool of participants; whereas (3) often requires further refinement of traditional methods to afford higher ecological validity. We separate (1) and (2) because they utilize different methodologies and experimental concerns, but many of the studies we discuss could fit into more than one category.

Our categorization is by no means exhaustive. Research outside of the lab can be conducted for other reasons, for example, theoretical reasons specific to a particular question. For example, researchers may seek alternative research settings in order to undertake manipulations not possible in the traditional psycholinguistic lab, such as manipulations of gravity in a space flight, which allowed Friederici and Levelt (1990) to investigate the perceptual cues used in order to determine spatial frames of reference in language. But the global issues—and potential benefits—of moving out of the lab should concern all psycholinguists.

Cross-Cultural Field Studies

Rationale

Why should we study diverse cultures and languages? It has long been recognized that the diversity of language is a window into diversity of thought. In the words of Wundt (1920)—the father of psycholinguistics—every language "represents its own characteristic organization of human thought" and as such may hide a "treasure" uniquely contributing to our understanding of how thought and language work (cf. Levelt, 2013).

Research in psychology has been heavily tilted toward a largely homogenous sample of Western undergraduate students (96% of study populations are from Western industrialized societies, who themselves only constitute 12% of the human population) (Arnett, 2008). This has been described as a "narrow" database (Sears, 1986).

The picture for psycholinguistics is largely similar. With the notable exception of language acquisition research (e.g, Bowerman & Brown, 2008; Slobin, 1985), most psycholinguistic research has been done with speakers of English, or other European languages. Jaeger and Norcliffe (2009), for instance, found sentence production research relies on data from only 0.6% of the world's languages (cf. Norcliffe, Harris, & Jaeger, 2015). This is problematic because English, and other "Standard Average European" languages, do not adequately portray the world's linguistic diversity (Dahl, 2015), and this leads researchers to disproportionately focus on patterns imposed by Eurocentric linguistic traditions (Whorf, 1944; Gil, 2001). Similarly, the sociodemographic characteristics of speakers typically participating in psycholinguistic experiments—that is, "WEIRD": Western, Educated, Industrialized, Rich and Democratic—make them unusual when compared to the rest of the world (Henrich, Heine, & Norenzayan, 2010; Majid & Levinson, 2010). For instance, there is a strong focus on monolinguals in psycholinguistic studies, which ignores the fact that worldwide multilingualism is rampant. In sum, an approach restricted to a largely homogenous sample fails to recognize the world's vast cultural and linguistic diversity (Evans & Levinson, 2009; Malt & Majid, 2013), tacitly assuming psycholinguistic universalism.

In reality differences in grammatical and semantic structure have differential consequences for the encoding and decoding of utterances (e.g., Norcliffe, Harris, & Jaeger, 2015; Levinson, 2012), and can affect general cognitive processes (e.g., Majid *et al.*, 2004; Wolff & Holmes, 2011). We focus here on the lesser-known languages spoken outside urban areas, but since cross-linguistic psycholinguistics is in its infancy, even relatively well-described languages (e.g., Tagalog), can offer novel insights (e.g., Sauppe *et al.*, 2013).

What Does It Entail? Best Practice

Each language presents a unique set of challenges to a researcher. The requirements and procedure followed in a field study will thus vary considerably from place to place depending on a number of practical and theoretical issues related to the field site logistics, sociocultural and linguistic background of the study population, state of language documentation, research questions, and so on. There are a number of

excellent guides (e.g., Bowern, 2008; Crowley, 2007; Sakel & Everett, 2012), and handbooks (e.g., Gippert, Himmelmann, & Mosel, 2006; Newman & Ratliff, 2001; Thieberger, 2011) on linguistic fieldwork, so we will only flag some key general issues, focusing specifically on psycholinguistic methods in the field.

The first prerequisite for successful psycholinguistic research in the field is familiarity with the language and culture under study. What this means in practice is long-term involvement with the community. If a language has not been previously studied, fieldwork will also require doing basic description to provide the groundwork for pursuing more advanced questions. If, on the other hand, a sufficiently good grammatical description already exists, getting to know the language will be easier. Knowing the language and culture is crucial not only because it enables you to interact with speakers and carry out experiments, but also to ensure you do not overlook important links. Since it is impossible to determine a priori how an underdescribed language works, fieldworkers cannot allow themselves the luxury of being interested only in syntax or only in morphology, but need a general mastery of the "whole language" (Hyman, 2001), and an understanding of its fit within the culture. For instance, sentence formulation is affected by word order, but at the same time it might also be driven by verb morphology (Norcliffe *et al*., 2015), while perceptual vocabulary might be intimately tied to cultural practices (e.g., Burenhult & Majid, 2011; Wnuk & Majid, 2014).

Stimuli and data collection in the field need not differ very much from lab studies, insofar as the employed method is itself suitable for the study population. Classical psycholinguistic paradigms (e.g., self-paced reading, lexical decision) were developed with literate populations in mind, so many standard methods need adaptation for cross-cultural usability (e.g., Wagers, Borja, & Chung, 2015). In principle, any task administered on a simple computer can be run in the field on a laptop. Needless to say, other (non-electronic), easily transportable stimuli such as pictures, booklets, small 3D objects can also be used in a field experiment. Transport and storage often requires careful planning—as does ensuring regular access to electricity—but there are a number of tips to deal with such practical considerations, for example, use of protective bags/boxes, lightweight solar powers, carrying backup equipment (e.g., Bowern, 2008). Thanks to the rapid development of technology, some specialized techniques—for example, ultrasound (Gick, 2002), eye-trackers (Norcliffe *et al*., 2015), EEG systems—have become portable and can also be used in psycholinguistic field studies. In some situations, it might also be possible to create field labs—enclosed quiet spaces—to approximate lab-testing conditions. So rather than moving the researcher out of the lab, we can now move the lab to the outside world.

Disadvantages and Pitfalls

As already mentioned, no two field sites are identical, so there is no single set of pitfalls for psycholinguistic field research. There are, however, some general issues to keep in mind. Of these, we would like to single out three we consider most important in the context of the present discussion: the practicalities of working with naive participants, small participant pools, and limited experimental control. For an extensive discussion of the general challenges of carrying out linguistic fieldwork, see Crowley (2007).

One important concern to keep in mind is the practicality of working with people who are not used to being tested. Many non-urban communities do not have formal education, and are not socialized into being compliant responders. Things that seem unproblematic from the point of view of university students, who spend hours listening to lectures and writing exams on a daily basis (e.g., performing repetitive tasks), can be highly demanding for other people (see also Whalen & McDonough, 2015). Care also has to be taken that modern equipment and testing are not intimidating to participants. So avoid straining participants with endless questionnaires or tedious procedures.

A second issue to consider is the limited common ground between the experimenter and participants, for example resulting from distinct cultural backgrounds. Sometimes, conveying the point of an experiment might be difficult, especially if it includes concepts with no direct translation equivalents in the target language. For these reasons, it is important to keep the design as clear and simple as possible: pilot the task and include a training phase. With growing knowledge of the language and community, researchers learn to anticipate participants' reactions and potential misunderstandings, so challenges of this kind usually become easier to navigate.

Another issue concerns the difficulty of recruiting large numbers of participants in the field. Understudied languages are often spoken by small communities so the participant pool can be relatively small. A possible solution is to increase the number of stimuli, so there are more critical data-points to feed into the analysis. Note, though, there is a trade-off between the duration of the experiment and data quality, as people might become tired more easily or even be reluctant to participate. To maximize the chances of recruiting people, it is important to plan the field trip at the right time. It may not be a good idea to visit a farmer community during harvest, for instance. Another related constraint has to do with potential societal stratification along gender or class lines. It might be socially inappropriate for fieldworkers to talk to community members of the opposite gender or of certain social classes. In these cases, it can help to recruit a local third person to accompany you, or perhaps even administer the task.

Finally, it can be difficult to have full experimental control in the field. Many fieldwork locations have little or no infrastructure. There is often no available separate, enclosed space for testing. So disruptions can include background noise and inquisitive observers. You can take various precautions to avoid these—for example, find a quiet spot out of the way, politely ask not to be disturbed, and so on. Again, further familiarity with the people and local environment can help optimise testing conditions.

Exemplary Studies

An example of a psycholinguistic study employing a diverse sample is the "Cut & Break" project (Majid *et al.*, 2007; Majid, Boster, & Bowerman, 2008). The project investigated event categorization across 28 diverse languages using a set of video clips depicting physical separation events (cutting and breaking). Speakers—interviewed in their native languages by a team of expert linguists—were asked to view the clips and provide free descriptions of each event. From the full descriptions, the verbs describing the target physical separation events were used to create a clip-by-clip similarity matrix for each language. Pairs of events were deemed similar (i.e., assigned a similarity score of 1) if they were ever described with the same verb, otherwise they were deemed dissimilar (i.e., assigned a score of 0). The stacked

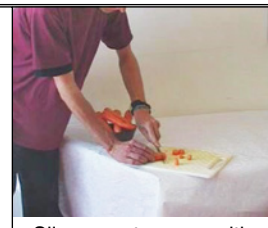
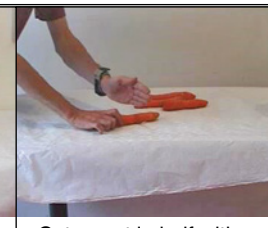
	 Slice carrot across with knife	 Cut carrot in half with a karate chop of hand	 Snap twig with two hands
Chontal	*te-k'e*		*tyof'ñi-*
Hindi	*kaaT*	*toDl*	
Jalonke	*i-xaba*	*i-sɛgɛ*	*gira*

Figure 10.1 Comparison of cut and break verbs in Chontal, Hindi, and Jalonke (adapted from Majid *et al.*, 2007). (*See insert for color representation of the figure.*)

similarity data was then fed into a correspondence analysis to extract the main dimensions of variance. The analysis revealed that although languages vary considerably in how they categorize events (see Figure 10.1), there is a common core underlying the structure of the domain across languages. To verify the results, the authors correlated the dimensions extracted by the general solution across languages with those for each individual language. Overall, the individual languages correlated highly, as reflected in high mean correlations and small standard deviations. Additional analyses with factor analysis and cluster analysis further confirmed a common space of event categorization across languages.

Thanks to the approach involving an "etic" grid—a standardized, language-independent stimulus set—it was possible to carry out a large-scale comparison at a general level, while the specialized expertise of the team of fieldworkers also enabled researchers to include the "emic" perspective—that is, a language- and culture-specific internal perspective (cf. the contributions in Majid *et al.*, 2007).

Studies Conducted Online and in Museums

Rationale

If as a psycholinguist you are not ready to pack your bags and jet-off to remote destinations to test the generalizability of your studies, you can still make efforts to broaden your participant sample so it is more inclusive and representative. Online studies and museums have both been the locus of a flurry of studies recently. Although on the surface they seem quite different, they are motivated by the same considerations so we discuss them here together.

Placing an experiment or survey online allows access to an impressively large number of participants, at all times of the day, every day of the week. Amazon

Mechanical Turk (MTurk), an online crowdsourcing site, permits the researcher to test over 100,000 participants in over 100 different countries (although the majority are based in the USA). Burhmester, Kwang, and Gosling (2011) report participants on MTurk are significantly more diverse than typical samples from American universities. Similarly, museums have a continuous flow of visitors almost every day, providing access to an impressively large number of people during opening hours. London's Science Museum has around 2.7 million visitors each year. Participants recruited online and in museums will represent a more diverse sample than typical psycholinguistic studies, and may even provide access to specialist populations, such as individuals with rare cases of synesthesia who are otherwise difficult to reach.

There may also be qualitative differences between participants recruited in universities and those recruited online and at museums. Participants from universities are likely to represent a volunteer bias. Results from people putting themselves forward for experiments might not be representative of the general population. Ganguli *et al.* (2015) found study volunteers tend to be younger, better educated, healthier, and have fewer cognitive impairments than participants randomly selected from the population. In addition, participants in universities typically get paid for participation, but museum visitors do not. Although studies online and in museums do not completely solve such a volunteer bias (visitors to a science museum are obviously interested in science, for example), they at least go a step toward diversifying the pool of participants.

Participants in the lab may also be particularly prone to experimenter demand characteristics. Recruiting participants online, therefore, has the additional advantage of anonymity, as pointed out by Bargh and McKenna (2004). Participants may be less inclined to try and figure out the "correct answer," or otherwise behave in a way they think will please the experimenter. Overall, recruiting participants online may improve the diversity and quality of the sample in a number of ways.

Along with access to larger and more diverse samples of participants, data collection can be much expedited if experimenters use these alternative locations. Recruiting and running individuals in a university setting is difficult and hampered by a number of factors, including the local population size (typically undergraduate psychology or linguistics students), university holidays, exam times, and so on. By moving out of the university setting both the researcher and the participant will be less disrupted. For example, Dufau *et al.* (2011) collected data from 4,157 participants in only four months using an experiment conducted on a smartphone. A comparably sized study conducted in a lab took almost three years (Balota *et al.*, 2007).

A further benefit is data is cheap. Cost per participant on Amazon Turk begins at one cent, with an additional fee to Amazon of 20% (https://requester.mturk.com/pricing). Costs are also reduced in terms of lab space, labor, and data entry (Birnbaum, 2004). Costs for experiments in museums are also lower. Participants typically volunteer for free. For them, participation is a fun and educational experience—another aspect of their museum visit.

Finally, research of this nature, particularly research conducted in museums, has additional benefits, for example, public engagement. By conducting research in a public setting one can promote a research program, institute, or university, and simultaneously educate the public about the research process and research findings.

What Does It Entail? Best Practice

In the last 5 years or so, research conducted online has expanded dramatically. With the development of crowdsourcing services such as MTurk and Crowdflower, or online experimental software such as WebExp, online research is easy. Many standard psycholinguistic studies involving visual and auditory stimuli, for example, pictures, words and sentences, are possible; and data can include ratings, written and spoken responses, and even reaction times. For example, Dufau *et al*. (2011) presented English words and nonwords and collected accuracy and response times for lexical decisions (i.e., "Is this a real English word?"). There are a number of standard templates on MTurk available, such as surveys and Likert scales, which can be easily adapted to suit the researcher's needs.

When building an online experiment there are a number of things to keep in mind. It is important to ensure all variables of interest are identified and coded to allow efficient data processing and analysis. A mistake in variable labelling could lead to weeks of additional work once large volumes of data have been collected. Since the participant will be completing the task away from the experimental lab, ways of reducing fatigue and sustaining motivation also need to be considered, such as a progress bar indicating the length of the study (Keuleers *et al*., 2015). Similarly, removing a "time-out" feature that ends the experiment after a period of inactivity means participants can take a break whenever they want and hence reduces the number of dropouts (Keuleers *et al*., 2015). However, as with all forms of experiments, the participant must be informed about their right to withdraw from participation at any point without consequence. After the experiment is completed, response times can be measured to assess concentration on the task. Participants with extremely long or short response times, or with large gaps during the experiment, probably were distracted or unmotivated, and so should be removed from analysis.

Online studies are now branching out into mobile devices, with a number of experiment applications ("apps") emerging. Smartphones are a fundamental feature of many people's daily lives and offer a great opportunity for research, with high spatial and temporal resolution making them appropriate for experiment presentation (Dufau *et al*., 2011). One example is the app "SynQuiz" designed by the research consortium Language in Interaction (2015). It is quick and easy to download and use, and presents participants with a number of fun tasks to test whether an individual has grapheme-color synesthesia (where individuals automatically and involuntarily experience color sensations to letters or numbers). The Language in Interaction consortium has also developed "WoordWolk," an app designed to aid aphasia patients with word finding, and "LingQuest," a game to educate players about the world's languages, and so are also applying and disseminating research.

Researchers have also been availing themselves of opportunities to run studies in museums (e.g., Simner *et al*., 2006), and other public events such as science festivals (e.g., Verhoef, Roberts, & Dingemanse, 2015). A research study in a museum will typically involve a museum residence for a period of time (i.e., days or weeks), but it is also possible to have short data collection sessions, such as at a special event or a museum "Late night" opening. Visitors to museums include individuals of all ages and backgrounds, so it is imperative this wide population is kept in mind and instructions are written in a clear and comprehensible manner. The experiment itself should be fun and educational. It is important participants leave the museum feeling happy

and other visitors feel encouraged to participate. For the same reason, experimental tasks should not be too long or difficult. Naturally, museums can be noisy and unexpected things occur, so keep a record of any such extraneous factors to take into account during analysis.

Disadvantages, Problems, and Pitfalls

Despite the excitement surrounding online studies and the potential for rapid data collection of large and diverse samples, there are, of course, a number of disadvantages to take into consideration. There are three main classes of problems centring around participants, the amount of control the experimenter has over the situation, and the types of studies that can be conducted.

First, although moving experiments online has the potential to increase the diversity of the participant pool, experimenters must be careful to understand the limitations of this type of sampling too. People with access to internet technology are part of an increasingly homogenized globalization culture, dominated by Western consumer values. They are likely infected by English too. So although participants may come from diverse nations, they may not reflect the cultural or linguistic diversity the researcher hopes to tap. Knowing the relevant demographic facts about the participants is important for interpreting any results.

Second, although researchers may carefully compose instructions, there will no doubt be room for misinterpretation and confusion. An online participant cannot ask clarification questions if something is unclear. So, there is no guarantee the instructions will be followed as carefully as they would be in a lab where the researcher is on hand to ensure comprehension. At the same time the experimenter has little control over who is participating in the study. The same people can take part in a study multiple times under different usernames (although this could be avoided by allowing participation from an IP address only once). Participants who do not meet the study's requirement can also sign up to a study (e.g., being a native speaker of a language), or they can "cheat" by working on an experiment collaboratively, for example. At the same time, the dropout rate may be higher than for studies conducted in person because there is no immediate social consequence, or simply because other events intervene for the participant.

This leads to a related issue—that is, the extent to which controlled experimental conditions are observed. In a lab, experimental cubicles are soundproofed and bare, with minimal distraction, so full attention is given to the task. Completing an experiment at home, on the other hand, instead lends itself to distraction. There may be music or a television playing in the background, telephone calls, children demanding attention, etc. The researcher has no control over this. Similarly, in a museum or other public space, participants are there to enjoy themselves, so they might not adhere to experimental conditions as would a paid participant in a university. On the other hand, "real-world variability" could be seen as an advantage because it simulates conditions closer to the way we naturally process language every day (Moroney, 2003). Interestingly, Enochson and Culbertson (2005) compared response time data collected online to an identical task in the lab, and found greater variability in the data from the lab (larger standard error). So perhaps people online are not as prone to succumbing to distractions as one might fear.

A corollary to the lack of control over the environment is a lack of control over the equipment used in online experiments. Different computers, different operating systems,

and different internet servers can add variance to the timing of both experimental stimuli presentation and participant reaction times. In psycholinguistics many robust phenomena, such as semantic priming, are observed in small but significant differences in reaction times, so any additional variance in the data could wash out effects. Enochson and Culbertson (2015), however, have replicated three classic psycholinguistic effects with small differences in reaction times using MTurk: faster processing of pronouns compared to determiner phrases, processing costs for filler-gap dependencies, and agreement attraction, when a verb spuriously agrees with a nearby noun, instead of its grammatical subject. Moreover, Germine *et al.* (2012) compared the quality of data (i.e., mean performance, variance, and internal reliability) collected from online studies with typical lab experiments, and found negligible differences.

Finally, in addition to the issues above, there are limits to the types of studies that can be conducted online or in public places. Experiments requiring behaviors more complex than pushing buttons on a keyboard, or which require non-visual or auditory stimuli (e.g., odors), are not possible online. Studies taking place in museums are constrained in terms of time and difficulty, as museum visitors are primarily there to have fun and learn.

Exemplary Studies

One of the largest online studies to date was conducted by Keuleers *et al.* (2015). Nearly 300,000 participants took part in an adapted lexical decision test online, in which participants had to judge if letter strings were real words or not, producing accuracy and response time data for tens of thousands of words. Data from such a large number of participants allowed the researchers to more reliably estimate variability in language processing in the general (Dutch speaking) population. Additionally, it provided the opportunity to investigate effects of age, education, multilingualism, and location on vocabulary size. This study also serves as a good example of public engagement. After completing the test, participants could share their scores on social media, which, the researchers believed, led to increased participation rates and participant satisfaction. Furthermore, participants could go back to their responses given in the lexical decision task and look up word meanings in an online dictionary. The educational aspect was not one-way either. Participants had the opportunity to comment on items used in the task, so experimenters were informed about a number of nonwords being too similar to real words.

To build such an online experiment, one can use a program like WebExp (Keller, Gunasekhran, Mayo, & Corley, 2009). WebExp utilizes a server that hosts experimental stimuli and results, and a connected client applet that runs in the browser of the participant. An experiment is written in XML, a programming language familiar to users of HTML, and requires a timeline describing the stages of the experiment (e.g., introduction, practice). Further specified in each stage are individual slides and components such as text, image, buttons, each with defined properties. Data such as button press and timing information can be recorded and stored on the server using numbered files in a data directory.

An excellent example highlighting the advantages of conducting studies in museums is provided by Simner and colleagues (2006). In 3 months, 1,190 English-speaking visitors to London's Science museum took part in a computerized letter/number-to-color matching task in order to estimate the prevalence of grapheme-color synesthesia. The most significant finding from this research was a female to

male ratio of synesthesia of 0.9:1. Previous studies had estimated a much higher ratio of 6:1. Collecting data from a wider pool of participants (museum visitors of many ages instead of just university students) provided results against the strongly held belief of a greater prevalence of synesthesia in females. The research suggested previous estimates reflected a study bias in which males are much less likely than females to come forward and report their synesthetic experience.

Conducting Studies in Real-World Settings

Rationale

Traditional studies within psycholinguistics tend to take a "narrow" view of language (Port, 2010), focusing on speech or written text, while leaving out rich contextual features—such as the physical context, the discourse context, and the social context; as well as other features of communication, such as hand and body gestures and facial expressions. Since much of the psychology of language has only focused on a constrained portion of communication this raises the question to what extent psycholinguistic findings reflect the way language is actually used by people.

Studies conducted in more "real-world" settings—that is, situations more closely reflecting how language is used in daily life—can be a step toward addressing the problem of ecological validity. This has also been described as the "scaling problem" (Zwaan, 2014): do results from psycholinguistic studies "scale up" to the real world? The study of natural language use has typically been side-stepped in traditional psycholinguistics most likely because of the difficulty involved in studying language in its fully embedded and multimodal context.

Traditional psycholinguistic experiments are conducted in controlled settings with real-world factors removed or radically simplified so variables of interest can be carefully manipulated. They take place in soundproofed laboratory cubicles. The participant is encouraged to focus solely on the language task at hand. The linguistic stimuli are often presented context free. Responding to decontextualized single words presented in the centre of a computer screen, or reading a single sentence about an unknown agent in an unknown situation is arguably a different matter than speaking and understanding in everyday life.

Language use in daily life is accompanied by a wealth of context. Consider chatting to your family over dinner, talking to friends as you take a stroll, or catching up with a cousin after a long separation. Speakers have common ground with their interlocutors. There are people-centred—rather than experimenter-driven—motivations and intentions for comprehending and producing language. There are contextual factors at play from multiple modalities.

In addition to external context—such as objects in the environment or ongoing activity—additional aspects of the communicative signal are often neglected in psycholinguistic studies and theories too. When talking, speakers use hand and body gestures, for example, via iconic gestures or by using beat gestures as a prosodic cue (e.g., McNeill, 1992). Research has shown speech and gesture are an "integrated system" (Kelly, Özyürek, & Maris, 2010): gestures congruent with speech (e.g.,

cutting gesture with "*chop*") facilitate speech comprehension compared to gestures incongruent with speech (e.g., twisting gesture with "*chop*"). With the advent of the embodied cognition paradigm (e.g., Barsalou, 1999), researchers are now also investigating how external factors in the communicative situation, such as the body and ongoing actions, affect the comprehension and production of language (for a review see Fischer & Zwaan, 2008). This highlights the potential impact of real-world body movement on language comprehension.

What Does It Entail? Best Practice

To reduce the artificiality of experimental manipulations and increase the ecological validity of results, researchers can use real-world situations to assess how various factors affect language processing. The concern for ecological validity is by no means new. One of the first examples of a psycholinguistic experiment conducted in a natural setting is by Clark (1979). In order to investigate responses to indirect requests, across five experiments a researcher telephoned 950 local businesses and asked simple direct and indirect questions such as "Could you tell me the time you close tonight?", and recorded the responses given. Based on the results, Clark outlined six sources of information addressees use to determine whether indirect questions should be interpreted in the literal form or not.

Today, researchers are beginning to record lengthy periods of real-world interaction. There are now recording devices children can wear all day so recordings of the child's utterances, and of those around her, can be collected and automatically analyzed when connected to specialized computer software (e.g., Kimbrough Oller, 2010). Similarly, children can wear lightweight head-cameras that enable researchers to see the world through a child's eyes and assess the role of real-world features on language acquisition (Smith, Yu, Yoshida, & Fausey, 2014).

Experiments conducted in real-world situations can be difficult and potentially problematic. So another way forward is to bring richer contextual cues into the lab. Experiments could investigate speech processing with simultaneous gestures or facial expressions, language comprehension whilst completing manual tasks or other forms of ongoing action such as by using a virtual reality environment (see Chapter 9, this volume), or conversations among friends with topics relevant to the individuals.

Disadvantages, Problems, and Pitfalls

Many of the disadvantages of conducting studies outside the lab reflect the trade-off between ecological validity and experimental control. In addition, there are specific ethical issues raised.

First, let's consider the lack of experimental control. Having a fairly context-free setting for an experiment enables the researcher to identify the effect of an experimental manipulation with more certainty. Within the real world, it is difficult to ensure the experimental manipulation occurred under the same conditions at all times. In an external context precise measurements are more difficult, which can be problematic for certain psycholinguistic phenomena that occur in the order of milliseconds. Real-world environments are noisy and so the range of psycholinguistic phenomena amenable to rigorous testing in this context may be limited.

A more practical consideration concerns problems recording data with specialist equipment. Many experimental methods now popular in psycholinguistics, such as EEG, eye tracking, and fMRI are difficult, if not impossible, to use outside of the typical laboratory purely due to the requirements of the equipment. However, recent developments have overcome some of these problems—such as mobile eye-trackers in wireless glasses (www.smivision.com). In addition, including records of the non-linguistic situational context can be expensive in terms of the time required to analyze and code such features (particularly if in video format), and also disruptive if video equipment needs to be installed into environments, such as people's homes (Roy, 2009). However, methods to reduce such costs are being developed; for example, the development of fast and accurate speech and video transcription and annotation (Roy, 2009) or virtual reality systems (see Chapter 9, this volume).

Second, we turn to the ethics of conducting experiments under more naturalistic contexts. When conducting studies in a university, research proposals have to be carefully reviewed by an ethics committee to monitor for likely risks, and make sure sufficient information is given to participants. By conducting an experiment outside of the lab, the researcher cannot anticipate all potential problems and risks. In addition, some studies may rely on the participant not knowing they are part of an experimental manipulation, since knowing you are in an experiment may make you behave differently. This means participants lose the opportunity to give informed consent. However, ethical guidelines set out by the American Psychological Association indicate it is acceptable to dispense with informed consent provided certain conditions are met, such as there being no risk of harm or distress to the participant, and participant confidentiality being protected (http://www.apa.org/ethics/code/). In sum, researchers must respect participants' freedom and privacy, and take care not to disrupt people's daily lives.

Since studies completed in real-world environments can contain a large amount of variance and potential confounding factors, researchers must take careful and thorough records of events. Overall, it is probably still the case that any phenomena will have to be investigated using multiple methodologies (i.e., in typical experimental settings and in ecologically valid settings). Such data can be used to provide converging evidence for specific psycholinguistic phenomena.

Exemplary Studies

Boroditsky and Ramscar (2009) present a good example of a study conducted in an everyday situation with rich context. The researchers wished to address the effect of spatial position on the conceptualization of time, so they took advantage of real-world situations that could serve as experimental manipulations. For example, individuals in an airport who were waiting to depart or who had just arrived were asked the question "*Next Wednesday's meeting has been moved forward two days. What day is the meeting now that is has been rescheduled?*". The extent to which people took an ego-moving perspective (thinking of themselves moving through time and thus answering "*Friday*") or a time-moving perspective (thinking of time moving toward them and thus answering "*Monday*") was affected by their real-world spatial experience: people who had just arrived on a flight were more likely to take the ego-moving perspective (and answer *Friday*) than those just about to depart.

Although more an observational study than an experiment, an impressive example of rich, ecologically valid data comes from Roy (2009). In the "Human Speechome Project" cameras were fitted in Roy's own home so a comprehensive recording of language acquisition in the natural context of Roy's son could be collected from birth to age three. This resulted in over 230,000 hours of recordings. From the recordings numerous features could be extracted using human-machine transcription and annotation systems, such as words, prosodic features, and speaker identification from the audio; and person/object information, actions, and manner of actions from video. After processing this perceptual information, it can be fed into a machine learner that computationally models and predicts the language acquisition process. Initial findings from this rich data suggest the importance of the caregiver in language acquisition. For example, the first reliable utterance of a new word by the child occurred once the caregiver had reduced the complexity of utterances containing that word. There are many further possibilities for the Speechome project, taking into account semantic and pragmatic contexts and assessing the role of eye-gaze and body movements on production, for example. Overall the project reveals how children learn to understand the meaning of words within meaningful contexts.

Conclusions

The lab experiment remains a crucial home for psycholinguistics. But there are a number of factors which together call for a greater participation of a wider-selection of people, and a more contextualized notion of language. An informed choice of methods, weighing up the advantages and pitfalls specific to each of them, offers remedy to some of the problems haunting psycholinguistic research. After all, our theories should apply to all of humanity, and all of language use in its rich and varied guises. It's time for psycholinguists to venture out of the lab.

Acknowledgments

All authors are supported by The Netherlands Organization for Scientific Research: NWO VICI grant "Human olfaction at the intersection of language, culture and biology." We thank Ilja Croijmans, Josje de Valk, Elisabeth Norcliffe and Sebastian Sauppe for comments on an earlier draft.

Key Terms

Crowdsourcing The process of collecting responses from large groups of people in an online community.

Ecological validity The extent to which research findings can be generalized to real-world settings.

External validity The extent to which research findings can be generalized to other populations and situations.

Linguistic fieldwork Collection of primary language data outside of a workplace setting, typically associated with long-term investigation of lesser-known and under-described languages.

Linguistic relativity The hypothesis, associated most strongly with Benjamin Lee Whorf and Edward Sapir, which proposes that language can affect the way reality is viewed by its speakers.

Standard Average European (SAE) A term used with reference to modern Indo-European languages of Europe to highlight similarities in their linguistic features.

References

Arnett, J. J. (2008). The neglected 95%: Why American psychology needs to become less American. *American Psychologist*, *63*, 602–614. doi: 10.1037/0003-066X.63.7.602

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, *39*, 445–459.

Bargh, J. A., & McKenna, K. Y. A. (2004). The internet and social life. *Annual Review of Psychology*, *55*, 573–590. doi:10.1146/annurev.psych.55.090902.141922.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*, 577–660. doi: 10.1017/S0140525X99532147

Birnbaum, M. H. (2004). Human research and data collection via the internet. *Annual Review of Psychology*, *55*, 803–832. doi:10.1146/annurev.psych.55.090902.141601.

Boroditsky, L., & Ramscar, M. (2002). The roles of body and mind in abstract thought. *Psychological Science*, *13*, 185–189. doi: 10.1111/1467-9280.00434

Bowerman, M., & Brown, P. (Eds.). (2008). *Crosslinguistic perspectives on argument structure*. New York: Lawrence Erlbaum Associates.

Bowern, C. (2008). *Linguistic fieldwork: A practical guide*. Basingstoke: Palgrave Macmillan.

Burenhult, N., & Majid, A. (2011). Olfaction in Aslian ideology and language. *The Senses & Society*, *6*, 19–29. doi: 10.2752/174589311X12893982233597

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*, 3–5. http://doi.org/10.1177/1745691610393980

Clark, H. (1979) Responding to indirect speech acts. *Cognitive psychology*, *11*(4), 430–477.

Crowley, T. (2007). *Field linguistics: A beginner's guide*. New York: Oxford University Press.

Dahl, Ö. (2015). *How WEIRD are WALS languages?* Presented at the Diversity Linguistics: Retrospect and Prospect, Leipzig.

Dufau, S., Duñabeitia, J. A., Moret-Tatay, C., McGonigal, A., Peeters, D., Alario, F.-X., … Grainger, J. (2011). Smart phone, smart science: How the use of smartphones can revolutionize research in cognitive science. *PLoS ONE*, *6*, e24974. doi:10.1371/journal.pone.0024974.

Enochson, K., & Culbertson, J. (2015). Collecting psycholinguistic response time data using Amazon Mechanical Turk. *PLOS ONE*, *10*, e0116946. doi:10.1371/journal.pone.0116946.

Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, *32*, 429–448. doi:10.1017/S0140525X0999094X.

Fischer, M. H., & Zwaan, R. A. (2008). Embodied language: A review of the role of the motor system in language comprehension. *The Quarterly Journal of Experimental Psychology*, *61*, 825–850. doi:10.1080/17470210701623605.

Friederici, A. D., & Levelt, W. J. M. (1990). Spatial reference in weightlessness: Perceptual factors and mental representations. *Perception & Psychophysics*, *47*, 253–266. doi:10.3758/BF03205000.

Ganguli, M., Lee, C.-W., Hughes, T., Snitz, B. E., Jakubcak, J., Duara, R., & Chang, C.-C. H. (2014). Who wants a free brain scan? Assessing and correcting for recruitment biases in a population-based fMRI pilot study. *Brain Imaging and Behavior*, *9*, 204–212. doi:10.1007/s11682-014-9297-9.

Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., & Wilmer, J. B. (2012). Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review*, *19*, 847–857. doi:10.3758/s13423-012-0296-9.

Gick, B. (2002). The use of ultrasound for linguistic phonetic fieldwork. *Journal of the International Phonetic Association*, *32*, 113–121. doi:10.1017/S0025100302001007.

Gil, D. (2001). Escaping Eurocentrism: Fieldwork as a process of unlearning. In P. Newman & M. S. Ratliff (Eds.), *Linguistic fieldwork* (pp. 102–132). Cambridge, UK ; New York, NY: Cambridge University Press.

Gippert, J., Himmelmann, N., & Mosel, U. (Eds.). (2006). *Essentials of language documentation*. Berlin ; New York: Mouton de Gruyter.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world. *Behavioral and Brain Sciences*, *33*, 1–75. doi: 10.1017/S0140525X0999152X

Hyman, L. M. (2001). Fieldwork as a state of mind. In P. Newman & M. S. Ratliff (Eds.), *Linguistic fieldwork* (pp. 15–33). Cambridge, UK ; New York, NY: Cambridge University Press.

Jaeger, T. F., & Norcliffe, E. J. (2009). The cross-linguistic study of sentence production. *Language and Linguistics Compass*, *3*(4), 866–887.

Keller, F., Gunasekharan, S., Mayo, N., & Corley, M. (2009). Timing accuracy of web experiments: A case study using the WebExp software package. *Behavior Research Methods*, *41*, 1–12.

Kelly, S. D., Ozyurek, A., & Maris, E. (2010). Two sides of the same coin: Speech and gesture mutually interact to enhance comprehension. *Psychological Science*, *21*, 260–267. doi:10.1177/0956797609357327.

Keuleers, E., Stevens, M., Mandera, P., & Brysbaert, M. (2015). Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *The Quarterly Journal of Experimental Psychology*, *68*, 1665–1692. doi:10.1080/17470218.2015.1022560.

Kimbrough Oller, D. (2010). All-day recordings to investigate vocabulary development: A case study of a trilingual toddler. *Communication Disorders Quarterly*, *31*, 213–222. doi:10.1177/1525740109358628.

Language in Interaction Consortium (2015). LingQues t(1.1). [Mobile application software]. Retrieved from http://itunes.apple.com

Language in Interaction Consortium (2015). Syn Quiz (1.1.161). [Mobile application software]. Retrieved from http://itunes.apple.com

Language in Interaction Consortium (2015). WoordWolk (1.3). [Mobile application software]. Retrieved from http://itunes.apple.com

Levelt, W. J. M. (2013). *A history of psycholinguistics: The pre-Chomskyan era*. Oxford: Oxford University Press.

Levinson, S. C. (2012). The original sin of cognitive science. *Topics in Cognitive Science*, *4*, 396–403. doi:10.1111/j.1756-8765.2012.01195.x.

Majid, A., Boster, J. S., & Bowerman, M. (2008). The cross-linguistic categorization of everyday events: A study of cutting and breaking. *Cognition*, *109*, 235–250. doi: 10.1016/j.cognition.2008.08.009

Majid, A., Bowerman, M., Kita, S., Haun, D. B. M., & Levinson, S. C. (2004). Can language restructure cognition? The case for space. *Trends in Cognitive Sciences*, *8*, 108–114. doi:10.1016/j.tics.2004.01.003.

Majid, A., Bowerman, M., Staden, M. van, & Boster, J. S. (2007). The semantic categories of cutting and breaking events: A crosslinguistic perspective. *Cognitive Linguistics*, *18*, 133–152. doi: 10.1515/COG.2007.005

Majid, A., & Levinson, S. C. (2010). WEIRD languages have misled us, too. *Behavioral and Brain Sciences*, *33*, 103. doi: 10.1017/S0140525X1000018X

Malt, B. C., & Majid, A. (2013). How thought is mapped into words. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*, 583–597. doi:10.1002/wcs.1251.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago: University of Chicago Press.

Moroney, N. (2003). Unconstrained web-based color naming experiment. In *Electronic Imaging 2003* (pp. 36–46). International Society for Optics and Photonics. doi:10.1117/12.472013

Newman, P., & Ratliff, M. S. (Eds.). (2001). *Linguistic fieldwork*. Cambridge: Cambridge University Press.

Norcliffe, E., Harris, A. C., & Jaeger, T. F. (2015). Cross-linguistic psycholinguistics and its critical role in theory development: Early beginnings and recent advances. *Language, Cognition and Neuroscience*, *30*, 1009–1032. doi: 10.1080/23273798.2015.1080373

Norcliffe, E., Konopka, A. E., Brown, P., & Levinson, S. C. (2015). Word order affects the time course of sentence formulation in Tzeltal. *Language, Cognition and Neuroscience*, *30*, 1187–1208. doi:10.1080/23273798.2015.1006238.

Port, R. F. (2010). Language as a social institution: Why phonemes and words do not live in the brain. *Ecological Psychology*, *22*, 304–326. http://doi.org/10.1080/10407413.2010.517122

Roy, D. (2009). New horizons in the study of child language acquisition. *Proceedings of Interspeech 2009*, 13–20.

Sakel, J., & Everett, D. L. (2012). *Linguistic fieldwork: A student guide*. Cambridge ; New York: Cambridge University Press.

Sauppe, S., Norcliffe, E., Konopka, A. E., Van Valin, R. D., & Levinson, S. C. (2013). Dependencies first: Eye tracking evidence from sentence production in Tagalog. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Meeting of the Cognitive Science Society(CogSci 2013)* (pp. 1265–1270). Austin, TX: Cognitive Science Society.

Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology*, *51*, 515. doi: 10.1037/0022-3514.51.3.515

Simner, J., Mulvenna, C., Sagiv, N., Tsakanikos, E., Witherby, S. A., Fraser, C., Scott, K., & Ward, J. (2006). Synaesthesia: The prevalence of atypical cross-modal experiences. *Perception*, *35*, 1024. doi: 10.1068/p5469

Slobin, D. I. (Ed.). (1985). *The crosslinguistic study of language acquisition*. Hillsdale, N.J: L. Erlbaum Associates.

Smith, L. B., Yu, C., Yoshida, H., & Fausey, C. M. (2015). Contributions of head-mounted cameras to studying the visual environments of infants and young children. *Journal of Cognition and Development*, *16*, 407–419. doi:10.1080/15248372.2014.933430.

Thieberger, N. (Ed.). (2011). *The Oxford handbook of linguistic fieldwork*. Oxford: Oxford University Press.

Verhoef, T., Roberts, S. G., & Dingemanse, M. (2015). Emergence of systematic iconicity: Transmission, interaction and analogy. In In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *The 37th annual meeting of the Cognitive Science Society (CogSci 2015)*. Cognitive Science Society.

Wagers, M, Borja, M. F., & Chung, S. (2015). The real-time comprehension of WH-dependencies in a WH-agreement language. *Language*, *91*, 109–144.

Whalen, D. H., & McDonough, J. (2015). Taking the laboratory into the field. *Annual Review of Linguistics*, *1*, 395–415. http://doi.org/10.1146/annurev-linguist-030514-124915

Whorf, B. L. (1944). The relation of habitual thought and behavior to language. *ETC: A Review of General Semantics*, 197–215.
Wnuk, E., & Majid, A. (2014). Revisiting the limits of language: The odor lexicon of Maniq. *Cognition*, *131*, 125–138. doi: j.cognition.2013.12.008
Wolff, P., & Holmes, K. J. (2011). Linguistic relativity. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*, 253–265. doi:10.1002/wcs.104.
Wundt, W. M. (1920). *Erlebtes und Erkanntes*. Stuttgart: A. Kröner.
Zwaan, R. A. (2014). Embodiment and language comprehension: Reframing the discussion. *Trends in Cognitive Sciences*, *18*, 229–234. doi:10.1016/j.tics.2014.02.008.

Further Reading and Resources

Comprehensive reference information for the world's languages, especially the lesser known languages:
http://glottolog.org/

Database of structural (phonological, grammatical, lexical) properties of languages gathered from descriptive materials:
http://wals.info/

Stimulus material for the elicitation of semantic categories by the Language and Cognition department at the Max Planck Institute for Psycholinguistics:
http://fieldmanuals.mpi.nl/

A comprehensive and practical guide to designing and conducting semantic elicitation studies:
Majid, A. (2012). A guide to stimulus-based elicitation for semantic categories. In N. Thieberger (Ed.), *The Oxford handbook of linguistic fieldwork* (pp. 54–71). New York: Oxford University Press.

A collection of useful databases of various linguistic measures from Ghent University. Includes software such as nonword generators, and data from online vocabulary tests:
crr.ugent.be/programs-data

Amazon's Mechanical Turk, an online crowdsourcing site that allows collection of data from a large number of participants, such as using questionnaires and experiments:
www.mturk.com

Home of WebExp, a system for conducting experiments on the internet and storing results:
http://groups.inf.ed.ac.uk/webexp/

Information on how to apply to conduct research in London's Science Museum:
http://www.sciencemuseum.org.uk/about_us/new_research_folder/livescience.aspx