Wagers, M., & S. Chung. Language processing experiments in the field.
To appear in J. Sprouse, Ed., Oxford Handbook of Experimental Syntax.

Language processing experiments in the field*

Matthew Wagers and Sandra Chung

Department of Linguistics, UC Santa Cruz

Abstract: This chapter discusses some of the circumstances and opportunities that arise when doing language processing experiments in the field. We focus on how resource and cultural challenges can shape the design of experiments. Because of the diversity of small-language communities, we avoid generalizing very broadly and instead draw principally on our own experiences working with the Chamorro community in the Northern Marianas Islands.

Key words: sentence processing, fieldwork, small languages, Chamorro, Austronesian, self-paced listening, preferential looking, touch-tracking

1.1 Challenges and opportunities in field-based experiments

In principle, there are no differences between an experiment in the lab and an experiment in the field: both ought to rely on theoretically-driven, precise hypotheses, and predictions; good design, adequate sample sizes, and appropriate analysis; and ethical treatment of human subjects. In practice, there are many more obstacles in the field than in the lab to meeting the standard of a valuable experiment. There are, at the same time, many more opportunities and room for discovery. Our goal in this chapter is to respond to these practicalities, and signpost both the obstacles and the opportunities, based on our own experience conducting psycholinguistic studies on Chamorro over six years in the U.S. Commonwealth of the Northern Mariana Islands (CNMI). Our focus will be on what are sometimes called 'small language' communities – language communities which have relatively few speakers and which typically lack socio-economic and political power. The majority of the languages of the world are small languages.

Experimental linguistics is typically carried out in the context of a university laboratory or, increasingly, online via social media or crowd sourcing platforms. Researchers need to compile datasets of considerable size to allow them to draw reliable conclusions, with several scores of items and participants usually the norm[1]. In the physical lab, sophisticated devices are often used, such as eye-tracking cameras or brain scanners. The availability of participants, and the security and infrastructure required for costly instrumentation, are two considerations that make it practical to conduct experimental research in a university laboratory or online. An obvious challenge in field-based experimental research is simply that it provides different set of practical circumstances to which researchers must adapt. Increasingly, researchers are overcoming these challenges. For example, technological improvements have given rise to more portable equipment which can be brought to the field; see Bennett, et al. (2017) for an example

of an ultrasound imaging investigation of Irish consonants, and Norcliffe, et al. (2015) for an example of an eye-tracking experiment on Tzeltal sentence planning. See Polinsky (this volume) for a comparison of field work and experimental linguistics that also provides a broader summary of recent, relevant research.

A less-often recognized challenge, but in our view often the most serious challenge, is that the experimental method we've inherited from decades of doing research in university laboratories is itself heavily culturally circumscribed. For a researcher planning a field-based experimental linguistics study, this may be the source of more surprises and obstacles than the practical difficulties that must be overcome. Anand, Chung, and Wagers (2011) describe some of the 'cultural felicity' conditions which can typically be met in the lab, but are not guaranteed in the field: the priority attached to test-taking; an individual's willingness to maintain exclusive focus on an unnatural, usually solitary, task; the expectation of accommodation to out-of-context language material (usually presented by a machine). There is a kind of social contract with the experimenter that only makes sense enmeshed in certain cultural standards of authority and evaluation (Rosenthal and Rosnow 2009). Participants at western-style universities volunteer to participate, are acculturated to test-taking environments, and often appear to be motivated to comply with the experimenter's expectations and wishes. In a linguistic community in the field, none of this can be taken for granted, nor should the university context necessarily be prized as providing better operating conditions (Henrich, Heine, and Norenzayan 2010).

The fieldwork tradition in linguistics, inherited from anthropology, provides a basis for developing a more culturally-sensitive ethos in experimental linguistics. However, it is also labor-intensive and typically centered around the partnership between the linguist and a single individual, or a small number of individuals. Thus it typically provides few answers, and

occasionally opposing presuppositions, to some of the bread-and-butter issues of the experimentalist, such as achieving adequate statistical power or disguising the intention of the experimental design. In our experience in conducting language processing experiments on Chamorro, the most effective way forward was to pursue a team-based approach, one which combined the expertise of an experimentalist (Wagers), a fieldworker (Chung), and, crucially, a member of the language community under investigation (Borja). By devolving some responsibilities and recombining others, this arrangement gave us a way to find realistic, site-specific solutions to the practical and cultural 'scalability' issues introduced above.

But there is ultimately no 'one-size-fits-all' approach to those issues because there's no 'one-size-fits-all' small language community. Small language communities are extraordinarily diverse – in population size, wealth, political structure, level of education and industrialization, and cultural and societal norms. We believe that this diversity makes it unproductive, at this point in time, to try to generalize about successful strategies for conducting language processing experiments in the field. The line of research is still too new for anyone to be in a position to enumerate in detail either its challenges or the best practices that would respond to them. So we will limit ourselves to talking about, and talking through, our own experience with Chamorro, an Austronesian language of the Mariana Islands (in Micronesia).

From 2011 to 2016 we conducted seven experiments on the processing of Chamorro in the three inhabited islands of the U.S. Commonwealth of the Northern Mariana Islands (henceforth CNMI): six comprehension experiments, each of which involved 80-120 participants, and one production experiment, which involved 43 participants. Our aim was to achieve experimental results that provided meaningful information about the time course of language processing, including reaction times on a par with those typically seen in experiments

conducted in western-style universities. To do this, we had to negotiate the many issues that arose working with a broad community of speakers in a different social-cultural context. Our expectations were violated on numerous occasions; sometimes we managed to find a workable solution, other times, we did not. We describe here what worked, what didn't work, and our diagnosis of why.

If we have general advice to offer, it is this: the experimental linguist in the field must adopt an outlook that is at once holistic and minimalist – holistic in recognizing the interdependency between experimental and social constraints, and minimalist in understanding that complicated procedures or designs can be more easily up-ended on the ground.

1.2 Materials

Improved experimental design, better digital resources, more accurate measurement devices, more sophisticated analytic techniques: these are just some of the ways in which researchers are constantly innovating and, in doing so, sharpening the questions they can ask. The linguist in the laboratory inherits these cumulative efforts and typically makes small, incremental changes from project to project. Linguists in the field, who must port these cumulative efforts to a different social context, will often find themselves innovating on several fronts at once. The challenges that spur these innovations stem from how a particular experimental design should be moored into the social context of the community whose language is being investigated. We will discuss two ways in which these challenges manifest themselves, in terms of *resources* and *function*:

• Do the *resources* exist to implement the experimental design in a way that is culturally relevant or legible to the community of interest? If not, what has to be created?

• Are there any aspects of the task and its *function* which are not consistent with community values, or which otherwise conflict with community presuppositions? If so, how can the task be adapted?

For the sake of concreteness, let us consider one specific experiment and explore the ways in which it could be adapted to the field. The experiment we will model is Sussman & Sedivy (2003), a visual world eye-tracking experiment that traced the time-course of filler-gap dependency formation. The goal of this experiment was to test whether speakers entertained incremental interpretations of English wh-questions (e.g. 'What did Jody squash the spider with?') before the linguistic input signaled any information about the location of the gap. Participants first heard a story and then had to answer a comprehension question. During this time, participants looked at a visual display and their gaze was monitored with a head-mounted eye-tracker. Figure 1 illustrates a sample item set, which consists of <*story, question, display*> combinations. The display depicts items related to the story. Crucially, it contains the core arguments of one critical proposition; here, that Jody (AGENT) squashed the spider (THEME) with a shoe (INSTRUMENT). As expected, Sussman & Sedivy found that when these arguments were mentioned, participants tended to look at their depictions. But they also found that when the verb was mentioned, the participants' looks anticipated the upcoming theme argument. This anticipation was amplified in wh-questions compared to polar questions (e.g. 'Did Jody squash the spider with her shoe?'). This finding – that participants were selectively 'eager' to look at the depiction of a particular unmentioned argument while processing a filler-gap dependency – converged with evidence from reading-time studies that filler-gap processing is especially predictive.

< insert Figure 1 here>

Several features of this study make it a compelling design for use in the field. Language processing behavior was measured in response to connected text as opposed to isolated sentences; the auditory modality was used; and the crucial measurement, probability of fixating a particular image, could be made before participants had to give the response that was ostensibly desired (an answer to the question). What, then, are the challenges?

Firstly, there are *resource challenges* around compiling the physical and digital materials required to carry out a particular design. In the case of a visual-world study like Sussman & Sedivy (2003), the resources required are the stories themselves, the recordings of the stories, and the pictures comprising the visual world.

1.2.1 Visual materials

Many methods for investigating language processing involve pictures, such as the visual-world paradigm, naming, picture matching, etc. The laboratory best practice would be to use pictures that are normed in a task-appropriate way. For example, it is usually important that the pictures in a study elicit consistent labels across participants, and a great deal of work has been invested in collecting measurements that an experimenter might use either to identify a homogeneous set of candidate images or to model the variability among images (e.g. as part of a regression analysis). For one recent example, see Moreno-Martínez & Montoro (2012), who created a collection of 360 color images rated by 36 Spanish speakers along such dimensions as familiarity, typicality, and manipulability. There is now a plethora of high-quality sources of

normed images and related resources (e.g. https://www.cogsci.nl/stimulus-sets; last accessed September 30, 2017).

There were two problems with using images from the existing picture databases for our experimental research on Chamorro. Firstly and most obviously, the images were normed for languages other than the language we were investigating. Although Chamorro translations could be found for many of the specific names in these databases, there was no guarantee that the Chamorro translation would be the best name for the picture given. In principle, this problem is easy to fix. We could have taken illustrations from the databases, which are often controlled along non-linguistic dimensions as well, and elicited their names and other sorts of judgments from an appropriately-sized sample of Chamorro speakers. Practically, we did not want to do this. Chamorro is a small language, with a total of some 35,000-40,000 speakers in the CNMI and Guam combined. We believed that locating participants to norm the illustrations would effectively sap, for the period of time we were in the field, the limited pool of Chamorro speakers who would be willing to participate in the main study

Secondly, the existing images were generally not culturally relevant. They did not depict people with the appearance or clothing typically seen in the CNMI; they did not show the kinds of flora or fauna found there; and they did not show common culturally-specific situations and events in Chamorro life. Naturally, many Chamorros have been exposed to mainland U.S. culture through the internet, television, and life abroad, but we wanted our experiments on the Chamorro language to be engaging in Chamorro terms. So we elicited our linguistic stimuli in Chamorro first, in order to decide what needed to be depicted.

For example, we wanted to use the verb *ngingi'*, which refers to sniffing or kissing the back of the hand – a traditional sign of respect when one encounters a Chamorro elder in a social

situation. We found a few photos on the internet of the *ngingi'*, and some illustrations in older printed matter on Saipan, but nothing that would suffice for our experiment. Likewise, we needed a drawing of a *sihik*, a species of Micronesian kingfisher (*Todiramphus cinnamominus*), as it appeared on the islands of Saipan and Rota. We could not find many images on the internet that looked just like the local birds; most depicted a differently-colored variety of kingfisher found in Guam. So, working with our Chamorro team member, we created a detailed 'Request for Proposals' to circulate to potential illustrators, in which we described how a *ngingi'* was performed, what the specific features were of kingfishers that our Chamorro team member observed flying around his home, etc. In some other cases we did find good internet resources we could offer as a guide, such as the Wikipedia entry for 'Saipan Jungle Fowl', or the excellent Guampedia (http://www.guampedia.com/).

Figure 2 shows some of the resulting illustrations, which were commissioned for different studies in our project. As we learned from debriefings, many participants were pleasantly surprised to encounter drawings that were locally and culturally specific.

<insert Figure 2 here>

At the same time, a few illustrations were problematic, and there were several instances in which drawing conventions familiar to us were not interpreted as we intended. For example, in the image of a rooster pecking the *sihik* (Figure 2, top right panel), small lines were used to indicate the impact of the pecking. These were not widely understood by our participants, particularly older Chamorros. It was instructive to learn how many of our presumptions could be frustrated, in ways we could not expect. For example, the very same illustration of the *sihik*

elicited a surprising response from several farmers, who claimed that roosters would never attack a kingfisher, although hens might. Most of the concerns about the adequacy of particular drawings were minor, but in a few cases they were serious enough to cause us to set the item aside.

1.2.2 Audio materials

In addition to pictures, an experiment in the field will often need audio recordings. This was a necessity in our own research; most Chamorros are not skilled readers of Chamorro, in part because there are several standard and nonstandard orthographies in use (Chung & Rechebei, 2014). Given the number of individual tokens required in many experimental designs – such as in the fully-crossed, within-subjects designs we used – recording the stimuli is one of the most arduous aspects of preparing for the experiment. We had limited time in the CNMI and thought it would greatly prolong the study to work with another speaker, so we once again made a trade-off and decided to use only our Chamorro team member's voice. There are definitely perils involved in making this choice. Because the experimenter knows the design of the experiment, s/he might read the materials in a way that is unintentionally informative about a desired mode of task performance. However, the choice also granted us some important flexibility, because it made it easier to re-record our stimuli on the fly. We took care in inspecting and editing our recordings. We measured several relevant acoustic cues and compared them across conditions, to ensure that our designs were varying what we intended; that there wasn't too much unintended variation across an item set; and that any unintended variation was random and not correlated unintentionally with condition. In many instances we re-recorded specific stimuli. And unlike an experiment in the lab, where a researcher may have access to a

sound booth or a reliably quiet room, we had to contend with background noise, although it was often from the natural world.

Whether our Chamorro team member used a substantially non-standard pronunciation was a concern, since that would be difficult for us to detect. We knew that there were two major dialects in Chamorro – one spoken in Rota (and the southern part of Guam), and the other – the majority dialect – spoken elsewhere in the Mariana Islands. These dialects are mutually intelligible, and the differences between them are mostly phonological. For example, the majority dialect distinguishes between geminate and nongeminate consonants, whereas Rotanese Chamorro has no geminates. However, the majority dialect is recognized as the standard, and there was no problem in using stimuli recorded in this dialect to collect data from speakers on Rota. As it turned out, lexical variation – tied not only to island, but also to age – was the greatest barrier to comprehension for our participants. We were less prepared for this kind of variation because it had not been described in any detail by other linguists. For example, the word for 'frog' is *kaheru'* on Rota but *kairu'* on Saipan (both forms are evidently borrowings from Japanese). Possibly because this sort of word wouldn't typically occur on a news broadcast, say, it was a point of variation of which almost all speakers were unaware. Relatedly, we found that younger speakers' lexical knowledge of names of animals was extremely limited.

A final resource challenge we had to solve involved composing the stimuli themselves. This was, in many ways, purely a fieldwork task. That is to say, we did not attempt to translate 'targets' from English to Chamorro ourselves, but instead generated the stimuli through elicitation and translation directly with our Chamorro team member. This was a crucial part of the design phase of the experiment. There is a mode of constructing stimuli for experiments that one might call the 'Mad Libs approach,' according to which item templates are designed and the

experimenter fills in slots in the templates with lexical material as if they were pigeon-holes. While this characterization is somewhat cartoonish, it is not an entirely inaccurate rendition of how many researchers design experiments in their own language when those experiments require large numbers of items. This was not feasible for our Chamorro experiments, because we often did not know *a priori* whether a particular factorial design we envisioned could be implemented generally. It would have been easy to be misled by the idiosyncrasies of a few lexical items. Every <*item, condition*> pair needed to be elicited singly to make sure it was acceptable in Chamorro. Not only that, we discovered frequently, but unsurprisingly, that lexical items had unintended connotations in some constructions.

There is a potential reward in attempting to generate large numbers of sentences from scratch with a native speaker. The act of searching for lexical items and trying out new combinations of them often led to unexpected ungrammaticality or unpredictable complexity. An exigency of experimental research, i.e., large numbers of items, can thus become an asset in the fieldwork context. In our case, we discovered a number of novel grammatical generalizations, including a complex constraint on wh-dependencies formed on the possessor (Wagers, Borja, and Chung 2015) and the optionality of wh-agreement in certain relative clauses (described in Wagers, Borja, and Chung 2018). Our Chamorro team member had strong but difficult-to-pinpoint intuitions about the infelicity of certain passives in prenominal relative clauses – intuitions that would be supported and amplified by high error rates in a comprehension task (Wagers, Borja, and Chung 2018).

Finally, we did occasionally present written materials to experimental participants. We conducted a word familiarity study as a pencil-and-paper survey with a small subset of participants in one of our first experiments (Wagers, Borja, and Chung 2015), as well as some

word order preference surveys. Because of the variation in reading skill we alluded to above, the use of written materials required careful administration and instruction. Often, we simply ended up reading the survey aloud and, in many interactions, the administration of the survey effectively became an elicitation session. The data we gleaned was valuable but acquired at a relatively steep cost.

1.3 Methods

1.3.1 Self-paced listening

The goal of our first two studies was to learn something about the incremental processing of wh-agreement, the special agreement found in Chamorro filler-gap dependencies that signals the grammatical relation of the gap. We were interested, in particular, in whether the information that wh-agreement provides about the gap is used by comprehenders to interpret wh-questions in advance of unambiguous bottom-up evidence for the gap site. Although we were inspired by Sussman & Sedivy (2003)'s visual-world paradigm experiment, we did not think we could marshal the required resources to mount that design. Instead we used an anomaly design, of the sort used by other researchers who have investigated wh-dependencies, such as Boland, et al. (1995) in their research on argument structure. We compared sentences in a design that crossed the plausibility of the filler as a direct object of the verb (here, *prensa,* 'iron') with the presence or absence of wh-agreement morphology, illustrated in (1-2) for overt wh-agreement only.

(1) *Plausible*

Kuåntu	**na chinina**	prinensåm-mu	nigap	gi talu'åni?
how.many?	shirts	WH[OBJ].iron-AGR	yesterday	in afternoon

'How many shirts did you iron __ yesterday afternoon?'

(2) *Implausible*

Kuåntu na patgun låhi	prinensåm-mu	nigap	gi talu'åni?	
how.many? child male	WH[OBJ].iron-AGR	yesterday	in afternoon	

'How many boys did you iron __ yesterday afternoon?'

In a reading time version of this design (Traxler and Pickering 1996; Wagers and Phillips 2014), the point at which enough information has been amassed to form a dependency between the filler and the gap is indexed by increased reading times for the implausible object conditions. We needed to adapt this to auditory presentation because of the high variability in Chamorro reading skill. One straightforward way to port a reading-time task into auditory presentation is the auditory moving window technique, also called *self-paced listening* (SPL), described first in detail by Ferreira, et al. (1996) (but cf. Pynte, 1978). Participants 'listen' to a sentence by pressing a button to iteratively advance through a series of segments which were spliced from whole sentence recordings.

Compared to reading-time studies, there are many fewer SPL studies in the adult psycholinguistics literature, and so fewer established findings to guide experiment design. SPL has been used more commonly to investigate populations where literacy is an issue, such as children (e.g. Bavin and Kidd 2007) or second language learners (Papadopoulou, Tsimpli, and Amvrazis 2014). Probably the most common concern for SPL is simply that it is an awkward way of listening to language. By segmenting a sentence and relying on participants' button presses, it introduces timing discontinuities in the acoustic signal that could distort or corrupt prosodic cues to lexical or syntactic processing. Indeed, Ferreira, Anes, and Horine (1996) showed that, in a task in which participants must leverage prosodic cues to disambiguate an otherwise globally ambiguous sentence, SPL is detrimental to performance compared to the

presentation of unsegmented recordings (though only somewhat). So, naturally, attention must be paid to the prosodic features of the phenomenon under investigation, and a judgment must be made about how likely it is that injection of noise into that process would lead to undesirable or misleading consequences.

How did participants react to this technique? The first time we used it, 7 of the 40 participants reported during the debriefing that they had had substantial difficulties with words being 'cut off' (*ha u'utut*) or the sound 'dying' (*måtai*). Another 6 reported problems with a small subset of words, but generally found that the listening technique became easier as the experiment progressed (*gi tutuhun kulan makkat, lao klumåklaru* 'at the beginning it was a little tricky, but it started to get clearer'). Finally, 27 reported little to no difficulty understanding. It is hard to directly interpret these numbers. Language processing experiments and debriefing sessions were novel experiences for virtually all of our participants. Even an ostensible perceptual report, like the words sounding 'cut off', could be the conflation of a number of factors, deriving not only from acoustic quality but also language experience and expectations about how they should respond in the debriefing.

Our experience suggests that, when all appropriate care is taken, SPL is a valuable technique for the experiment in the field. Yet we ended up only using it for two experiments, each with only 30-40 participants (the first experiment is reported in Wagers, Borja, & Chung, 2015). The reason for this was, essentially, a hunch that – for some speakers – it would either be too taxing, too uninteresting, or too unfamiliar. In our first two studies on wh-agreement, we worked with about 200 unique speakers, spanning ages from 19 to 81 (median age: 43). We only administered SPL to those speakers who seemed familiar with computers, were younger, or worked in an office.

1.3.2 Preferential looking.

For the rest of the participants, we needed a technique that would seem less onerous. The technique that we developed is a variant of inter-modal preferential looking (Golinkoff, et al. 1987) in which two different response categories were displayed on-screen while participants heard a sentence play over speakers (Wagers, Borja, and Chung 2015). The sentences followed the same anomaly design as in SPL, and the two response categories were simply 'Good' (*Måolik)* and 'Bad' (*Ti måolik*). We reasoned that participants would preferentially look at one of the response categories as evidence accumulated in its favor – in our case, coinciding with dependency formation and interpretation. We were aware that other researchers had used relatively simple technology – a hidden camera – to record and then code point of gaze to a manipulable, physical display using frame-by-frame annotation (Snedeker and Trueswell 2004). So we decided to simply pair our response collection software with a laptop-embedded webcam, and later have annotators align and code the webcam recordings with the simultaneously recorded audio. This was appealing, since we were wary that using an actual eye-tracking camera would be overly intrusive[2]. We aimed to be able to set up, and tear down, quickly and not take more than 15 minutes of anyone's time. We obtained explicit verbal consent to make the recordings. A handful of participants declined to be recorded (fewer than 5%), and we simply covered the camera with a sticker for those sessions.

Our data ultimately showed that our idea – that participants would selectively look at response categories as the sentence wore on – was only weakly supported. A better indicator was participants' looks away from the screen and down toward the keypad as they prepared to make a response. And while that measurement did end up being interpretable and consistent with the SPL data (Wagers, Borja, and Chung 2015, Figure 6), it came at a high cost. We had effectively

traded labor on the data collection end for labor on the analysis end. We trained several undergraduate RAs to do frame-by-frame annotation. Each video was multiply coded, and it was possible, even with this simple data, to achieve high inter-annotator agreement (comparable to Snedeker and Trueswell 2004, Appendix D). Unfortunately only 45 of 72 original videos (62.5%) were codeable. There were two main reasons for this. Perhaps as a consequence of our deliberately impromptu interactions, participants often felt free to look away for extended periods of time, chat with someone across the table, or generally not pay attention to the screen. In addition, there were many instances when we conducted the experiment at a participant's home or some other venue they had selected, and it was impossible to exercise adequate control over the illumination of the face.

Generally, our experience with the wh-agreement study was mixed. The Chamorro instructions delivered at the beginning of the experiment emphasized relaxed, brief, and non-judgmental interactions. In doing so, it seems probable that we simultaneously limited the scientific value of some of our participants' data (as evidenced by the attrition rate in our codeable videos), while also unexpectedly placing greater burdens on the 'cleaner' self-paced listening data drawn from the more demographically-biased sample. If those imperfect datasets had not pointed to the same conclusions, it is not clear we would have had anything to show for our efforts. On the other hand, our open recruitment standards brought in 112 participants, a number greater than we had imagined possible. 112 participants is a more-than-healthy sample for a lab-based psycholinguistics experiment, but, to put it in the perspective of the small language community, it represents at least 0.3% of the entire Chamorro-speaking population in the Mariana Islands (and nearly 3%, on the island of Rota)! The benefits of a large sample were not only statistical, but also social. In future experiments the percentage of participants who had

taken part in one of our previous experiments was usually a minority and ranged from 25%-60%. But many new participants 'had heard' about the experiment from others and were interested in joining in.

1.3.3 Tablets

After alternating between the SPL and modified preferential looking task for two experiments, we switched to a simpler – and ultimately more engaging – task: sentence-picture matching on a tablet computer. In a series of experiments on the comprehension of relative clauses (Borja, Chung, and Wagers 2016; Wagers, Borja, and Chung, 2018), we asked participants to select from one of two pictures that could depict an individual denoted by a relative clause. An example of a stimulus, translated into English, is 'Push the star over to the kingfisher that the rooster is pecking.' We would then depict two eventualities: one of a rooster pecking a kingfisher, and the other of a kingfisher pecking a rooster (see Figure 2). Previous researchers had used picture matching as an effective technique for studying relative clause parsing in populations for whom literacy could not be presupposed; see e.g. Caplan, Waters, & Hildebrandt (1997) for a study with adults who are aphasic, Clemens, et al. (2015) for speakers of Ch'ol and Q'anjoba'l, and Grüter (2005) for children who are second language learners of French or who have specific language impairment.

The use of the tablet computers opened the way to an innovation, namely, that we could collect data not only about what pictures participants selected but also *how* they selected it. We were inspired here by research using *mouse tracking* (Freeman and Ambady, 2010) as a stand-in for eye-tracking in the visual world paradigm. In mouse tracking, the inflection of the trajectory – how much it bends toward an alternative picture – has been used to gauge degree of

competition between two response alternatives (Freeman, Dale, and Farmer 2011). In our design[3], we asked participants to move around a small icon called the puck. The puck was initially situated near the bottom of the screen and needed to be moved to one of two pictures situated equidistant from it near the top of the screen. We were able to analyze not only when participants selected a picture, but also when they first touched the cursor; moreover, we could visualize and quantify the trajectory from initial to final position.

It had already been noted that swiping on a touch screen is much more 'ballistic' (Freeman and Ambady 2010) and thus there is less variability in the trajectories. Our research basically confirmed this observation, although we were able to find some clear competitive effects (reported primarily in Borja, Chung, and Wagers 2015). We also found that the point at which our participants initially touched the puck correlated strongly with their final selection time. The usefulness of this finding is that sentence-picture matching times are often quite long and variable – and this can severely limit the conclusions that can be drawn about incremental processing. For example, Clemens, et al. (2015) reports button-pressing times from sentence-picture matching experiments that range from an average of 6000 ms in an experiment with Russian speakers, to 3100 ms in an experiment with Ch'ol speakers, and 1200 ms in an experiment with Q'anjob'al speakers. Our Chamorro experiments which use initial touch times, instead of final selection times, routinely deliver results at the lower end of this spectrum, with correct answers ranging from 800 to 1600ms (medians across conditions). The distribution of these reaction times is potentially more plausibly linked to comprehension processes at the final word in the sentence than would be the case for higher RTs. More research is required to substantiate this claim in greater detail.

Setting aside the promise of collecting relatively short RTs, participants nearly uniformly found the tablets easy and intuitive to use and the task relatively pleasant to complete. Use of the tablets also opened up the opportunity for more substantive debriefings. In contrast to SPL or preferential looking, our participants could see potential applications for the tablet computers in Chamorro language teaching in the schools. This further dimension of engagement meant that our participants talked longer, and more concretely, about the task and the materials.

1.3.4 Debriefing

The debriefings turned out to be central to our experimental protocol. Participants were debriefed individually or—if several had finished the experimental task at the same time—in small groups, usually by our Chamorro team member together with one of us. The conversation began in Chamorro and usually continued in Chamorro, although some participants switched to English or alternated between the two languages. Some debriefing questions provided more information about the participant's fluency in Chamorro (e.g. 'Were there any words you didn't recognize?') or about which stimuli had worked or not worked (e.g. 'Were there any pictures that didn't make sense?'). Information about lexical variation is an example of something that routinely emerged in debriefings and something that we could incorporate into our analysis or future design. We found that when speakers could be encouraged to talk on concrete topics, especially whether they recognized particular words, they would often segue into more subtle observations about word order, say, or ambiguity.

Other debriefing questions invited participants to reflect on the experience of completing the task, what they thought its purpose was, and how it might be made more enjoyable (e.g. 'What did you think of the experiment?', 'Did you like it?', 'Would you take another experiment

like this?'). Still other questions were simply invitations to talk (e.g. 'Which picture did you like the most?'). Although some debriefings were perfunctory, others evolved into extended conversations about the state of the Chamorro language, the need to preserve it, the purpose of our research, and how some of our materials could be used in the schools. These conversations strengthened our personal relationships with community members and encouraged some of them to return to participate in our later studies.

1.4 In the field

Many of the same sorts of issues that arise in the design phase can also arise in the field, when the experiment is actually being conducted. In addition to finding the best way, for a given time and place, to recruit participants, the team must be able to resolve cultural issues that are uncovered only as the experiment is being conducted, interact with participants in ways that strike all parties as ethical and respectful, and encourage participants' interest in continuing to be involved in future research. One memorable illustration of this point comes from our first study, in which we had to re-program and re-record parts of the experiment on the fly, when we learned that an instruction to 'look at the cross in the center of the screen' could only be translated with the Chamorro word *kilu'us*. We quickly learned that the most prominent sense of this word, and its most immediate translation, was 'crucifix,' which elicited a strong reaction in the experimental context. In the end we simply rotated the graphic 45° and replaced the relevant word in the instruction with *ekkis* (the letter 'x').

1.4.1 Recruitment strategies

In our work in the CNMI we found that a 'one-size-fits-all' recruitment strategy would not work:

there had to be multiple recruitment strategies that emphasized personal connections and were tailored for the cultural setting. Given that the CNMI is a multilingual, multicultural society in which fluent Chamorro speakers form a minority of the population, it would not have worked to simply post a sign-up sheet at the local public library. Instead, our Chamorro team member used his extensive network of personal connections to make contact with potential participants on his home island. On the other two islands, he found a local Chamorro who agreed to contact potential participants in the same way. Over and above this, whenever we arrived on an island, we paid visits to local officials—members of congress, mayors, administrators, school principals—to talk about our project and ask them to encourage their Chamorro-speaking staff to participate. We were fortunate enough to be interviewed from time to time on local radio and television programs, and were able to use those interviews to announce (in Chamorro) our interest in recruiting Chamorro-speaking participants. Finally, we did not hesitate to turn random social encounters into on-the-spot invitations to participate in the experiment ('Oh—would you like to take the experiment? We could do it right now...').

Similarly, we found that there had to be multiple types of venues where the experiment could be conducted. It occasionally worked for us to conduct the experiment at a participant's home, but more often the venue was our Chamorro team member's home, or a workplace, public library, government office, restaurant, or other more neutral setting.

The inclusive character of Chamorro culture made it almost impossible to turn away potential participants, even those who were not native speakers of Chamorro. So we minimally screened participants by asking them three or four questions in Chamorro (e.g. 'How old are you?' 'How old were you when you began speaking Chamorro?'). Everyone who could answer the questions in Chamorro was invited to serve as a participant, and all data files were initially

included in the analysis. We set aside data files only on the basis of automatic criteria, such as average reaction times that were extremely long, or high error rates in answers to comprehension questions. Across several experimental studies we excluded, on average, 10% of participant data files.

1.4.2 Issues in delivering the experiment

The most sustained issue we confronted while conducting our experimental studies was that participants wanted to give their reactions as a group. Often when several participants were taking the experiment at the same time, they would want to consult with one another or compare their responses. In the debriefings, many participants said they enjoyed the experiment but would prefer a task they could collaborate on. We never managed to design a collaborative experimental task, although we devoted much thought to the issue. It may be that we were hampered by our specific research questions, which dealt with the comprehension of syntax and morphology, and that other designs could more fruitfully take advantage of the desire for group responses (for example, if the experimental study involved production; see, e.g., Brown-Schmidt and Konopka 2011). In several instances, we were able to debrief multiple participants together. That generated more concrete feedback and appeared to be gratifying for the participants.

A different issue concerned the trade-off between informed consent and disseminating information about the experiment's purpose. Before our research began, few if any Chamorros in the CNMI had participated in an experimental study. The IRB at our university agreed to waive the requirement for written informed consent on the grounds that the need to sign a consent form might frighten potential participants or discourage them from participating. We did, however,

obtain positive oral consent, and participants were informed of their right to discontinue participation at any time.

Participants nonetheless evinced some anxiety, and many expressed the belief that they were somehow being tested on their knowledge of Chamorro. Despite explicit statements to the contrary in the instructions, which were delivered in Chamorro (e.g. 'This is not a test. There are no right or wrong answers.'), this proved to be a difficult presupposition to defeat, and participants often asked for their score immediately afterwards. We suspect that several factors may have contributed to this presupposition. First and foremost, participants were often surprised to learn that a small language like Chamorro could be worthy of scientific study, and relatedly, that someone who had not studied their first language in school could be viewed as a competent speaker. This did not come as a surprise: one often heard Chamorro spoken of as a creole or a kind of corrupted version of Spanish (due to its rich lexical stratum of borrowings from the Spanish colonial era) or a language without a grammar.

In the third year of our experimental studies in the CNMI, it came to our attention that although participants did not want to sign a consent form, they wanted to be informed of the purpose of our research and to be assured that their participation was anonymous. The issue of anonymity arose partly because the version of preferential looking we had used involved videotaping not just the eyes but the entire face. Because of these privacy concerns and the labor-intensive character of the initial stages of the data analysis, we did not use this version of preferential looking in our later experiments. We addressed the more general concerns about our research by developing an information sheet for each experiment which we distributed to participants as part of the debriefing. The information sheet, which was written in Chamorro and English, gave a brief description in lay terms of the purpose of our research and the particular

experiment, stated that participation was anonymous, and provided contact information for the three researchers.

1.4.3 Sustaining the pool of participants

In experimental studies in Western-style universities, the pool of participants is typically regulated by a system that requires undergraduates to participate in order to complete particular courses or certain fields of study, or induces participation by offering extra course credit or sufficient financial compensation. In the field, no such system is in place. This means that an important task for a research team in the field is to figure out what causes community members to participate in an experiment and what would encourage them to continue to do so in the future. This is particularly important when the language has a small population of speakers and so the number of potential participants is intrinsically limited. It is made more challenging by the fact that cultures, societies, and communities clearly differ from one another along this dimension.

We were aware before we began our research that the financial compensation we could provide for our experiments would induce very few Chamorros to serve as participants. The money economy of the CNMI, together with the high cost of most goods, which are imported, made it impossible for us to consider paying community members at a rate that would justify their participation time. However, we also knew that flash drives were expensive, hard to obtain in the CNMI, prized by younger Chamorros, and much in demand. In our initial study we offered each participant a flash drive or else $10 as compensation. Flash drives, which were chosen by almost all participants, proved to be a great incentive, both because of their high storage capacity and because they were imprinted with the word ‘Chamorro.’ Phone cards were also successful.

But overall, flash drives were our most effective method of compensation, and we have returned to them again and again.

Over and above this, people agreed to participate for intangible reasons. The most important of these was their respect for our Chamorro team member, an educator and author who is known throughout the CNMI as a highly skilled, unusually generous community member who is committed to advancing indigenous languages and cultures. Many Chamorros who participated in more than one experiment were members of his extended family, his co-workers, his former students, or had collaborated with him in community endeavors. The novelty of participating in a psycholinguistics experiment was another draw. Some people participated because they wanted to help us, because they believed that our work would advance the study of the Chamorro language, or because they were curious to be part of an event that was conducted in Chamorro and involved outsiders. Finally, it was helpful that two members of our team are involved in a long-term, community-based effort to revise the Chamorro-English dictionary. This meant that dictionary group members were particularly willing to serve as participants and to help recruit other participants by spreading the word about our work.

It is harder to identify factors that discouraged people from serving as participants in our experiments. Length of time was clearly a potentially discouraging factor. During the instruction phase, participants were told the length of time that the task would probably take (10 to 20 minutes, depending on the experiment). While almost no one was deterred by this, some people commented in the debriefing that the task seemed long, or observed that the task took longer than it had in previous experiments. Our local contacts advised us not to tell participants during the instruction phase how many stimuli would be presented, on the grounds that this number (e.g. 40) would be viewed as a disincentive. In fact, very few individuals opted out at the instruction

phase, or began the experimental task but left before completing it. It was more common for individuals to show up at a testing site expecting to take the experiment, but leave when they learned there would be a 10-20 minute wait. Unsurprisingly, individuals were more willing to participate when they had been contacted by our Chamorro team member in advance and we could conduct the experiment in a setting they had chosen. The number, and engagement, of participants was more variable when the experiment was delivered in a more anonymous setting, such as a library or government office.

1.4.4 Community engagement

From the beginning we had planned to inform the community about the results of our research, and encourage their involvement, by giving public presentations on each of the three islands every few years. The idea was that these presentations would introduce the audience to the scientific study of language through Chamorro data that the community itself had provided. The first set of presentations, on community-based research on Chamorro, focused on the results of our first comprehension study and, separately, on the online parser and search engine that had been developed by Boris Harizanov for the revised Chamorro-English dictionary (Chung and Rechebei 2014). The discussion period touched briefly on psycholinguistics but then turned into a wide-ranging discussion of the need to preserve and maintain the Chamorro language. This audience response caused us to frame the second set of presentations to highlight what our research revealed about the changing nature of the Chamorro language. One unintended consequence of the inclusive approach to recruiting experimental participants was that our studies collected data from many different types of Chamorro speakers. Some interesting variation was revealed when the data were sorted by the participant's age or home island. For

instance, wh-questions in which the gap was a possessor were comprehended far more accurately by older generations than by younger generations; relative clauses in which the gap could be construed as a subject or an object were interpreted differently across islands. More surprising, to us, was the *lack* of age-related variation in the comprehension of certain types of relative clauses involving complex verb morphology (i.e. wh-agreement). Our second set of presentations described these findings and used them to point out that younger generations of Chamorro speakers know more about the language's complex verb morphology than they are usually given credit for. This set of presentations was well-attended and highly successful on one island and minimally attended on the others. On all three islands, the community's later interactions with us revealed that they found these presentations important even if they themselves had not been in the audience.

1.5 Conclusion

We purposely resist drawing too many conclusions from our particular experience working with the Chamorro community in Saipan, Tinian and Rota. However, if there's one lesson we think will have broad applicability, it is that the involvement of community members at different levels is indispensable. Language processing experiments are fundamentally unusual activities. In the context of a small language community, finding the right settings of cultural parameters will be a process of iterative discovery and adaptation. Rarely will it succeed to port an existing study directly into the language of interest without altering it. The need to involve native speakers as stakeholders as well as experimental participants flows from the fact that experiments have substantial practical requirements that cannot be met responsibly by a researcher who does not fully control the language. For example, experiments require the generation and fine-tuning of

large sets of high-quality materials which have the design features they are intended to have. In our case, it was even better that a native speaker had ownership over the project. Likewise, the need to recruit substantial numbers of participants meant developing a network of community members who had a positive, informed disposition toward what we hoped to accomplish. We had to become comfortable explaining and re-explaining the goals and outcomes of our project to sustain our presence in the community. In the end, this helped us not only to achieve greater focus on the scientific issues at stake and but also to find new ways to view the language.

References

Anand, Pranav, Chung, Sandra, & Matthew Wagers (2011). Widening the net: Challenges for gathering linguistic data in the digital age. In *NSF SBE 2020: Future research in the social, behavioral, & economic sciences*. http://www.nsf.gov/sbe/sbe_2020/submission_detail.cfm?upld_id=121.

Kidd, Evan, and Edith L. Bavin (2007). 'Lexical and referential influences on on-line spoken language comprehension: A comparison of adults and primary-school-age children', *First Language* 27: 29-52.

Bennett, Ryan, Ní Chiosáin, Máire, Padgett, Jaye, and Grant McGuire (2017). 'An ultrasound study of Connemara Irish palatalization and velarization', *Journal of the International Phonetic Association,* 1-44. doi:10.1017/S0025100317000494.

Boland, Julie E., Tanenhaus, Michael K., Garnsey, Susan M., and Greg N. Carlson (1995). 'Verb argument structure in parsing and interpretation: Evidence from wh-questions', *Journal of Memory and Language* 34*:* 774-806.

Borja, Manuel F., Chung, Sandra, and Matthew Wagers. (2015, January). Filler-gap order and online licensing of grammatical relations: Evidence from Chamorro. Paper presented at the Eighty-ninth Annual Meeting of the Linguistic Society of America, Portland, OR.

Borja, Manuel F., Chung, Sandra, and Matthew Wagers. (2016). 'Constituent order and parser control processes in Chamorro', in Amber Camp, Yuko Otsuka, Claire Stabile and Nozomi Tanaka (eds), *AFLA 21: The Proceedings of the 21st Meeting of the Austronesian Formal Linguistics Association*. Canberra: Asia-Pacific Linguistics, 15-32.

Brown-Schmidt, Sarah, and Agnieszka E. Konopka, A. E. (2011). 'Experimental approaches to referential domains and the on-line processing of referring expressions in unscripted conversations', *Information* 2*:* 302-326.

Caplan, David, Waters, Gloria S., and Nancy Hildebrandt. (1997). 'Determinants of sentence comprehension in aphasic patients in sentence-picture matching tasks,' *Journal of Speech, Language, and Hearing Research* 40: 542–555.

Chung, Sandra, and Elizabeth D. Rechebei (2014). 'Community Engagement in the Revised Chamorro-English Dictionary', *Dictionaries: Journal of the Dictionary Society of North America* 35: 308–317.

Clemens, Lauren Eby, Coon, Jessica, Pedro, Pedro Mateo, Morgan, Adam Milton, Polinsky, Maria, Tandet, Gabrielle, and Matthew Wagers (2015). 'Ergativity and the complexity of extraction: A view from Mayan', *Natural Language & Linguistic Theory* 33: 417–467.

Kidd, Evan, and Edith L. Bavin (2007). 'Lexical and referential influences on on-line spoken language comprehension: A comparison of adults and primary-school-age children', *First Language* 27: 29–52.

Ferreira, Fernanda, Anes, Michael D., and Matthew D. Horine (1996). 'Exploring the use of prosody during language using the auditory moving window technique', *Journal of Psycholinguistic Research* 25: 273–290.

Ferreira, Fernanda, Henderson, John M., Anes, Michael D., Weeks, Phillip A., Jr., and David K. McFarlane (1996). 'Effects of lexical frequency and syntactic complexity in spoken-language comprehension: Evidence from the auditory moving-window technique', *Journal of Experimental Psychology: Learning, Memory, and Cognition* 22: 324–355.

Freeman, Jonathan B., and Nalini Ambady (2010). 'MouseTracker: Software for studying real-time mental processing using a computer mouse-tracking method', *Behavior Research Methods* 42: 226-241.

Freeman, Jonathan B., Dale, Rick, and Thomas A. Farmer (2011). 'Hand in motion reveals mind in motion', *Frontiers in Psychology* 2. doi:10.3389/fpsyg.2011.00059.

Golinkoff, Roberta M., Hirsh-Pasek, Kathryn, Cauley, Kathleen M., and Laura Gordon (1987). 'The eyes have it: Lexical and syntactic comprehension in a new paradigm', *Journal of Child Language* 14: 23–45.

Grüter, Theres. (2005). 'Comprehension and production of French object clitics by child second language learners and children with specific language impairment', *Applied Psycholinguistics* 26: 363–391.

Henrich, Joseph, Heine, Steven J., and Ara Norenzayan (2010). 'Most people are not WEIRD', *Nature*, 466: 29.

Jäger, Lena A., Engelmann, Felix, and Shravan Vasishth (2017). 'Similarity-based interference in sentence comprehension: Literature review and Bayesian meta-analysis', *Journal of Memory and Language* 94: 316–339.

Mathôt, Sebastiaan, Schreij, Daniel, and Jan Theeuwes (2012). 'OpenSesame: An open-source, graphical experiment builder for the social sciences', *Behavior Research Methods* 44: 314–324.

Norcliffe, Elisabeth, Konopka, Agnieszka E., Brown, Penelope, and Stephen C. Levinson (2015). 'Word order affects the time course of sentence formulation in Tzeltal', *Language, Cognition and Neuroscience* 30: 1187–1208.

Moreno-Martínez, Francisco Javier, and Pedro R. Montoro (2012). 'An ecological alternative to Snodgrass & Vanderwart: 360 high quality colour images with norms for seven psycholinguistic variables', *PloS One* 7: e37527.

Papadopoulou, Despina, Tsimpli, Ianthi, and Nikos Amvrazis (2014). 'Self-paced listening', in Jill Jergerski and Bill Van Patten (eds.), *Research Methods in Second Language Psycholinguistics*. New York: Routledge, 50-68.

Pynte, Joel (1978). 'The intra-clausal syntactic processing of ambiguous sentences', in Willem J.M. Levelt and Giovanni B. Flores d'Arcais (eds.), *Studies in the perception of language*. New York: John Wiley & Sons, 109-127.

Rosenthal, Robert, and Ralph L. Rosnow (2009). *Artifacts in Behavioral Research.* Oxford: Oxford University Press.

Snedeker, Jesse, and John C. Trueswell (2004). 'The developing constraints on parsing decisions: The role of lexical-biases and referential scenes in child and adult sentence processing', *Cognitive Psychology,* 49: 238–299.

Sussman, Rachel S., and Julie Sedivy (2003). 'The time-course of processing syntactic dependencies: Evidence from eye movements', *Language and Cognitive Processes*, 18: 143-163.

Traxler, Matthew J., and Martin J. Pickering (1996). 'Plausibility and the processing of unbounded dependencies: An eye-tracking study', *Journal of Memory and Language* 35: 454-475.

Wagers, Matthew, Borja, Manuel F., and Sandra Chung (2015). 'The real-time comprehension of WH-dependencies in a WH-agreement language', *Language* 91: 109–144.

Wagers, Matthew, Borja, Manuel F., and Sandra Chung (2018). 'Grammatical licensing and relative clause parsing in a flexible word-order language', *Cognition* 178: 207-221.

Wagers, Matthew W., and Colin Phillips (2014). 'Going the distance: memory and control processes in active dependency construction', *The Quarterly Journal of Experimental Psychology* 67*:* 1274-1304.

Notes

* We are indebted to Manuel Flores Borja, the third member of our research team, for his many insights and his collaborative spirit. This work was supported in part by NSF Project BCS-1251429 to the University of California, Santa Cruz.

[1] Even in laboratory-based experiments, common sample sizes are regrettably not the optimum for many designs. See, e.g., Jäger, Engelmann and Vasishth, 2017, Appendix B.

[2] We stress that those desiderata reflect trade-offs we chose to commit to in the Chamorro milieu, which was grounded in our hope that we could keep coming around for future projects and keep recruiting participants. For other kinds of research questions, or in other kinds of communities, the use of an actual portable eye-tracker may very well make better sense (as in, e.g., Norcliffe, et al., 2015).

[3] We wrote custom software for this project in OpenSesame (Mathôt, Schreij and Theeuwes, 2012), an open-source experiment-building environment based in Python. It has an easy-to-use Android run-time module and we were able to collect all of our participants' interactions with the tablet.

Captions

WagersChung Figure 2: Some culturally-specific illustrations created for the Chamorro Psycholinguistics na Project. *Clockwise from top-left:* A doctor sniffs (*ngingi')* the hand of an elder, a traditional Chamorro sign of respect. A rooster pecks a Micronesian kingfisher (*sihik*). The Liberation Day queen (*raraina*) is photographed holding a trumpet shell (*kulu*'). A coconut grater (*kåmyu*) rests against a large water bottle – both are very common household items. These illustrations, which were created by California-based artist Nicole Goux, are available for anyone to download and freely use from our project website (http://chamorro.sites.ucsc.edu).