Language and Linguistics Compass 7/4 (2013): 221–232, 10.1111/lnc3.12021

Experimental Methods for Linguists

Sudha Arunachalam*
Speech, Language, & Hearing Sciences, Boston University

Abstract

Linguists are increasingly using experiments to provide insight into linguistic representations and linguistic processing. But linguists are rarely trained to think experimentally, and designing a carefully controlled study is not trivial. This paper provides a practical introduction to experiments. We examine issues in experimental design and survey several methodologies. The goal is to provide readers with some tools for understanding and evaluating the rapidly growing literature using experimental methods, as well as for beginning to design experiments in their own research. © 2013 The Author. Language and Linguistics Compass © 2013 Blackwell Publishing Ltd.

1. Introduction: Why Do Experiments?

Linguists are increasingly embracing experimental methods to augment traditional approaches such as grammaticality judgments. Well-designed experiments can offer insight beyond that available from grammaticality judgments alone, for example by providing converging evidence from multiple types of data, or testing subtle phenomena for which grammaticality judgments are difficult. Experimental work also contributes to a complete cognitive model extending beyond the competence–performance divide to investigate not only the nature of linguistic representations but also the computations involved when speakers, hearers, and readers access and manipulate these representations. For an ungrammatical sentence, we can ask *where* in the sentence the ungrammaticality arises, *when* in the processing of the sentence the ungrammaticality is noted, and *how* listeners/readers deal with the problem (e.g. If reading, do they backtrack, and if so, to which part of the sentence?). For a grammatical sentence, we can study the processes involved in assigning interpretation and integrating it into a discourse model.

Because experiments are becoming more common, it is critical that linguists understand the basics of experimental method and design. This paper provides an introduction. We first consider general aspects of design and method, and then survey a sampling of methodologies, focusing on syntax, semantics, and pragmatics. This review is intended as a practical introduction for the linguist who is beginning to conduct experiments as well as for readers of experimental papers.

2. Experimental Design

A thorough understanding of design issues requires more background than we have space for; a psychology textbook on research methods is recommended (for example, see Elmes et al. 2012). Here, we make some of the terms and concepts concrete with examples relevant to linguistics.

The steps in carrying out an experimental study are the following: (1) Formulate the hypothesis, (2) Determine the experimental design, (3) Get participants and run the study, (4) Analyze and interpret the data. Each step presents several factors to consider, so we go through them in turn.

2.1. FORMULATING THE HYPOTHESIS

The first step is formulating a specific and testable scientific hypothesis that can be tested experimentally. Suppose we are interested in the semantics of two quantifiers, and we believe that one has a more complex denotation than the other. To design an experiment to address this issue, we will first identify the *independent* and *dependent variables*.

The goal of an experimental study is to investigate how some element or elements, the *independent variable(s)*, affect some behavior or outcome, the *dependent variable*. The *independent variable* defines groups across which we look for differences, and we can intentionally manipulate these groups to test our hypothesis. The independent variable in this case is the quantifier; that is, we compare participants' behavior with one quantifier to their behavior with the other. Each quantifier represents a *level* of the independent variable, and the segment of the experimental procedure testing each level is called a *condition*.

Dependent variables are the outcome measures of the experiment – their value *depends on* the independent variables – and are specific, measurable behaviors. Dependent variables common in linguistics include reaction time (RT) and direction of eye gaze. Some dependent measures are *on-line* measures, which tap into language processing as it is happening, and others are *off-line* measures, which tap into the outcomes of that processing. RTs, obtained while the participant reads a sentence phrase-by-phrase, are an on-line measure. Acceptability ratings, which the participant provides after having read and processed a sentence, are an example of an off-line measure. Both measures are useful and can in fact be used in the same experiment, with online measures used to study the timecourse of interpretation, and off-line measures used to identify what interpretation was ultimately arrived at.

In the present case, we need a dependent variable that we expect to be affected by semantic complexity. RT is a common dependent measure for studying complexity (e.g. Frisson and Frazier 2005; Gennari and Poeppel 2003; Gillon et al. 1999; Kintsch 1974; McKoon and MacFarland 2002, Schmauder et al. 1991; Shapiro et al. 1987; Shapiro et al. 1991), with the prediction that more complex representations are associated with longer RT. A number of tasks measure RT, including lexical decision and self-paced reading. For example, in a self-paced reading task, participants read sentences that include one of the quantifiers, and we calculate how long they take to read the particular phrase containing that quantifier (see Section 3.2).

Having identified a measurable behavior to serve as dependent variable (RT) and identified the levels of the independent variable (quantifier 1 and quantifier 2), we can explicitly formulate the experiment's hypothesis in terms of the relationship we expect between them. Almost always, we hypothesize that *there is a difference between conditions*. Here, we hypothesize that RTs for the more complex quantifier will be longer than those for the less complex quantifier. Unintuitively, to investigate this hypothesis, we test its opposite, the *null hypothesis*, which states that *there is no difference between conditions*. If, after performing statistical tests, we decide to reject the null hypothesis, we say that the data support the hypothesis of interest. In this case, the null hypothesis we test experimentally is that there will be no difference in reading times between phrases containing the two quantifiers.

2.2. DESIGN AND METHOD

At this point, we have identified the main aspects of the design and method. But there are several further factors to consider.

©2013 The Author
Language and Linguistics Compass © 2013 Blackwell Publishing Ltd

2.2.1. Number of Conditions

To compare two quantifiers, a *1* × *2* design is the simplest. That is, there is one independent variable (quantifier) with two levels. But if we also are interested in how these quantifiers are interpreted in different sentence contexts, we can add sentence context as a second independent variable. If there are two sentence contexts to test, this becomes a *2* × *2* design; if three, a *2* × *3* design. More complex designs allow you to identify interactions and obtain a more sophisticated picture of the phenomenon – but also typically require more participants and more complex analyses. Simple designs are preferable for early investigations of a phenomenon; if an interesting effect surfaces, it can be fleshed out in subsequent experiments.

2.2.2. Within- and Between-subject Variables

Within-subject variables are manipulated by testing each participant at each level of the variable: each participant reads sentences with Quantifier 1 and sentences with Quantifier 2. Between-subject variables are manipulated by testing different participants at each level: half of the participants only read sentences with Quantifier 1, and half only sentences with Quantifier 2. How should we choose which to use?

Within-subject variables are generally preferable from an experimental design perspective. This is because differences between individual participants vary across multiple features, most of which are not the focus of our experiment. If we find a difference between conditions, it might be due to a difference in the quantifiers' complexity as we hypothesized, but there is also the possibility that it is due to a different, untested, variable (which we might call a confound, see below). For example, participants vary greatly in their baseline reading speed, which would affect overall RT. In a within-subject design, because the participants are the same in each condition, the conditions are equivalent with respect to participants' overall reading speed. Therefore, by using a within-subject design, we worry less about the possibility that a difference between conditions is due to participant-level variability.

There are, however, two reasons to consider using between-subject variables. One is concern that participants will guess the experimental variable because they see all variants of it. If participants are reading a large set of sentences, each containing one of two quantifiers, they may notice that there are two sentence types, and this may influence their responses. Filler items or analysis of order effects can alleviate this concern (see below). Another reason to use between-subject variables is if experience on one item is likely to affect performance on subsequent items, even without participants' conscious awareness. Syntactic priming, for example, is a robust phenomenon in which participants are more likely to use a syntactic construction if they have recently heard a sentence in that construction, and in some cases, it has long-lasting effects (Bock and Griffin 2000). We may unwittingly bias participants' responses on subsequent trials because of priming effects, even if the intention was not to conduct a priming study.

2.2.3. Confounds

Confounds occur when the levels of the independent variable vary directly with some other factor that is not of interest to the hypothesis of the study but nevertheless has an impact on the dependent variable. Suppose we are studying two groups of nouns, predicting that Group A nouns will produce shorter RTs than Group B nouns because the former are more imageable. But, it so happens that Group A nouns are also on average more frequent in the language than Group B nouns. Lexical frequency is a big contributor to RT: the higher the frequency, the lower the RTs. We may find the hypothesized difference in RT between

©2013 The Author
Language and Linguistics Compass © 2013 Blackwell Publishing Ltd

conditions, but this could be because of frequency rather than imageability. In this design, imageability and frequency are confounded.

To avoid confounds, we must identify factors that may be relevant and control for them. For example, we can choose a subset of the nouns in each group that are matched for lexical frequency. That is, our stimuli consist of pairs of nouns, one from each group, with similar lexical frequencies. We can also use statistical modeling to see if imageability has an effect over and above frequency. In either case, we would use corpora to determine the frequencies of these nouns (e.g. British National Corpus and Corpus of Contemporary American English). Many other factors affect processing, including semantic and phonological relatedness, word length (Baddeley et al. 1975), neighborhood density (e.g. Luce 1986), and age of acquisition (e.g. Brysbaert et al. 2000; Juhasz 2005).

2.2.4. Counterbalancing

Some confounds can be avoided by counterbalancing. Order of presentation is a common experimental design feature that is controlled by counterbalancing. If bilingual participants are completing a task in each of their two languages, we can counterbalance the order in which the tasks occur, such that half of the participants first do the task in Language X, and the other half first in Language Y. Then, if participants in the X-first group perform similarly to those in the Y-first group, we infer that differences in performance on the two tasks are not due to the order in which they were completed.

2.2.5. Randomization

Counterbalancing is sometimes impractical because of the number of factors to be controlled; in these cases, randomization is preferable. Suppose that participants are completing a picture-naming task in which they see 100 pictures and are asked to name them. We can randomly assign each trial a number from 1–100 and present the trials in this order, with different random number assignment for each participant. In psycholinguistic experiments, trials are often *pseudorandomized*: randomized lists are created, but are then edited to ensure that the same trial type does not appear too often in succession. In a *randomized block design*, participants or trials are first separated into homogeneous blocks, and randomization is done within those blocks. For example, if we believe that gender may play a significant role in behavior in our task, we can first divide participants into a group of men and one of women, and within each group, randomly assign individuals to condition. This design reduces variability, potentially providing a better estimate of the effects of the variable of interest.

2.2.6. Fillers

Filler trials can be included, interspersed with target trials. These are typically similar to the experimental stimuli but do not contain the particular element being studied (e.g. they lack quantifiers, or nouns from either Group A or Group B). Fillers are used to distract participants from the true purpose of the experiment and to screen participants for attentiveness. For example, comprehension questions asked during filler trials can reveal whether participants are attending to the task. Typically, there are twice as many fillers as targets in a task, although the length of the session should be kept reasonably short; fatigue can affect RTs (Milroy 1909).

2.2.7. A Note about Context

A common decision to make in designing experiments that present words or sentences as stimuli is whether to embed these materials in a relevant context or to present them in isolation. In experimental studies, we seek to control all aspects of the environment except for the variable of interest. But this takes language out of context. Reading times for sentences in isolation will differ from reading times for those same sentences in a discourse because expectations based on prior discourse, as well as extralinguistic context, affect linguistic processing (e.g. Altmann and Steedman 1988, Tanenhaus et al. 1995). This does not necessarily mean that all materials must be embedded in context. There may be good reasons to provide linguistic stimuli in isolation. However, it is important to anticipate how processing may be affected by this aspect of the design.

2.3. GET PARTICIPANTS AND RUN THE STUDY

2.3.1. Choosing a Sample

Although a given experiment tests a small number of participants, we hope the results are generalizable to a broader population, say, all adult native speakers of English. By hypothesis, all humans have the same capacity for language, barring certain pathologies. This means that we tend to worry less about how the sample is selected than if we were studying traits that are expected to be more heavily influenced by personal experiences, and psycholinguistic experiments commonly use a sample of undergraduate students. The choice to use undergraduates as participants can be problematic, of course, and the results may not be widely generalizable (see Henrich et al., 2010 for discussion).

Potentially more dangerous than testing a sample of undergraduates is testing a sample of linguists (Gibson and Fedorenko 2013: but see Phillips 2009). Linguistic training requires you to poke and prod at meaningless sentences to identify their structure and to challenge your introspective abilities to determine whether sentences are ungrammatical or simply odd. These practices undoubtedly affect language processing in an experimental task, as well as the ability to guess the experimenter's hypotheses. In most cases, a naïve sample is best.

2.3.2. Number of Participants and Trials

We are ultimately going to make inferences about the target population using statistical analyses, so we need to be sure that the sample size is large enough to detect underlying differences between conditions. The researcher is advised to look at existing research using similar methods to decide on a sample size. Some kinds of studies yield large effects, but more commonly effect sizes are small, and large sample sizes are needed to detect them. A rule of thumb for studies measuring RT: the number of subjects and the number of trials should be about four times the number of conditions.

2.3.3. Random Assignment

As mentioned above, if a variable is between-subject, we must do our best to ensure that any group differences in the data are due to the manipulated variable and not to other underlying differences between the participants. For example, an obvious flaw would be testing only males in one condition and only females in the other. But subtler errors are also possible. If we run one condition in the middle of the semester and the second before exams, performance might differ due simply to participants' general fatigue. By randomly assigning each participant to a level of between-subject variables, we minimize this kind of error.

2.4. ANALYZE AND INTERPRET THE DATA

We do not discuss statistical analysis here; there are some excellent textbooks specifically for linguists: Baayen 2008, Gries 2010, and Johnson 2008. Instead, we consider some general issues that arise in interpreting experimental results.

2.4.1. Fixed and Random Factors

A random factor is one for which we sample non-exhaustively from a larger population, with the goal of generalizing to that population. For example, we test a group of participants, but we are interested in generalizing beyond those specific individuals. Similarly, our experimental stimuli do not include all of the words or sentences in the language, but we want to generalize to the larger language. A fixed factor is an independent variable that we want to include in our analysis, that is, the manipulation of interest.

2.4.2. Statistical Significance

The result of performing a statistical test is a test statistic with a known distribution; our goal is to compare the test statistic to the distribution to determine whether this result could have come about by chance or is extreme enough to be unlikely to be attributable to chance. Significance level is the cutoff probability for determining that a test statistic is extreme and is typically set at 0.05 (that is, the probability that we would have achieved this result or a more extreme result *if the null hypothesis were true* is 5%). We say that the outcome is statistically significant if the *p*-value obtained is less than this cutoff.

2.4.3. Interactions

In a factorial design involving more than one independent variable, we can examine *interaction* effects in addition to *main effects* of each variable by itself. Suppose we are interested in the effects of word length and word frequency on reading times. By including them both in the same experimental design rather than investigating each in separate experiments, we can ask whether the effect of one of these variables is somehow qualified by the other. For example, if high frequency short words yield faster RTs than high frequency long words, but word length has no effect for low frequency words, then an experiment studying word length alone might yield no effect. A factorial design including both variables, however, could reveal an interaction between them. Of course, without a theoretical basis for believing such an interaction exists, you might not think to include both variables in a factorial design. Interaction effects, then, typically involve more complex explanations than main effects.

2.4.4. Ceiling and floor effects

Ceiling effects occur when a task is easy and participants perform so well that no differences between groups emerge. For example, if RTs are extremely fast (close to the fastest possible human response time), there will be no differences between the two quantifiers, even if one is in fact slightly more difficult to process than the other (note that for RTs, good performance means shorter times, so the ‘ceiling’ corresponds to low values). If you suspect ceiling effects, you can make the task more difficult, by, for example, adding a simultaneous task that participants must perform, like remembering a string of digits.

Floor effects are the opposite of ceiling effects; they occur when a task is so difficult that participants perform poorly across the board, and again, no differences between groups are evident. Studies with children can suffer particularly from floor effects if studies are not well designed. Children may perform poorly if asked for grammaticality judgments but may perform better if the task is situated in a game in which a puppet produces a sentence, and the child rewards the puppet for speaking well or punishes it for speaking poorly (Stromswold, 1990). For adults, tasks requiring speed or intense concentration for extended periods may be too demanding.

©2013 The Author
Language and Linguistics Compass © 2013 Blackwell Publishing Ltd

2.4.5. Type I and Type II Error

After performing a statistical test, we determine whether there is a statistically significant difference between conditions. If so, we reject the null hypothesis; if not, we fail to reject it. What do these outcomes reveal about linguistic knowledge? Consider the possible outcomes in Table 1.

If you find yourself in the top row, congratulations! You should prepare a manuscript to submit to a journal. If in the bottom row, you have a 'null result'. You failed to find a difference between conditions. When you obtain a null result, you may have learned something of value to science, but it is nevertheless difficult to get published (except perhaps in the *Journal of Articles in Support of the Null Hypothesis*). You can avoid this outcome by designing the study to be interesting whether or not you reject the null hypothesis. For example, in Experiment 1, replicate an attested effect, and in Experiment 2, use very similar methods to test something related but slightly different; the positive result in Experiment 1 can make Experiment 2 interesting for the field whether or not you find an effect.

But it is important to consider the columns in Table 1 as well. The columns relate experimental findings to reality, to the true presence or absence of the effect you found. The upper left cell represents an erroneous conclusion; although you found a difference between conditions, in reality there is no such difference. Our field uses a fairly strict criterion for deciding that a difference is statistically significant ($\alpha = 0.05$), so such errors are rare. But to minimize Type I error, you are also responsible for doing good science; that is, frame good hypotheses that are more likely to be true than false.

The lower right cell is another instance of an erroneous conclusion, called Type II error. Here, you have decided not to reject the null hypothesis, when it is in fact false; that is, you failed to detect a real effect. This means you lacked sufficient *power*. The power of a statistical test is the probability that the test will reject the null hypothesis when the null hypothesis is false. Increasing the sample size increases the power of a test (see Cohen 1992 for a quick introduction to power). But if an effect is small, you are also more likely to detect it if the data is not noisy. It is therefore imperative to control for variables that introduce noise. If you administer an RT study over the web, for example, you may obtain noisier data than if you test all participants in the same quiet room.

3. Methods

This section surveys some methods that are either common or new but particularly promising.[1] We continue our focus on syntactic, semantic, and pragmatic knowledge, emphasizing practical concerns. The methods are simply mentioned here, with a brief description or examples of studies that have used them. Note that many studies use multiple methods (e.g. sentence completion and reading time) to provide support from converging evidence or to illuminate distinct aspects of the same phenomenon (McKoon and Ratcliff (2003), for example, use five different methods and a corpus study to examine reduced relative clauses).

Table 1. Possible outcomes, including errors, of significance testing.

	Null hypothesis is in fact true	**Null hypothesis is in fact false**
Decide to reject null hypothesis	Type I error	Correct
Decide *not* to reject null hypothesis	Correct	Type II error

©2013 The Author
Language and Linguistics Compass © 2013 Blackwell Publishing Ltd

3.1. BEYOND INTROSPECTIVE JUDGMENTS: AN EXPERIMENTAL APPROACH

The traditional measures in linguistics, introspective acceptability and truth-value judgments obtained from the linguist herself, are, in a sense, experiments; they simply have a sample size of one and lack the careful controls discussed here. Taking an experimental approach to judgment data means, at the very least, using a larger sample, which allows for a number of naïve participants, provides statistical power, and avoids investigator biases (Gibson and Fedorenko 2013). An understanding of experimental design can also clarify what factors should be controlled for, whether fillers are needed to distract the participant from the study's purpose, or whether narratives or videos should be included to control the contexts in which judgments are provided (e.g. Arunachalam and Kothari 2012). Other variations include using rating scales instead of yes/no responses and magnitude estimation. See Schütze and Sprouse (2012) for detailed and practical discussion of these techniques.

3.2. MEASURES OF READING

Several reading techniques are used to study language comprehension. In self-paced reading, participants determine the rate at which text is presented on a monitor by pressing a button. We infer the reader's difficulty in interpreting the text from the rate of button presses. Techniques differ on whether a single word, a phrase (or 4–5 word chunk), or an entire sentence is presented on each screen. They also differ in whether already read words remain on the screen when the subsequent words are presented (cumulative technique) or whether they are replaced by dashes with the subsequent words appearing to their right (moving window technique). The moving window technique has been argued to most closely resemble natural reading (Just et al. 1982). A further variant, rapid serial visual presentation, is not self-paced; the text is presented at a fast, fixed rate. Reading time measures have been used in studies of many phenomena. We list just a few here, with a reference that clearly describes the method for each: filler-gap dependencies (e.g. Aoshima et al. 2004), parasitic gaps (e.g. Phillips 2006), negative polarity items (e.g. Szabolcsi et al. 2008), and anaphor resolution (e.g. Clifton and Ferreira 1987).

One drawback of reading time measures is that presenting text in chunks interferes with natural reading. When reading naturally, in addition to moving our eyes forward – in English, rightward – to read new text on a line, we also make *regressions*, or saccades backward to previously read material. Self-paced reading and rapid serial visual presentation interfere with natural regressions. Eye tracking during reading is a more naturalistic method for studying reading.

Eye tracking systems used in psycholinguistics typically involve the participant's eyes being video-recorded and the image processed with image-processing software to determine the coordinates of gaze. There are head-mounted systems, in which the participant wears the camera assembly on her head; fixed-head systems, in which the participant rests her head in a chin rest; and remote systems, which neither restrain the participant nor require her to wear anything. Many measures can be analyzed, including duration and number of fixations, and duration and number of regressions, which may reflect different aspects of processing (see Pickering et al. 2004 for case studies).

Though commonly used to study syntactic processing, eye tracking has also provided insights on many other issues, including lexical ambiguity, morphological processing, and discourse effects on sentence processing (see Rayner 1998, for a classic review, and Staub and Rayner 2007 for a more recent one). Perhaps the biggest drawback of eye tracking during reading is the expense of the equipment.

©2013 The Author
Language and Linguistics Compass © 2013 Blackwell Publishing Ltd

3.3. EYE TRACKING DURING LISTENING

Differing somewhat from eye tracking during reading in equipment, methods, and analysis techniques, visual world eye tracking allows researchers to study eye gaze while participants inspect visual scenes, either in the world or on a screen. When looking at a scene while hearing language, the listener's direction of gaze is closely *time-locked* to her interpretation of the sentence (Cooper 1974), allowing us to study how listeners interpret auditorily presented sentences online, as they unfold (e.g. Tanenhaus et al. 1995). This is commonly done with a remote eye-tracking system to record participants' eye gaze as they view scenes on a monitor, but visual world eye tracking can also be done with a video camera. Sometimes called 'poor man's' eye tracking, in this method a camera is directed at the participant's face, and the video is later played back frame-by-frame, and direction of gaze manually coded. Though labor intensive, this method requires little equipment and is therefore well-suited for experiments conducted in the field.

3.4. SELF-PACED COUNTING

Hackl (2009) presents a clever RT paradigm for studying words with quantificational semantics: self-paced counting. In this paradigm, the participant advances a display from one screen to the next as in self-paced reading. Each screen depicts a number of dots, all of which are initially covered; on each subsequent screen, some dots are uncovered, and their color is revealed. The participant's task is to judge the truth/falsity of a statement that describes how many dots are of a particular color; RT to press the space bar is calculated. The dots are revealed incrementally such that the truth or falsity of the sentence cannot be judged until the last screen, but RT indicates how difficult the computation is for the participant. A long RT suggests that the participant is engaging in mental calculation to determine whether the statement is true, while a short RT suggests that the indeterminacy of the answer is readily apparent from a quick glance at the dots.

3.5. OTHER REACTION TIME METHODS: PRIMING AND LEXICAL DECISION

Priming techniques measure how the amount of time required to process and react to a *target* is affected by having previously processed a *prime*. It is well documented that processing is affected by the phonological, semantic, orthographic, and structural similarity of the target to the earlier prime. For example, readers are faster to read the word 'nurse' if they have just read 'doctor', than if they have just read 'chair'. This allows us to draw conclusions about how these representations are organized in the mental lexicon. This is an extremely common method in psycholinguistics; for a review of the classic findings, see Neely (1991).

Lexical decision is an RT technique in which participants are asked to judge whether a string of letters (presented visually) or phonemes (presented auditorily) compose a word or a non-word. This technique, often combined with semantic priming, is used to investigate how lexical representations are organized in memory and the processes involved in retrieving these representations from the lexicon.

3.6. WORKING WITH CHILDREN

There are many special populations that can be studied to gain insight into linguistic theory and behavior. But because observations and theories about language acquisition have played such a fundamental role in the development of modern linguistic theory, typically developing children are of special interest.

One popular method for studying language acquisition in children is the truth-value judgment task, in which the child is asked to judge whether a puppet's utterance is a true or false description of a depicted scene or story. There are many excellent resources on this task, so we will not go into it here. See for example, Gordon (1996) and Crain and Thornton (1998).

Eye tracking is a newer method with children but is particularly useful because it does not require verbal or pointing responses. Eye movement patterns are similar in children and adults (but see Kowler and Martins 1982), and as with adults, children's eye movements are time-locked with the speech they hear, so this technique can reveal incremental language processing (e.g. Trueswell et al. 1999). Both 'poor man's' and remote eye-tracking methods can be used even with infants. The primary concerns beyond those of eye tracking with adults are identifying visual stimuli that are comparably salient and interesting to the child, and creating a physical setup (chair and table height, etc.) that is likely to yield high quality data. See Fernald et al. (2008) and Trueswell (2008) for discussion of these and of analysis issues.

3.7. ADDITIONAL REFERENCES

For more comprehensive treatment of many of the methods described above, including theoretical as well as methodological issues, see, for example, Altmann (1998), Gibson and Pearlmutter (1998), and Trueswell and Tanenhaus (1995) on sentence comprehension, Grosjean and Frauenfelder (1997) on spoken word recognition, Adelman (2012) on visual word recognition, and chapters on a variety of topics in Carreiras and Clifton (2004), Gaskell (2007), Spivey et al. (2012), and Traxler and Gernsbacher (2006).

4. Conclusion

Experimental research has provided a richer understanding of the nature and organization of linguistic representations, and the nature and timing of linguistic processes, and – with increasing numbers of studies using brain imaging techniques – even a glimpse of their neural substrates. As research methods available to linguists become more sophisticated, it is increasingly important to have a thorough understanding of experimental methods and design.We hope that this review provides a helpful starting point for the linguist interested in exploring these techniques.

Short Biography

Sudha Arunachalam's research focuses on language acquisition and language processing, specifically the acquisition, representation, and retrieval of the meanings and syntactic properties of words. She is currently Assistant Professor in Boston University's Department of Speech, Language, & Hearing Sciences. She holds a B.A. in Linguistics and Psychology from the University of Southern California, an M.A. in Psychology from the University of Pennsylvania, and a Ph.D. in Linguistics also from the University of Pennsylvania.

Acknowledgments

This overview of experimental methods and design stems from a course at the European Summer School for Logic, Language, and Information in Ljubljana, Slovenia, in 2011. Thanks to the students who attended for their input.

Notes

* Correspondence to: Sudha Arunachalam, Speech, Language, & Hearing Sciences, Boston University, 635 Commonwealth Ave., Boston, MA 02215, USA. E-mail: sarunach@bu.edu

[1] Many of these methods are high-tech, requiring special equipment. But there are many low-tech methods too, including paper and pencil sentence-completion tasks and syntactic priming production paradigms (e.g. Pickering et al. 2002). Experiments can also be conducted over the internet, eliminating the need for lab space or sophisticated equipment. See Keller et al. (2009) and Sprouse (2011) for discussion of validity of RT and judgment measures collected over the web.

Works Cited

Adelman, James S. 2012. Visual word recognition Vol. 1. New York: Psychology Press.

Altmann, Gerry T. M., and Mark J. Steedman. 1988. Interaction with context during human sentence processing. Cognition 30. 191–238.

——. 1998. Ambiguity and sentence processing. Trends in Cognitive Sciences 2. 146–152.

Aoshima, Sachiko, Colin Phillips, and Amy Weinberg. 2004. Processing filler-gap dependencies in a head-final language. Journal of Memory and Language 51. 23–54.

Arunachalam, Sudha, and Anubha Kothari. 2012. An experimental study of Hindi and English perfective interpretation. Journal of South Asian linguistics 4. 27–42.

Baayen, R. Harald. 2008. Analyzing linguistic data: a practical introduction to statistics using R. New York: Cambridge University Press.

Baddeley, Alan D., Neil Thomson, and Mary Buchanan. 1975. Word length and the structure of short-term memory. Journal of verbal learning and verbal behavior 14. 575–589.

Bock, J. Kathryn, and Zenzi M. Griffin. 2000. The persistence of structural priming: transient activation or implicit learning? Journal of Experimental Psychology. General 129. 177–192.

Brysbaert, Marc, Ilse Van Wijnendaele, and Simon De Deyne. 2000. Age-of-acquisition effects in semantic processing tasks. Acta Psychologica 104. 215–226.

Carreiras, Manuel, and Charles Clifton, Jr. 2004. The on-line study of sentence comprehension. New York: Psychology Press.

Clifton, Charles Jr., and Fernanda Ferreira. 1987. Discourse structure and anaphora: some experimental results. Attention and performance 12: the psychology of reading, ed. by Max Coltheart, 635–654. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, Jacob. 1992. A power primer. Psychological Bulletin 112. 155–159.

Cooper, Roger M. 1974. The control of eye fixation by the meaning of spoken language. Cognitive Psychology 6. 84–107.

Crain, Stephen, and Rosalind Thornton. 1998. Investigations in universal grammar: a guide to experiments on the acquisition of syntax and semantics. Cambridge, MA: MIT Press.

Elmes, David G., Barry H. Kantowitz, and Henry L. Roediger, III. 2012. Research methods in psychology, Ninth Edition. Belmont, CA: Wadsworth.

Fernald, Anne, Renate Zangl, Ana Luz Portillo, and Virginia A. Marchman. 2008. Looking while listening: using eye movements to monitor spoken language comprehension by infants and young children. Developmental psycholinguistics: online methods in children's language processing, ed. by Irina A. Sekerina, Eva M. Fernández, and Harald Clahsen, 97–136. Amsterdam: John Benjamins.

Frisson, Steven, and Lyn Frazier. 2005. Carving up word meaning: portioning and grinding. Journal of Memory and Language 53. 277–291.

Gaskell, M. Gareth. 2007. The oxford handbook of psycholinguistics. New York: Oxford University Press.

Gennari, Silvia, and David Poeppel. 2003. Processing correlates of lexical semantic complexity. Cognition 89. B27–B41.

Gibson, Edward, and Neal J. Pearlmutter. 1998. Constraints on sentence comprehension. Trends in Cognitive Sciences 2.7. 262–268.

——. and Evelina Fedorenko. 2013. The need for quantitative methods in syntax and semantics research. Language & Cognitive Processes, 28. 125–155.

Gillon, Brendan, Eva Kehayia, and Vanessa Taler. 1999. The mass/count distinction: evidence from online psycholinguistic performance. Brain and Language 68. 205–211.

Gordon, Peter. 1996. Methods for assessing children's syntax. Cambridge, MA: MIT Press.

Gries, Stefan Th. 2010. Statistics for linguistics with R: a practical introduction. Berlin: Mouton de Gruyter.

Grosjean, François, and Uli H. Frauenfelder. 1997. A guide to spoken word recognition paradigms. Hove, U.K.: Psychology Press.

Hackl, Martin. 2009. On the grammar and processing of proportional quantifiers: *most* versus *more than half*. Natural Language Semantic 17.63–98.

Henrich, Joseph, Heine, Steven J., and Norenzayan, Ara. 2010. The weirdest people in the world?. Behavioral and Brain Sciences 33. 83-135.

©2013 The Author
Language and Linguistics Compass © 2013 Blackwell Publishing Ltd

Johnson, Keith. 2008. Quantitative methods in linguistics. Oxford, UK: Blackwell Publishing.

Juhasz, Barbara J. 2005. Age-of-acquisition effects in word and picture identification. Psychological Bulletin 131. 684–712.

Just, Marcel A., Patricia A., Carpenter, and Jacqueline D., Woolley. 1982. Journal of Experimental Psychology: General 111. 228–238.

Keller, Frank, Subahshini Gunasekharan, Neil Mayo, and Martin Corley. 2009. Timing accuracy of Web experiments: a case study using the WebExp software package. Behavior Research Methods 41. 1–12.

Kintsch, Walter. 1974. The representation of meaning in memory. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Kowler, Eileen, and Albert J., Martins. 1982. Eye movements of preschool children. Science 215. 997–999.

Luce, Paul A. 1986. Neighborhoods of words in the mental lexicon. Research on speech perception, Technical report 6. Bloomington: Indiana University.

McKoon, Gail, and Talke Macfarland. 2002. Event templates in the lexical representations of verbs. Cognitive Psychology 45. 1–44.

——. and Roger Ratcliff. 2003. Meaning through syntax: language comprehension and the reduced relative clause construction. Psychological Review 110. 490–525.

Milroy, T. H. 1909. Fatigue studied in reaction time experiments. Experimental Psychology 2. 277–282.

Neely, James H. 1991. Semantic priming effects in visual word recognition: a selective review of current findings and theories. Basic processes in reading: visual word recognition, ed. by Derek Besner and Glyn W. Humphreys, 264–336. Hillsdale, NJ: Lawrence Erlbaum.

Phillips, Colin. 2006. The real-time status of island phenomena. Language 82. 795–823.

—— 2009. Should we impeach armchair linguists? Japanese–Korean linguistics (Vol. 17), ed. by Shoishi Iwasaki, Hajime Hoji, Patricia M. Clancy, and Sung–Ock Sohn. Stanford, CA: CSLI Publications.

Pickering, Martin J., Holly P. Branigan, and Janet F. McLean. 2002. Constituent structure is formulated in one stage. Journal of Memory and Language 46. 586–605.

——. Steven Frisson, Brian McElree, and Matthew Traxler. 2004. Eye movements and semantic composition. The online study of sentence comprehension: eyetracking, ERPs, and beyond, ed. by Manuel Carreiras and Charles Clifton, Jr. Hove: Psychology Press.

Rayner, Keith. 1998. Eye movements in reading and information processing: 20 years of research. Psychological Bulletin 124. 372–422.

Schmauder, A. René, Shelia M. Kennison, and Charles Clifton, Jr. 1991. On the conditions necessary for obtaining argument structure complexity effects. Journal of Experimental Psychology: Learning, Memory, and Cognition 16. 1188–1192.

Schütze, Carson, and Jon Sprouse. 2012. Judgment data. Research methods in linguistics, ed. by Robert J. Podesva and Devyani Sharma, to appear.

Shapiro, Lewis P., Edgar Zurif, and Jane Grimshaw. 1987. Sentence processing and the mental representation of verbs. Cognition 27. 219–246.

——. Bari Brookins, Betsy Gordon, and Nicholas Nagel. 1991. Verb effects during sentence processing. Journal of Experimental Psychology: Learning, Memory, and Cognition 17. 983–996.

Spivey, Michael, Ken McRae, and Marc Joanisse. 2012. The Cambridge handbook of psycholinguistics. New York: Cambridge University Press.

Sprouse, Jon. 2011. A validation of Amazon mechanical Turk for the collection of acceptability judgments in linguistic theory. Behavior Research Methods 43. 155–167.

Staub, Adrian, and Keith Rayner. 2007. Eye movements and online comprehension processes. The Oxford handbook of psycholinguistics, ed. by M. Gareth Gaskell. 327–342. New York: Oxford University Press.

Stromswold, Karin. 1990. Learnability and the acquisition of auxiliaries. Doctoral Dissertation, Massachusetts Institute of Technology.

Szabolcsi, Anna, Lewis L. Bott, and Brian McElree. 2008. The effect of negative polarity items on inference verification. Journal of Semantics 25. 411–450.

Tanenhaus, Michael K., Michael J. Spivey–Knowlton, Kathleen M. Eberhard, and Julie C. Sedivy. 1995. Integration of visual and linguistic information in spoken language comprehension. Science 268. 1632–1634.

Traxler, Matthew, and Morton Ann Gernsbacher. 2006. Handbook of psycholinguistics. New York: Academic Press, Elsevier.

Trueswell, John C. 2008. Using eye movements as a developmental measure within psycholinguistics. Developmental psycholinguistics: online methods in children's language processing, ed. by Irina A. Sekerina, Eva M. Fernández, and Harald Clahsen, 73–96. Amsterdam: John Benjamins.

——. and Michael K. Tanenhaus. 1995. Sentence comprehension. Handbook of perception and cognition Vol. 11: speech, language, and communication, ed. by Joanne L. Miller and Peter D. Eimas, 217–262.

Trueswell, John C., Irina, Sekerina, Nicole M., Hil, and Marian L., Logrip. 1999. The kindergarten-path effect: studying on-line sentence processing in young children. Cognition 73. 89–134.

©2013 The Author
Language and Linguistics Compass © 2013 Blackwell Publishing Ltd