This article surveys the major kinds of empirical evidence used by linguists, with a particular focus on the relevance of the evidence to the goals of generative grammar. After a background section overviewsing the objectives and assumptions of that framework, three broad kinds of data are considered in the three subsequent sections: corpus data, judgment data, and (other) experimental data. The perspective adopted is that all three have their place in the linguist’s toolbox: they have relative advantages and disadvantages that often complement one another, so converging evidence of more than one kind can reasonably be sought in many instances. Points are illustrated mainly with examples from syntax, but often can be easily translated to other levels (e.g., phonology, morphology, semantics, and pragmatics). © 2010 John Wiley & Sons, Ltd.

BACKGROUND

Goals of Linguistic Theory

Chomsky has often characterized the foundational goals of generative linguistics qua subfield of cognitive psychology (and cognitive neuroscience) in terms of one central question and three associated questions: the central question is ‘What constitutes knowledge of (a) language?’ and the associated questions are ‘How is that knowledge acquired?’, ‘How is that knowledge used (in comprehension, production, and other activities)?’, and ‘How is that knowledge encoded in the brain?’ To make sense of these questions, we need to be clear on what Chomsky means by ‘knowledge of a language’, also known as linguistic competence.1 (We shall return to the more abstract notion ‘knowledge of language’ shortly.) This is not a reference to grammar rules that we might have been taught in school, but rather to tacit knowledge that we have no conscious access to, knowledge that governs, among other things, how we interpret the sound waves that reach our ears as a message being communicated, much as our visual system embodies tacit knowledge that governs how we interpret the light waves striking our retinas as a three-dimensional environment surrounding us.

This article focuses on the kinds of linguistic evidence that are typically used in addressing the central question—that is, in determining the properties of linguistic expressions in an adult (native) speaker’s competence. It will thus be assumed, for purposes of discussion, that linguistic competence (in Chomsky’s sense, contrasted with performance) exists. Researchers who adopt different assumptions may use the same sorts of evidence to different ends, which I do not explore.

Why is the word ‘a’ parenthesized in the phrasing of the central question? Because it actually stands for two intimately related questions. The specific question about knowledge of a language concerns our tacit knowledge of the particular language(s) we speak. Generative linguistics seeks to identify what that knowledge is, but not in a vacuum. It also seeks to delimit the range of human languages—not just the set of all known languages, but the possible languages that could exist or could have existed throughout the history of the human species, i.e., the languages that our brains are capable of encoding and we are capable of learning and using. This is ‘knowledge of language’ in general, also known as Universal Grammar.

Grammars and Grammaticality

In generative linguistics, the term for the system that constitutes our knowledge of a particular language is a (mental) ‘grammar’; we each have the grammar of at least one language in our mind/brain. A hypothesis about the information contained therein, stated in
terms of any of a number of formalisms, is also called a ‘grammar’. (Some researchers might want to differentiate the latter by calling it a ‘model of the grammar’.) Universal Grammar is comprised of those properties that (by hypothesis) are necessarily common to all possible human language grammars, along with specifications of the extent to which grammars may vary on properties that are not common to all of them (sometimes called ‘parameters’). At the level of the word and above, a grammar can be minimally conceived of as a mapping (partial relation) between sequences of morphemes/words and meanings. A grammar proposed by a linguist might be intended as just some arbitrary way of defining the mapping that the mental grammar is hypothesized to define (i.e., any specification of that function would suffice), or it might be intended additionally as a hypothesis about the algorithm via which the human mind/brain computes that mapping. (This is a point on which even researchers within the same theoretical framework do not necessarily agree.)

Many generative linguists concur with Chomsky’s view that our mental grammar distinguishes at least two kinds of strings: those that are possible sentences of our language and those that are not. All speakers of English agree that the string in (1) is a well-formed sentence. Similarly, all speakers agree that the reverse string in (2) is not a well-formed sentence; this is conventionally indicated by preceding the sentence with an asterisk.

1. What did John eat for dinner?
2. *Dinner for eat John did what?

This notion of ‘well-formed’ or ‘possible’ versus ‘ill-formed’ or ‘not possible’ is not the same distinction as that between strings that can be interpreted and those that cannot. Although most speakers, if asked what was meant by (2), would probably feel there was no answer or they were guessing wildly, the same is not true for (3).

3. *What did John ate for dinner?

In this case, most speakers would say that (3) probably means what (1) means, but at the same time they would still say that (3) is not a possible sentence of English, just as they do for (2). This is not to deny that, if asked, they might rate (3) as ‘closer to English’ than (2). That would not undermine the claim that (2) and (3) share the status of being impossible, just as in the domain of color perception people might agree that an orange swatch is ‘closer to’ the color of a red swatch than a green swatch is, while still having no doubt that neither the orange nor the green swatch is red.

As it turns out, linguistic theory has devoted virtually no attention to the questions of which strings can be assigned a consistent interpretation or how the aforementioned notion of closeness might be reified. Rather, its focus has been on distinguishing (1) on the one hand from (2) and (3) on the other, i.e., delineating well-formed or ‘grammatical’ versus ill-formed or ‘ungrammatical’ sentences and explaining various properties of the grammatical ones (e.g., their meanings). According to Chomsky, there is no way to operationalize the notion of grammaticality in other terms. Not only does it not correspond to interpretability, it also does not correspond to probability of occurrence: speakers can consistently distinguish well-formed from ill-formed strings among those whose probabilities are zero, based on past occurrence (see examples in Box 2). This is not to deny that speakers, if asked, can detect probability/frequency differences among many sentence types; the claim is that grammaticality cannot be wholly reduced to such factors. There may even be cases (see Drawbacks under Section Acceptability Judgment Data) where the mental grammar defines some sentence as part of the language even though speakers, when asked for their impression of it (often called an ‘acceptability judgment’—see Section A Brief Digression on Terminology), find it ill-formed.

Although there are a great many strings like (1)–(3) whose status all native English speakers agree on (and likewise for every other language), there are also some strings that seem to cause considerable disagreement among speakers of ‘the same language’ as to whether they are possible. In some instances this disagreement is systematic, varying along geographical or socio-cultural dimensions. For example, the use of the word anymore in positive declarative sentences to mean something along the lines of ‘nowadays’, as in (4), is chiefly isolated to specific geographic regions, such as the greater Philadelphia metropolitan area; almost all speakers from other regions find sentences like (4) ill-formed.²

4. It gets dark early anymore.

However, there are other instances of inter-speaker disagreement that show no obvious signs of correlating with any independent properties of the speakers, such as (5), which is possible for many speakers but impossible for a large minority, who allow only John promised to wash… and John promised Mary that he would wash….
5. John promised Mary to wash the dishes.

In light of these observations, a single grammar of English cannot apply to all speakers, not even to all speakers of ‘the same dialect’, if one looks closely enough. But this yields no particular problem if we take the generativist perspective that there is no object in the world called ‘English’ or even ‘Southern Ontario English’. There are simply a number of speakers whose grammars are similar enough that they can understand one another (most of the time) and who we are therefore willing to casually describe as speaking the same language. Nonetheless, each person’s language is defined exclusively by what is inside his or her own head (I-language rather than E-language, in Chomsky’s terms), so these grammars can be somewhat different. [Even some proponents of a rich, innate Universal Grammar do not claim that its universal (i.e., nonparametric) components are invariant across the species, just as the genetic coding for certain fundamental properties, such as the number of limbs and organs, is invariant, putting aside pathology.)

One other respect in which establishing a grammar (for an individual or a group) is challenging is that there are some sentences that seem neither clearly possible nor clearly impossible, but somewhere ‘in between’; often, many speakers will agree on their middling status, evincing a systematic phenomenon. Examples of this kind, such as (6), are often designated by a preceding question mark.

6. ?The child was kept an eye on by the babysitter.5

As I discuss in Section Gradience in Acceptability and Grammaticality, such phenomena are compatible with Chomsky’s view that the mental grammar itself actually draws a sharp boundary between possible and impossible sentences, although some researchers take the same phenomena as evidence that grammar does not have this property.

A Brief Digression on Terminology

Speakers’ reactions to sentences as being possible or impossible (or somewhere in between) have traditionally been referred to as ‘grammaticality judgments’, but this term is misleading.6 Because a grammar is a mental construct not accessible to conscious awareness, speakers cannot have any impressions about the status of a sentence with respect to that grammar; rather, in Chomsky’s terms one should say their reactions concern acceptability, that is, the extent to which the sentence sounds ‘good’ or ‘acceptable’ to them. Therefore I will use the term ‘acceptability judgments’. Likewise, as properties of sentences I will reserve the terms ‘(un)grammatical’ and ‘(un)grammaticality’ to refer to their (hypothesized) status vis-à-vis the mental grammar, and the terms ‘(un)acceptable’ and ‘(un)acceptability’ to refer to their status according to a speaker who is asked about them.

CORPUS DATA

A corpus can be defined as a collection of language data used for linguistic study. It might consist of material that was produced at the request of a researcher, e.g., the telling of traditional stories by speakers of a language for which there is no written record. Outside the realm of field linguistics, however, it more typically consists of material that had already been created for some purpose other than linguistic research, e.g., the contents of newspapers, books, and other published writings, as well as spoken language in the form of television or radio broadcasts, parliamentary debates, political speeches, and so on. I distinguish two types of corpora: ‘traditional’ corpora (Section Traditional Corpora), which include any bodies of language of the types just mentioned when assembled by linguists (such as the Brown corpus, CELEX, the British National Corpus, the Corpus of Contemporary American English, etc.), versus the World Wide Web as a corpus (Section The World Wide Web as a Corpus).

Notwithstanding their pros and cons, there are situations where corpus data must be used. Corpora often constitute the only evidence available for investigating a language with no living speakers. (The number of such languages is growing rapidly: somewhere between 20 and 50% of the languages spoken in the world as of the year 2000 were not being learned by children and will thus cease to be spoken by 2100, unless dramatic changes occur.11 This would be the equivalent of half the species on earth becoming extinct over that same time period.)

Traditional Corpora

Benefits

Evidence from traditional corpora has numerous desirable properties (vis-à-vis the goals mentioned in Section Background), several of which are not shared by other kinds of evidence. In the typical scenario, the contents of a corpus could not have been biased by the researcher. Corpora can provide samples of language produced when the speakers/writers were not consciously focusing on the form of their utterances.
A corpus might bring to our attention phenomena that would not otherwise have come to light: for example, corpus data are increasingly being used to show that constructions that seem ill-formed when presented as isolated examples may nonetheless sound natural in everyday situations.\(^{12}\) Corpora may allow the study of phenomena not testable in sentences in isolation, e.g., how referents are identified throughout a narrative or discourse. If a corpus is in electronic form (either based on original written documents or transcribed from audio recordings), automatic searching and statistical analysis can be performed. A very limited number of such corpora, mostly smaller ones, have been parsed, allowing searches based on structural configurations. This is the only way to search for many of the phenomena of interest to linguists, because they often cannot be identified by any particular (string of) words.

**Drawbacks**

The aforementioned benefits must be considered in the light of certain limitations to corpus data. Most importantly, the absence of some phenomenon is hard to interpret: is it absent because it is not possible in the language, or because of the nature of the corpus? As discussed in Box 1, the former situation would be important to identify, but if it cannot be distinguished from the latter then a critical kind of evidence for theory construction will be lacking.

For example, suppose that a particular corpus contains no occurrences of a matrix question formed by extraction from three clauses down, as in (7). This could be an artifact of the corpus collection process: the frequency of the phenomenon may be too low for it to have arisen (because the corpus is not large enough), or the communicative situations represented in the corpus may not have lent themselves to the kind of content expressed by it (because the corpus is not diverse enough), or if we had no other kind of evidence concerning the structure in (7), it might not be a possible structure of English.

7. Which books do you think [the school board wants [its teachers to forbid their students [to read t]]]?

More often a linguist is concerned not with simple attestation of some phenomenon, but with establishing its properties once it has been attested. Because the phenomena of greatest interest to linguists are often extremely rare, even very large and diverse corpora may not contain the substantial number of exemplars required for this purpose. In parsed corpora a further problem can ensue, because the original design of the parsing scheme will constrain which structures can be searched for. That scheme may make too many or too few distinctions, or the wrong distinctions, relative to the structures that the linguist would like to identify as hits.

**The World Wide Web as a Corpus**

In this section, I discuss the properties of the (World Wide) Web when used as a linguistic corpus, in comparison with the sorts of corpora discussed in Section *Traditional Corpora*. One thing that is neither pro nor con but is unique to the Web is that one cannot meaningfully ask whether it is a representative sample of some larger body of language. That is because it is not a sample or subset of any such body (though it obviously contains samples of some larger bodies, e.g., newspaper text). Much of the material on the Web exists nowhere else, and much of it is written in various ‘codes’ that are used nowhere else. Some subset of it blurs the line between traditional spoken and written language in the sense that, while it is generated from a computer keyboard, it is part of a nearly real-time conversation and undergoes minimal editing or self-correction. Of course, language on the Web is still emanating from the same human minds as the language we encounter outside the Web, so it is surely relevant to the study of the latter, and can likely provide new kinds of evidence for that enterprise. But when it comes to fine details, one cannot assume that language on the Web is equivalent to spoken language produced in real time, nor to published written text—the two major sources of traditional corpora; the Web demands its own methods of analysis.

**Benefits**

Relative to traditional corpora, the greatest advantage to using the Web is that it is many orders of magnitude larger. All else equal, a larger corpus is always preferable, but in this case many things are not equal. There are some other benefits that the Web has over at least some linguist-assembled corpora: it can be searched for free (at least for now), searches are run very quickly (this being a point of commercial competition), it is stylistically more varied, it is more representative of the population of speakers of a given language (because the written content of traditional corpora comes mainly from professional writers), it tracks changes in the language with virtually no delay, and it contains plenty of unedited text. Although most Web searching to date has been string based, tools for structure-based searching are being developed.\(^{13}\)

**Drawbacks**

We can divide the problems with Web data into the qualitative and the quantitative. Beginning with the
quantitative, the overarching problem is that the true nature and source of the linguistic material that one finds there cannot be determined with certainty, and for most of the search ‘hits’ that contribute to the reported number of results it cannot be determined at all, at least using existing commercial search engines, because they provide links to only a fixed, small number of hits for each search. In particular, determining who originally created the contents of a given (part of a) webpage is often impossible, which makes it impossible to know what language and dialect that person was a native speaker of. (In fact, the contents might have been created by machine translation, voice recognition, or optical character recognition, each of which adds its own kinds of distortions.) Concomitantly, one cannot usually determine how many different speakers/writers contributed to a set of hits, which makes the interpretation of rare forms all the more difficult. The inability to individuate data by speakers also makes it impossible to ask vital linguistic questions like ‘Do speakers who accept X also accept Y?’ When errors are found, they could reflect typing slips as well as language productions slips (and see Ref 14 for the suggestion that the former may have been misanalyzed as a linguistic innovation by Manning15). On this question of errors, Kilgarriff and Grefenstette16 assert that ‘the Web is a dirty corpus, but expected usage is much more frequent than what might be considered noise’ (p. 342), citing hit counts for a particular phenomenon in Spanish in which ‘the correct form is always at least 500 times more common than any potentially incorrect form’. Curiously, though, the ‘incorrect’ forms they searched for to illustrate this generalization instantiate a proscribed hypercorrected construction so well-known that there is a name for it (deque´ismo); these are forms that linguists treat as attested.17

Let us turn now to problems in interpreting quantitative data from Web searches, data which linguists have used to argue, for example, that more frequent structures violate fewer grammatical constraints than less frequent ones. The overarching problem is that commercial search engines use proprietary algorithms, so one can never be sure how the numbers of hits are arrived at—this cannot be done by exhaustive counting, given the size of the Web: extrapolations based on sampling (non-random, apparently18) must be involved. The heuristics being used sometimes yield surprisingly large internal inconsistencies.19 Furthermore, these algorithms are continually being refined, and the Web is continually growing, meaning that counts are not replicable across time. Another problem is not knowing how many of the hits are genuinely distinct, rather than resulting from multiple copies of the same content.

**ACCEPTABILITY JUDGMENT DATA**

Acceptability judgments involve explicitly asking speakers (using a range of possible wordings) to judge (i.e., report their spontaneous reaction as to) whether a particular string of words is a possible utterance of their language, with an intended interpretation either implied or explicitly stated. (It is not known exactly what speakers do in response to such questions, nor whether different ways of asking for judgments engage different sorts of judgments; see Ref 20 for preliminary data.) Acceptability judgments have taken on particular importance in the study of language since the emergence of generative grammar,1,21 due to the value placed in this framework on delimiting the impossible forms of a given language (and language in general) (Box 1).

**BOX 1**

**THE IMPORTANCE OF NEGATIVE EVIDENCE**

The fact that speakers widely agree on what is not possible in their language is taken as particularly significant by generative linguists, because most language acquisition researchers now agree that children do not receive systematic negative evidence in the course of learning their native language(s), that is, they are not provided with sentences that are ill-formed, identified as such (e.g., grammatical corrections). This lack of negative evidence, combined with the observation that acquisition seems to entail the ability to distinguish well-formed from ill-formed utterances among those that have never been encountered before, is taken to imply that children must have prior knowledge (encoded somehow in their brains by virtue of the genotype) that constrains their view of what is possible in human languages.

In the following two sections I discuss benefits and drawbacks that are more or less inherent in judgment data, regardless of exactly how they are elicited or analyzed. In Section *Informal Versus Formal Judgment Collection*, I address choices of how to elicit judgment data.

**Benefits**

For information about what the ill-formed expressions of a language are, acceptability judgments are superior to other available sources of linguistic evidence. As
discussed in Section Corpus Data, no amount of spontaneously occurring language can definitively answer that question. There are also no known brain measures that are sensitive to all and only the ungrammatical sentences, and no alternative behavioral measures. In addition, judgments can provide immediate information about the status of any string a researcher happens to be interested in—there is no guarantee that such a string, even if grammatical, will occur in any body of spontaneous language within any finite amount of time. A further advantage of judgment data over spontaneous usage data is that the latter will include some proportion of production errors (slips of the tongue or pen, etc.), the vast majority of which will be judged as ill-formed even by the very speakers who produced them, and which therefore (on most views of the competence–performance distinction) should not be generated by the grammar. (See Ref 22, pp. 62–63 for additional comparison of acceptability judgments to alternative sources of evidence.)

There are also practical reasons why many linguists have made extensive use of acceptability judgments as a data source. Judgment data can be elicited from a small number of speakers in a matter of minutes at little or no cost, allowing for relatively rapid growth of the empirical base for grammatical theory, and they can be readily obtained in language communities where the use of expensive laboratory equipment is infeasible, where large numbers of speakers do not exist, or for which there are no corpora available. In light of all of these considerations, the increasingly common suggestion that acceptability judgments should be eliminated as a source of evidence for linguistics would be not only counterproductive, but in fact lethal to the field’s progress.

Drawbacks

There are potential confounds in the collection of acceptability judgments that should be taken into account when using them in studying linguistic competence. Perhaps the most obvious are certain aspects of the meaning of the sentence, such as its truth or plausibility in the real world, that most generative linguists take to be orthogonal to grammatical well-formedness. (It should be emphasized that many aspects of meaning are involved in grammatical well-formedness—see Box 2.) For instance, naive speakers might rate a well-formed expression of improbable content (e.g., My guppy admires illness) less acceptable than an ill-formed expression of an obvious, sensible message (“Whose did you admire necklace?”), which a syntactician would view as a confounded response pattern if syntactic well-formedness was what she was trying to measure (Box 2). That guppies are probably incapable of feeling admiration and that illness is an odd thing to admire are facts about the world, not about language; in contrast, the fact that whose cannot undergo WH-movement unaccompanied by the noun it modifies is a fact about (the English) language. Grammatical theory is generally held accountable for only the second fact, but acceptability judgments may reflect both facts to unpredictable degrees. As a result, whenever possible the plausibility of the (intended) meanings should be matched in critical comparisons. The same holds for the actual truth value, when applicable: children (and some adults) base their responses on truth when asked to judge well-formedness.24,25

BOX 2

WELL-FORMEDNESS AND MEANING

Chomsky argued that the syntactic well-formedness of a sentence is independent of its meaning, based on the observation that Colorless green ideas sleep furiously is acceptable in a way that Furiously sleep ideas green colorless is not, although both strings are ‘nonsensical’. (They were also both claimed to have zero prior probability of occurrence, cf. Section Grammars and Grammaticality.) The same is often asserted based on so-called Jabberwocky sentences such as Every wug gorped a blick, which most speakers find more acceptable than Wug every blick a gorped, although both are purportedly meaningless. These claims are misleading, however. The closed-class meanings in the well-formed strings are fully interpreted when they are encountered: in the Every wug sentence, we know that a relation called gorp existed in the past between each member of the set of countable entities that fall in the category wug and at least one member of the set of countable entities that fall in the category blick, and that it might have been the same blick for all the wugs or a different blick for each wug (i.e., there is a scope ambiguity). It is only the open-class aspects of the meaning (what sorts of things wugs and blicks are, what gorping is) that are unknown. English speakers know the sentence cannot be synonymous with Every wug was gorped by a blick, for example, no matter what the open-class words mean. An utterance will generally not be judged acceptable if its word order and closed-class morphemes cannot be given a consistent interpretation: for instance, Every wug were gorped is unacceptable because of a number mismatch.
(For this reason, language acquisition researchers have developed fairly complex methodologies for eliciting judgments from children.)

Another potential confound (when using judgments to study linguistic competence) is parseability, if we adopt the standard assumption that the grammar is not identical to the parser. (One can of course use judgment data to study parsing, in which case parseability would be an independent variable rather than a confound.) A perception of ill-formedness may arise due to properties of the comprehension process that are independent of grammatical knowledge, although exactly how to make the distinction is not evident a priori. For example, linguists disagree about the grammatical status of clausal subject embeddings like the one in (8), which are claimed to be ungrammatical by Koster and Stowell but grammatical (though hard to parse) by Delahunty.

8. If [that John likes Mary] surprises you, you obviously haven’t been paying much attention.

More dramatically, so-called garden path sentences such as (9) are often judged unacceptable because the parser cannot ‘find’ any grammatical parse of the string, in this example because of the parser’s propensity to (mis)interpret the verb raced as a main clause active verb.

9. The horse raced past the barn fell.

This apparent unacceptability disappears with a verb whose inflection is not ambiguous between past tense and passive participle:

10. The horse ridden past the barn fell.

Sentence (11) is a classic center-embedded structure in which a relative clause is embedded within another relative clause, which in turn is embedded in the middle of the main clause. Chomsky and Miller argue that this structure is licensed by the grammatical rules of English, but that (11) is perceived as unacceptable because it cannot be successfully parsed. Sentence (12), on the other hand, violates basic principles of syntax: there are too many subjects for the number of verb phrases. However, in controlled experiments, Gibson and Thomas found that speakers rated structures like (12) just as acceptable as ones like (11), and related work suggests that (12) itself may, by virtue of the particular lexical content, sound more acceptable than (11). This suggests that, although the parser evidently fails on both (11) and (12), the omission of the second verb phrase from (12) makes that failure less salient, perhaps because the need for three verb phrases cannot be kept track of.

Another case of a structure that almost all linguists agree is ungrammatical is (13) (attributed to the speech of Herman Schultze).

13. More people have been to Russia than I have.

Example (13) often sounds fine on first hearing, an opinion that changes after it is pointed out that this string has no possible meaning. This illusion may be due to the fact that (13) contains two parts, each of which occurs as a substring of other (grammatical) sentences.

Given these various challenges (see also Section Other Experimental Data), the question has arisen whether it is possible to obtain useful judgments on subtle issues from speakers with no linguistic training. In some cases, it is difficult to replicate relatively agreed-upon judgments of linguists when testing naive speakers. Partly for this reason, linguists often rely on other linguists for judgment data. It is not clear whether the net effect of doing so is to the field’s advantage (cf. Ref 39). Valian makes a case in favor of using such ‘expert’ judgments, based on an analogy to wine tasting, which relies on the acquired ability to detect subtle distinctions that inexperienced wine drinkers simply cannot make. On the other hand, it is possible that linguistic training induces systematic biases in one’s judgments. One practice that is clearly undesirable, however, is for researchers to present their own judgments as primary evidence. There is no reason to believe that people with a stake in the outcome of such judgments can remain unaffected by their theoretical stance.
Informal Versus Formal Judgment Collection

Although the elicitation of acceptability judgments is itself a behavioral experiment—the speaker is asked for a voluntary response to a stimulus—the majority of judgments that have been collected over the past 50 years have relied on relatively impoverished response options (such as just ‘acceptable’ or ‘unacceptable’), relatively few speakers, relatively few tokens of the structures of interest, and relatively informal data analysis. Recent work has investigated whether this informal data collection process may be profitably augmented by more rigorous experimental techniques: finer-grained response scales, larger sample sizes ($N > 20$), more tokens, and formal statistics such as ANOVA and linear mixed effects regression. I briefly discuss the latter three suggestions as a group, and then return to the first in more detail.

Methodological Rigor

Any experimental psychologist would take it as self-evident that when collecting behavioral data one should always use more than a couple of participants and more than one item per condition, and a psychology journal would not normally allow results to be published without formal statistical analysis. If generative linguistics is conceptually a subfield of cognitive psychology, why have these desiderata not been self-evident to linguists? Part of the answer is probably that acceptability judgments have often struck linguists as the counterpart of what psychologists sometimes call ‘sledgehammer effects’, which are so large and obvious that ‘you hardly need to do the experiment’ because it is so clear from one’s everyday experience how the result will come out. This seems to be the position of Phillips. Even if this were true, one might ask what harm there is in doing the experiment rigorously, just to be safe. To this a linguist might respond, as Phillips does, that it would be a waste of time and resources that would be better invested in exploring the theoretical consequences of the observation. Moreover, the linguist might continue, the potential danger to the field of reporting an observation that could turn out to be false is minimal, because in the process of peer review there are bound to be skeptics eager to point out that they disagree with the data or that there is some quirk in an example that would not generalize to all tokens of the intended type.

These arguments are not without merit, but they are not airtight. Of the three dimensions mentioned above on which linguistic methodology is arguably lax, the small number of participants is the most worrisome, because large amounts of variation are known to exist among speakers of the same language/dialect, and there is no reason to think that individuals, even linguists, have a good sense of whether their own judgments are shared uniformly among some population—indeed, most linguists have repeatedly had the experience of being surprised to discover that a judgment that seemed extremely clear to them was not shared by everyone (or even anyone) in the room. However, not all interspeaker differences in responses necessarily reflect genuine differences in judgments. An important finding stressed by Cowart is that while individuals may differ widely in how they use a given response scale, so that a particular sentence might be rated 5 out of 7 by one participant and 3 out of 7 by another, for instance, there is much less individual variation in the patterns of responses, that is, how sentences rate relative to one another. Cowart also observes a large amount of random trial-to-trial variation in acceptability judgments within individual participants, in the midst of a very reliable pattern of results across a whole experiment. Using a substantial number of tokens per condition is thus useful for factoring out such random noise, and potentially systematic quirks of individual items.

Response Scales and Judgment Tasks

Linguists have long been sensitive to the fact that many sentences seem to have a status somewhere between fully acceptable and fully unacceptable, as mentioned in Section Grammars and Grammaticality. A question that has been receiving increasing attention is what this implies about the nature of grammar itself (cf. Section Gradience in Acceptability and Grammaticality) and what the best way is to collect such judgments. For a long time, psycholinguists (and some theoretical linguists) have used five-or seven-point Likert scales of the type familiar from survey research. However, over the past decade, magnitude estimation has been advocated as a new standard among acceptability judgment tasks (Box 3). Bard et al. and Keller argue that this task is superior to the standard Likert scale task, for at least two reasons. First, it allows participants to distinguish as many levels of acceptability as they can perceive, unlike tasks in which they are limited to five or seven choices. Second, the intervals between levels of acceptability are measured in uniform units calibrated by the acceptability of the reference sentence, known as the ‘standard’ in psychophysics parlance. (In contrast, the intervals between numbered points on a Likert scale, although numerically equal, may in fact not be treated as equal by respondents, particularly if they are associated with vague labels such as ‘somewhat ungrammatical’, ‘extremely ungrammatical’, etc.) Advocates of magnitude estimation argue
HOW MAGNITUDE ESTIMATION WORKS

In magnitude estimation, participants rate the acceptability of a target sentence using the acceptability of a reference (or standard) sentence as a point of comparison, e.g., (1).

1. Reference: What do you wonder whether Mary bought?
   Acceptability: 100

Target Item: What did Lisa meet the man that bought?
Acceptability: ____

In this example, the reference has been assigned an arbitrary rating of 100, referred to as the modulus. The participant must then rate the target sentence in terms of the scale implied by this modulus, using any positive real number. Thus, if the target feels half as acceptable as the reference, it should be rated 50; if it feels twice as acceptable, 200. (Even proponents of the technique concede that this way of talking—treating the response scale as a ratio scale—does not make scientific sense: although one can speak of total absence of light in a psychophysics experiment, the notion of total absence of grammaticality for a string of words is hard to fathom.)

that these two properties combine to increase precision (the reproducibility of the measurement). Let us consider each of these purported advantages in turn.

First, although the availability of any positive real number as a response would in theory allow participants to rate every stimulus differently, in practice this is not what they do. Rather, they use a small set of (typically whole) numbers repeatedly, and (many or all of) the members of that set often stand in a salient relationship to one another that does not seem to depend on the stimuli (e.g., multiples of 5 or 10). Second, regarding the claim that the standard/reference sentence allows participants to calibrate a uniform scale, Sprouse investigated this by systematically varying the acceptability of the standard (i.e., changing the sentence) across four samples of speakers while holding the target sentences constant. The absolute ratings of the test sentences did not change as the acceptability of the standard varied, suggesting that the standard does not play any role in the rating of the target materials. Thus, it is unclear how participants are calibrating their responses, but it cannot be on the basis that has been claimed. Finally, concerning the claim of increased precision, Weskott and Fanselow compared magnitude estimation to a standard seven-point scale task using identical experimental materials. They found that the seven-point scale led to significantly less unexplained variance, suggesting that it may actually be more precise than magnitude estimation. [They, along with Myers, have also demonstrated that, with a sufficient number of observations, the proportions of responses in even binary (acceptable/unacceptable) response data can be used to detect small acceptability differences between conditions.] When combined with the common observation that magnitude estimation is less intuitive for many participants than traditional scales (and hence more time-consuming and labor-intensive for experimenters), these results make it unclear what the real advantages would be of adopting magnitude estimation in place of simpler alternatives (cf. also Ref 53; Box 4).

HYBRID TASKS

Some researchers have proposed new tasks that are intended to combine the intuitive nature of point scales with the sensitivity of magnitude estimation. For example, Featherston has proposed a ‘thermometer task’ in which participants are given two reference sentences with associated acceptability values, such as 20 and 40 (analogous to freezing and boiling points). They can then choose values for target sentences along the real number line relative to those two points by treating it as a linear scale: for example, a target that is halfway between the acceptability of the two reference sentences would be rated 30.

Gradience in Acceptability and Grammaticality

The freedom provided by magnitude estimation to distinguish a theoretically infinite number of levels of acceptability and the ability to quantify the distances between those levels have been a catalyst for some researchers to replace a ‘categorical’ model of grammar in which there are two distinct categories, grammatical and ungrammatical (possibly with distinctions among the latter), with a gradient model of grammar in which grammaticality is a continuous property. Although it is not surprising that magnitude estimation yields continuous acceptability values, what is nontrivial is that respondents are consistent in their use of the intermediate levels of acceptability, suggesting that it is indeed tapping into a robust cognitive system that yields gradient results.
The question is whether those gradient results are a reflection of grammatical knowledge on its own or grammatical knowledge in combination with factors that affect language processing, decision making, and so on, and are already known to display gradient behavior (working memory load, semantic plausibility, lexical and syntactic frequency, prototypicality, etc.) (cf. Ref 57).

We know that people can give systematic gradient judgments about virtually any concept, including concepts such as even number whose only meaning is formally defined as categorical. 58,59 This observation does not entail that our knowledge of mathematics fails to make a perfectly sharp distinction between even and odd numbers. Rather, our judgments can evidently be sensitive to factors other than our underlying competence. Putting it another way, when asked for gradient responses, participants will find some way to oblige the experimenter; if doing so is incompatible with the experimenter’s actual question, they apparently infer that she must have really intended to ask something slightly different. For instance, instead of rating the extent to which some number is even, they may (not necessarily consciously) reinterpret the task as seeking a rating of how representative or typical the properties of a particular number are as compared to identical positions (indicated by the underscore following for). Also, both involve a parsing difficulty: due to the Active Filler Strategy (cf. Refs 64 and 65), the parser initially hypothesizes that the WH-phrase which general has moved from the first underscore position (immediately following killed); reanalysis will therefore be required in both sentences.

Two Examples of Gradient Acceptability
An emerging benefit of the use of gradient acceptability measures has been an increased awareness of differences among sentences that had been described in the literature as fully acceptable. For example, Featherston60 demonstrates that even though native German speakers report that (14a) and (14b) are both acceptable,61 one can detect that (14b) is actually significantly less acceptable than (14a) using magnitude estimation.

14. (a) Wer hat was gelesen?
   who has what read
   ‘Who has read what?’
   (b) Was hat wer gelesen?
   what has who read
   ‘What has who read?’

This acceptability contrast parallels that between the equivalent English sentences [the translations in (14)], except that the English version of (14b) has been considered unacceptable, which prompted Chomsky62 to propose the Superiority Condition to rule it out. Featherston’s interpretation of his data is that some grammatical constraints may be operative in a weak form in languages where they were previously thought not to be operative; this analysis presupposes a gradient view of grammar. But the possibility that the degraded status of some sentences could be due to parsing factors rather than a grammatical constraint has not been ruled out. (See Ref 63 for a detailed argument of this kind concerning certain island constraint violations such as *What did Mary wonder whether John bought?*

A second strand of research involving comparisons among grammatical sentence types has found that temporary processing difficulties that arise during the parsing of what appear to be fully well-formed sentences can lead to differences in acceptability ratings. (This is a different scenario from the type in (9), where the processing difficulty is typically not resolved on first encounter.) For example, sentences (15) and (16) are equally grammatical according to all traditional syntactic theories, because their WH-extractions are from identical positions (indicated by the underscore following for). Also, both involve a parsing difficulty: due to the Active Filler Strategy (cf. Refs 64 and 65), the parser initially hypothesizes that the WH-phrase which general has moved from the first underscore position (immediately following killed); reanalysis will therefore be required in both sentences.

15. John wondered [which general] the soldier killed the enemy effectively and enthusiastically for ____ during the war in Korea.

16. John wondered [which general] the soldier killed (____) effectively and enthusiastically for ____ during the war in Korea.

Sentences like (15) are rated significantly less acceptable than ones like (16),66 and this cannot be explained by even a gradient theory of grammar—no known syntactic constraint distinguishes them. (There could be a lexical constraint favoring transitive over intransitive uses of kill, but that would predict (15) should be more acceptable, not less acceptable.) A plausible processing explanation is available in terms of the kinds of information from which the parser can deduce that reanalysis is required. Specifically, in (15) it encounters a noun phrase (the enemy) at a point where it is not expecting one (because its working hypothesis is that which general has moved from there), whereas in (16) it fails to encounter a noun phrase at a point where it is expecting one (after for, because it is still pursuing the hypothesis that which general has moved from the
parenthesized underscore position). Why the former situation is the more disruptive should be explained by a (gradient) theory of reanalysis. Together, this and the previous example illustrate approaches that posit a source for gradience in acceptability that lies outside the grammar proper.

**OTHER EXPERIMENTAL DATA**

**Motivations**

It has sometimes been suggested that claims made exclusively on the basis of acceptability judgment data do not necessarily bear on how the human language faculty is actually constructed unless their ‘psycho-logical reality’ has been tested via some experimental procedure using another dependent measure such as time, error rate, electrophysiological response, and so on. This view belies a misunderstanding: acceptability judgments are themselves data about human behavior and cognition that need to be accounted for; they are not intrinsically less informative than, say, reaction time measures—many linguists would argue that they are more informative.

A potentially more serious objection to judgment data is that they are not ‘real’ because they demand awareness of language as an object of attention and evaluation (‘metalinguistic awareness’), rather than in some more typical use. This is claimed to make them artificial and undermine their external validity. One response is simply to note that to the extent that acceptability judgments are systematic, it is unlikely that their underlying cognitive source would be something unrelated to grammar—the existence of a separate system that duplicates many aspects of language in use but is invoked only for metalinguistic tasks seems highly improbable (Box 5).

Nonetheless, any single kind of evidence has its flaws, so experiments that measure something other than explicit ratings—most often, reaction times or brain activity—certainly have their place in the empirical base of linguistics. They can be used as a source of converging evidence for claims made on the basis of judgment and/or corpus data, and they can sometimes be used as a source of decisive evidence on questions for which other kinds of data are equivocal.

**Examples**

The history of the use of laboratory measures to confirm the claims of generative grammar is almost as old as the field itself, and space considerations prohibit a comprehensive survey. The examples mentioned in this section have been chosen because of their historical importance or their relevance to general points that are being illustrated; no position is being taken on the current significance of the methods or results. Some of the earliest attempts involved subjective location of clicks as evidence for constituent boundaries and the application of the Derivational Theory of Complexity to reaction times as evidence for the application of transformations (see Ref 73, p. 481 for an early discussion; see Ref 74 for criticism and Ref 75 for a contemporary review). Another large research enterprise, which began in the 1980s, involved (dis)confirming the existence of the various types of so-called empty categories posited in Government-Binding theory (see Ref 76; and ‘Government-binding/principles and parameters theory’ in http://wires.wiley.com/WileyCDA/WiresArticle/wisId-WCS35.html) using priming paradigms (cf. Refs 77 and 78 for reviews and new experiments).

**BOX 5**

**THE LIMITS OF METALINGUISTIC AWARENESS**

Admittedly, cases have been documented where speakers’ behavior when consciously reflecting on their language seems sharply at odds with their subconscious linguistic knowledge. For instance, Labov has documented that speakers who denied that they ever produce a given construction (positive anymore, cf. Section Grammars and Grammaticality) or even knew what it meant went on to use it spontaneously during the very same interview; there is no evidence that this is the result of sociolinguistic stigmatization. But such cases are extraordinarily rare.

The more challenging variety of experiment to design is one that goes beyond issues for which traditional data have (so far) been able to tip the scales substantially. Most examples involve brain measures and take advantage of the presence of qualitatively different response patterns [most notably, ‘components’ in event-related potentials (ERPs) and event-related fields (ERFs)], rather than unidimensional behavioral measures such as reaction time. For example, an early ERP study, whose aims were more ambitious than many that have followed, found that unacceptable sentences that were analyzed as violating different constraints in Government-Binding theory, such as (17)–(19), elicited different ERP patterns, a finding that was taken to support the Government-Binding analysis over alternatives under which these ungrammaticalities might not be divvied up in the same way.
17. *What was a sketch of admired by the man? [Subjacency violation]

18. *What did the man admire Don’s sketch of? [Specificity violation]

19. *The man admired Don’s of sketch the landscape. [Phrase structure violation]

Within each of these conditions every item had the same syntactic form, differing only in lexical content; this was necessitated by the practicalities of ERP data analysis. In principle, the logic of the experiment makes additional predictions, which apparently have not been tested: namely, that all sentences violating a particular constraint would have an ERP signature in common. For example, one would expect sentences like those in (20) to pattern with (17).

20. (a) *Who did the boss forget when he was supposed to interview? [Subjacency violation]

(b) *Who did the boss make the claim that he had already interviewed? [Subjacency violation]

Until this is established, it is difficult to exclude the possibility that incidental properties on which strings like (17), (18), and (19) happen to differ could be responsible for their different ERP signatures; the existence of a general signature of subjacency violations, for example, has not been established.

More recent work has endeavored to find evidence for the same neurophysiological response to different syntactic configurations in order to associate it with an abstract grammatical construct. In the illustrative example here, the construct is a semantic one: ‘coercion’, a process whereby the basic semantic type of an expression must be changed in order to fit the linguistic context. Pylkkänen and colleagues have found evidence that this process is at work in three different syntactic configurations. In one, Brennan and Pylkkänen80 considered extant semantic analyses of the resolution of a mismatch between the lexically specified aspectual class (‘Aktionsart’) of a verb and the aspectual type required by the meaning of the sentence as a whole. For example, the verb sneeze often denotes a single instance of the event of sneezing (a punctual reading), as in (21a); however, some sentences require it to refer to repeated events (an iterative reading), as in (21b):

21. (a) After 20 minutes, the student sneezed in the back of the classroom. [Punctual]

(b) Throughout the day, the student sneezed in the back of the classroom. [Iterative]

The literature had debated whether the aspect of sneeze is (1) lexically specified as punctual, such that the iterative reading in (21b) must be derived through coercion; (2) lexically specified as iterative, such that the punctual reading in (21a) must be derived through coercion; or (3) underspecified, such that neither reading would require coercion. Brennan and Pylkkänen’s reaction time data provided evidence favoring possibility (1) over alternatives (2) and (3); their magnetoencephalography (MEG) data supported this conclusion, finding that sentences like (21b) showed an increase in amplitude of a component known as the anterior midline field (AMF), relative to sentences like (21a).

In two other, superficially quite dissimilar, syntactic contexts, the AMF had already been suggested to be a reflex of processes such as coercion. These were cases of complement coercion,81,82 which involves a mismatch between a noun phrase that usually refers to an entity and a verb (22) or adjective (23) that needs to semantically combine with an event.

22. (a) The journalist wrote the article after his coffee break. [Regular meaning]

(b) The journalist began the article after his coffee break. [Coercion]

23. (a) The nimble climber imagined the fall survivable even though the experts disagreed. [Regular meaning]

(b) The nimble climber imagined the ice survivable even though others did not. [Coercion]

Although in (22a) an article is the kind of thing one normally writes, one normally begins doing something, i.e., an activity, and an article is not an activity, so some activity must be added by inference in (22b), e.g., ‘writing the article’. Similarly, unlike the fall in (23a), the ice in (23b) does not describe an event until the comprehender supplies additional meaning, e.g., ‘his crossing the ice’. Finding the AMF in sentences like (21b), (22b), and (23b) supports the claim that a common operation is involved in all three instances of semantic mismatch, independent of their syntactic form. Importantly, most semantic theories, along with behavioral evidence, have placed the above coercion cases in a class with other cases of ‘type-shifting’ that do not elicit the AMF component,33 so the MEG data provide evidence for a nonobvious qualitative distinction among them.
CONCLUSION

As is often true for science and particularly for cognitive science, there is no single source of empirical evidence that is ideal in all respects for the study of (knowledge of) language. Any particular technique might provide the best ratio of benefits to drawbacks in a specific linguistic research situation, but the strongest results will emerge when the application of multiple research methods points to the same conclusion.

REFERENCES


FURTHER READING
Gerken L, Bever TG. Linguistic intuitions are the result of interactions between perceptual processes and linguistic universals. Cogn Sci 1986, 10:457–476.