Book Review

The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology, Carson T. Schütze

Frank Keller
Centre for Cognitive Science, University of Edinburgh
2 Buccleuch Place, Edinburgh EH8 9LW, United Kingdom
keller@cogsci.ed.ac.uk

The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methology, Carson T. Schütze, Chicago: University of Chicago Press, 1996. Price: $28.95/£23.25 (hardback), xvi + 238, index, ISBN: 0-226-74154-0.

1. Introduction

The data on which linguists base their theories typically consist of grammaticality judgments, i.e., intuitive judgments of the well-formedness of utterances in a given language. When a linguist obtains a grammaticality judgment, he or she performs a small experiment on a native speaker; the resulting data are behavioral data in the same way as other measurements of linguistic performance (e.g., the reaction time data used in psycholinguistics). However, in contrast to experimental psychologists, linguists are generally not concerned with methodological issues, and typically none of the standard experimental controls are imposed in collecting data for linguistic theory. Carson T. Schütze's *The Empirical Base of Linguistics* aims to show that such methodological negligence can seriously compromise the data obtained, and argues for a more reliable mode of data elicitation in linguistics, based on standard methods from experimental psychology.

Schütze reviews the literature on linguistic judgments and identifies a set of factors that influence the judgment process, and hence have to be controlled for when collecting linguistic data. Schütze aims to identify parallels between linguistic judgment behavior and other types of cognitive behavior, an approach that allows him to arrive at a model of the judgment process that explains linguistic intuitions as the result of the interaction of the language faculty with other cognitive faculties. Based on this model, Schütze puts forward a set of practical recommendations for eliciting more reliable linguistic data.

Journal of Logic, Language and Information 8: 1, 114–121, 1999.

2. Grammaticality judgments and linguistic theory

Chapter 1 provides a general motivation for studying the empirical properties of linguistic data: theoretical linguists typically collect grammaticality judgment data in a naive, informal way. Psycholinguistic findings, on the other hand, show that grammaticality judgments are subject to a considerable number of biases, for which a naive approach to judgment collection fails to control. The central question is therefore one of data validity: "[i]n the absence of anything approaching a rigorous methodology, we must seriously question whether the data gathered in this way are at all meaningful or useful to the linguistic enterprise" (p. 5).

The details of this problem are fleshed out in Chapter 2, where Schütze analyzes how linguists typically make use of grammaticality judgments. He points out that the difficulties with naive data collection are amplified by the fact that current linguistic research does not confine itself to cases of clear acceptability or unacceptability, but makes crucial use of subtle (and potentially controversial) judgments: "The days are over when linguistics had more than enough to worry about with uncontroversial, commonplace judgment data, and the sophisticated and complex judgments now in use by theoreticians assume much about human abilities that remains unproven, even unscrutinized" (p. 9). As a case study, Schütze discusses the use of subtle judgments in the widely cited articles by Aoun et al. (1987), Belletti and Rizzi (1988), and Lasnik and Saito (1984). Belletti and Rizzi (1988), for instance, make extensive use of relative grammaticality judgments, de facto employing a seven point scale for acceptability. However, no attempt is made to establish whether native speakers can reliably provide judgments of this granularity.

Chapter 2 also provides an outline of the historical and theoretical background of Schütze's study. The theoretical framework in which he operates is Chomsky's (1965: 10) classical competence/performance model: competence pertains to the knowledge of language a speaker has, whereas performance pertains to how this knowledge is put to use. In this terminology, the central concern of Schütze's book is: how can we use performance data (e.g., grammaticality judgments) to investigate linguistic competence? One has to bear in mind here that experimental data as such are not sufficient to determine the grammaticality status of a sentence. For theoretical reasons, a linguist might want to assume that certain sentences are grammatical, even though they are not accepted by native speakers: given that a set of examples is clearly grammatical, it can be concluded that other, structurally related examples should also be generated by the grammar. In such cases, the linguist's intuition about what grammars look like is more relevant than the native speaker's intuition about acceptability.

However, the assumptions about grammaticality may vary from one theoretical framework to another, which raises the problem of the immunization of theories. Schütze identifies three strategies for protecting a theory from data that seems to falsify it:

1. Dispute the validity of the data, e.g., claim that certain sentences are not really acceptable (or unacceptable).
2. Claim that the data is not relevant to the theoretical issue at hand, e.g., stipulate that some other part of the grammar accounts for it.
3. Claim that the data is correctly accounted for by the theory, but the judgments do not reflect this, e.g., due to extragrammatical factors that cause a grammatical (or ungrammatical) sentence to be unacceptable (or acceptable).

Psycholinguistic experimentation allows an evaluation of the validity of strategies 1 and 3: "this is precisely why we *should* strive for a better understanding of acceptability judgments. It would allow us a *principled* way to establish to what extent any such piece of evidence should be considered to bear on the grammar. We will still not be able to draw direct conclusions from such data, but it will at least be a matter of objective fact what the relevant data are" (p. 30).

3. Factors influencing grammaticality judgments

"A great deal is known about the instability and unreliability of judgments" (p. 1), and Schütze devotes Chapters 3–5 of his book to reviewing the linguistic and non-linguistic factors that might influence judgment behavior and hence engender such instability and unreliability.[1] Based on this review, he then proposes a model of the judgment process and formulates a set of recommendations for collecting more reliable data.

3.1. Measurement scales

Chapter 3 includes a discussion of measurement scales: if grammaticality judgments are to be considered empirical data in the sense of experimental psychology, then the measurement scale used for judgment elicitation is of crucial importance, as it determines what type of

[1] For reviews of the use of grammaticality judgments in second language research see Birdsong, 1989 and Chaudron, 1983.

data is obtained and which mathematical (statistical) operations can be carried out on the data.

A *nominal scale* consists of a set of category labels representing the possible values of the property to be measured. The categories are discrete and the only formal relation defined on categories is equality: two stimuli can be compared as to whether or not they fall into the same category with respect to a given property. No ordering relation is defined for a nominal scale, and the only mathematical operation that can be performed is counting. Hence the statistics for nominal scales has to be carried out on category frequencies. Traditionally, linguistic examples are assigned labels like "acceptable" and "unacceptable", i.e., they are measured on a nominal scale.

An *ordinal scale* has the same properties as a nominal scale, and in addition, an ordering relation is defined over the categories: two stimuli can be compared in terms of their rank on the scale with respect to the measured property. However, no commitment is made as to the distance of the points on the scale, and again the only mathematical operation defined is counting, allowing for frequency statistics only. Acceptability is measured on an ordinal scale if the traditional binary categories are complemented by intermediate ones. These are typically notated as "?", "??", or "?*", allowing to record gradient acceptability judgments. This practice can be systematized by defining a consistent ordinal scale for acceptability, and much of the experimental literature on linguistic judgments has followed this approach. However, it is unclear "how many meaningful distinctions of levels of acceptability (relative or absolute) can be made" (p. 77), and different experimental studies have used different scales. This lack of agreement is problematic, as using the right scale is crucial for obtaining consistent data: if there are too few levels, then subjects might collapse true distinctions arbitrarily, if there are too many, they might create spurious distinctions.

Just like an ordinal scale, an *interval scale* presupposes an ordering over the measured categories. In addition, a distance relation is defined that specifies the difference between any two points on the scale. Typically, an interval scale is used for properties which can be measured numerically. Admissible mathematical operations include addition and multiplication, which allows for means to be calculated and for parametric statistics to be carried out. Standardly, linguistic data is not measured on an ordinal scale: it is determined whether an example is more or less acceptable than another one, but not how much more or less acceptable it is. Recently, however, a number of researchers have argued that linguistic intuitions should be measured using magnitude estimation, an experimental paradigm that yields judgment data on an interval scale (Bard et al., 1996; Cowart, 1997; Sorace, 1996). The magnitude estimation approach allows to address the problems raised by the use of gradient judgments and other subtle data in linguistic theory.[2]

3.2. Subject-related factors

Individual differences occur in many aspects of human cognitive behavior, and in Chapter 4, Schütze discusses the ones that have been shown to influence grammaticality judgments. A standard example for such an individual factor is field dependence, for which Nagata (1989b) demonstrated an influence on linguistic judgments.[3] Another factor known to influence judgment behavior is handedness: Cowart (1989b) demonstrates effects of familial handedness on judgments of sentences with subjacency violations.

A contentious issue is whether linguists and non-linguists differ in their judgments. Schütze (pp. 113–122) discusses this question in some detail and concludes that the available experimental evidence is not sufficient to establish systematic differences between the judgments of linguists and naive speakers. However, he contends that "we have enough reasons to *expect* [judgments of linguists] to be different that linguists simply ought to be excluded [as informants]" (p. 187). Cowart (1997: 60) seconds: "Although it might be that sustained practice can sharpen an individual's ability to give reliable judgments, there are also reasons to suspect (as has often been suggested) that training can produce some theory-motivated bias." Both authors conclude that only data from naive speakers should be used.

[2] Magnitude estimation (ME) is standardly used in psychophysics to measure judgments of sensory stimuli (Stevens, 1975). It requires subjects to estimate the magnitude of physical stimuli by assigning numerical values proportional to the stimulus magnitude they perceive. Highly stable judgments can be achieved for a whole range of sensory modalities, such as brightness, loudness, or tactile stimulation. ME has been extended to the psychosocial domain (Lodge, 1981), and recently Bard et al. (1996) demonstrated that linguistic judgments can be elicited in the same way as judgments of sensory or social stimuli. ME has been demonstrated to yield reliable and fine-grained measurements of linguistic intuitions and has been applied to a number of linguistic phenomena (see Cowart, 1997 and Sorace, 1996 for an overview).

[3] Field dependence is a concept used in personality assessment and can be measured using several standard tests, such as the embedded figures test. "A field dependent person fuses aspects of the world and experiences it globally, whereas a field independent person is analytical, differentiating information and experiences into components" (p. 177).

3.3. Task-related factors

Schütze's review of subject-related factors influencing grammaticality judgments is followed by a discussion of task-related factors (in Chapter 5). The factor Schütze considers most crucial are the instructions given to the subjects. Judgment experiments typically employ naive subjects, who are likely to be unfamiliar with the linguistic concepts that they are supposed to apply in rating the stimuli. If no definitions for "grammaticality" or "acceptability" are provided, each subject will use his or her own interpretation of these concepts, and the resulting data is likely to be noisy.

In this context, Schütze describes an experiment by Cowart that used two types of instructions for judging the same set of sentences (reported also in Cowart, 1997). The first, "intuitive", set of instructions asked subjects to base their ratings on their own reactions to a sentence, and stressed that there are no right or wrong answers. The second, "prescriptive" set of instructions evoked the scenario of an English professor marking term papers, and required subjects to judge whether a sentence would be considered right or wrong in such a context. No significant difference was found between the judgments for the two types of instructions, which leads Cowart (1997: 58) to suppose that "informants have very little ability to deliberately adjust the criteria they apply in giving judgments." Schütze concludes that "as long as subjects are given *some* explicit set of instructions, the exact contents of those instructions might not matter a great deal, at least for some classes of sentence types" (p. 133).

Another task-related factor reviewed by Schütze is order of presentation. Order effects were reported by Greenbaum (1976, 1977), and Schütze recommends that "sentence order should be controlled for, either by randomization or counterbalancing" (p. 134). Cowart (1997: 94) agrees and points out that "the informant's state of mind may well change in relevant ways as she proceeds through the [grammaticality judgment] questionnaire. Fatigue, boredom, and response strategies the informant may develop over the course of the experiment can have differing effects on sentences judged at various points in the entire procedure."

Another well-established influence on judgment behavior is repetition. Repetition effects were examined extensively by Nagata (1989a, 1989b, 1989c), whose results show that repetition within a short interval leads to lower grammaticality ratings, while repetition after a long interval (four months) has no significant influence on judgments. Schütze notes that repetition effects also manifest themselves in so-called "linguists' disease", i.e., the phenomenon that one's grammaticality judgments become increasingly blurred and uncertain when one ponders long enough over many examples of the same type.

A rather unexpected effect has been demonstrated by Carroll et al. (1981) and Nagata (1989a): speakers' grammaticality ratings are higher when they face a mirror while making their judgments. Finally, a number of studies have investigated the so-called anchoring effect (Cowart, 1994, Nagata, 1992): if a sentence is judged as part of a set of severely ungrammatical sentences it will receive a higher rating than if it is part of a set of grammatical (or mildly ungrammatical) stimuli.

4. Modeling the judgment process

In Chapter 6, Schütze proposes a model of the judgment process that tries to incorporate most of what is known about the psychological properties of grammaticality judgments. In developing this model, he relies on the assumption that linguistic judgment behavior is not due to a special cognitive component dedicated to linguistic intuition, but rather is the result of an interaction between the language faculty and general properties of the mind. Hence his key claim is that "for any effect on a language (judgment) tasks, there could be an analogous effect on a similar nonlinguistic cognitive (judgment) tasks." (p. 14).

This is certainly a plausible assumption, and seems to be justified by the experimental findings that Schütze reviews. However, the model he proposes provides only a high-level account of the judgment process, as Schütze does not flesh out the interaction of its components in any detail. In particular, the model lacks a precise specification (it is presented only diagrammatically) as well as a computational implementation.[4] Another problem that Schütze (p. 201) acknowledges himself is the absence of experimental data that specifically test his model (as all his evidence is drawn from the existing literature). Therefore he is careful to emphasize the preliminary and speculative nature of this model and concedes that "[m]uch more experimental work is needed before we can begin to have any real confidence in our knowledge about the way the mind works in this regard" (p. 172). It seems that Schütze's model is little more than an attempt to systematize existing experimental findings on judgment behavior, thereby potentially inspiring further research on this subject.

[4] Schütze (pp. 181–183) includes a section on implementational issues, where he suggests an implementation within a spreading activation/parallel processing framework, which however remains largely speculative.

5. Eliciting reliable grammaticality judgments

Grammaticality judgment behavior is influenced by a diverse number of factors, both task-related and subject-related. Unless these factors are properly controlled for, they can introduce a considerable amount of variance into the data, which leads Schütze to urge the use of experimental methods to obtain reliable judgments: “considerable care and effort must be put into the elicitation of grammaticality judgments if we are to stand a chance of getting consistent, meaningful and accurate results” (p. 171).[5]

To minimize potential biases, Schütze suggests a number of basic controls for the design of judgment experiments (in Chapter 6). Firstly, confounds from presentation order should be avoided by counterbalancing or randomizing stimulus presentation across subjects. Also, it is important to use a sufficient number of filler sentences, i.e., to present the experimental items interspersed in a list of sentences that are unrelated to the constructions under investigation. The fillers prevent subjects from becoming aware of the issue the experimenter is interested in (as this might bias their judgments). To avoid anchoring effects, one should make sure that the set of stimuli and fillers does not contain substantially more grammatical than ungrammatical sentences (or vice versa).

To guard against lexical effects, different lexicalizations for each sentence type should be used, and the frequency of the lexical items should be controlled for. Also, Schütze recommends the use of contextualized experimental sentences, as “there are numerous ways that context can influence grammaticality, from bringing out rare word meanings to priming certain parsing procedures” (p. 185). If no context is provided subjects might make up their own contexts, thus potentially increasing inter-subject variance in the ratings. Also, sentences that might trigger processing problems should be excluded from the test materials, as they are likely to confound grammaticality ratings (examples are center-embeddings and garden path sentences).

Once steps have been taken to reduce confounds in the materials, the experimenter has to minimize biases in the procedure of judgment collection. Here, Schütze considers the selection of subjects the most important issue. “If it is the competence of normal native speakers that we claim to be investigating, we need to study random samples of normal native speakers” (pp. 186–187). In particular, linguists should be excluded as informants, as their judgments are likely be confounded by

[5] Note that Schütze does not claim that *all* linguistic data have to be collected under experimental conditions, rather “this will only be required when we have reason to believe that there is disagreement” (p. 211).

theoretical bias. Of course, the number of subjects used has to be large enough so that statistical test can be carried out on the data. Schütze also recommends that potentially relevant individual differences should be recorded on a questionnaire accompanying the experiment, to allow for later analysis for these factors.

Schütze gives no clear recommendation as to the rating scale that should be used. He holds that both relative and absolute ratings can be appropriate, depending on the issue under investigation. Recent studies, however, favor the use of an interval scale based on the magnitude estimation methodology. Magnitude estimation has been shown to yield highly reliable and maximally fine-grained judgment data (Bard et al., 1996; Cowart, 1997; Sorace, 1996), thus avoiding the problems with conventional ordinal scales.

A certain amount of variance will remain in the experimental data, even if all necessary controls are applied. This variance could either be due to chance or could result from an experimental manipulation, i.e., from a factor that the experiment is meant to investigate (such as the violation of a certain grammatical constraint). In the latter case, the effect (e.g., a difference in grammaticality) is significant, in the former case non-significant. The only way of determining the significance of an effect is by performing statistical tests on the data, and so Schütze's most important recommendation the use of statistics, a suggestion that "linguists consistently ignore" (p. 195). This point is particularly important if degrees of grammaticality are used as evidence: mere intuition is not sufficient for determining whether small differences in acceptability are reliable or not (Cowart, 1989a, 1997 demonstrates this point with respect to gradience in extraction from picture NPs).

Schütze (pp. 186–201) also considers the problem of inconsistencies in judgments, i.e., how to interpret disagreements between speakers or changes over time in the ratings of a single speaker. Regarding this issue, Cowart (1997) demonstrates that the overall judgment pattern for a given structure can be highly stable within a group of speakers, while at the same time, the judgments of individual speakers show considerable variance. Cowart concludes that, similar to other types of behavioral data, linguistic judgments seem to exhibit a certain amount of random variance around a stable mean, which he takes as a strong arguments for collecting judgment data experimentally.

6. Conclusion

On the whole, *The Empirical Base of Linguistics* is a valuable guide to the elicitation and use of linguistic judgments. Schütze has produced an impressive survey of the relevant literature, and his volume will certainly serve as a reference work for theoretical linguists and psycholinguists alike. A particular achievement of Schütze's is to make the relevant psychological literature accessible to linguists without experimental background.

By discussing the potential problems with grammaticality judgments, Schütze makes a strong case for the use of experimental methods for eliciting linguistic data. However, the conventional informal approach will probably remain standard for the bulk of linguistic data, in spite of its serious shortcomings (due to practical reasons such as the lack of training and resources for experimental work). But one might hope that Schütze's argument for the use of experimental methods will be followed at least for those phenomena that involve subtle or gradient judgments, where experimentation is essential to obtain reliable data.

On the more practical side, Schütze provides an excellent set of recommendations on how to control for the most serious biases in judgment behavior. For those who want to embark on experimental data collection, however, Schütze's recommendations are not explicit enough; he is mainly concerned with fundamental issues, and providing explicit guidelines for psycholinguistic experimentation is outside the scope of his book. This is a gap that was filled recently by the excellent studies by Bard et al. (1996) and Cowart (1997).

References

Aoun, J., N. Hornstein, D. Lightfoot, and A. Weinberg: 1987, 'Two Types of Locality'. *Linguistic Inquiry* **18**(4), 537–577.

Bard, E. G., D. Robertson, and A. Sorace: 1996, 'Magnitude Estimation of Linguistic Acceptability'. *Language* **72**(1), 32–68.

Belletti, A. and L. Rizzi: 1988, 'Psych-Verbs and θ-Theory'. *Natural Language and Linguistic Theory* **6**(3), 291–352.

Birdsong, D.: 1989, *Metalinguistic Performance and Interlinguistic Competence*. Berlin: Springer.

Carroll, J. M., T. G. Bever, and C. R. Pollack: 1981, 'The Non-Uniqueness of Linguistic Intuitions'. *Language* **57**(2), 368–383.

Chaudron, C.: 1983, 'Research on Metalinguistic Judgments: A Review of Theory, Methods, and Results'. *Language Learning* **33**(3), 343–377.

Chomsky, N.: 1965, *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Cowart, W.: 1989a, 'Illicit Acceptability in *picture* NPs'. In: C. Wiltshire, R. Graczyk, and B. Music (eds.): *Papers from the 25th Annual Meeting of the Chicago Linguistic Society*, Vol. 1: The General Session. Chicago, pp. 27–40.

Cowart, W.: 1989b, 'Notes on the Biology of Syntactic Processing'. *Journal of Psycholinguistic Research* **18**(1), 89–103.

Cowart, W.: 1994, 'Anchoring and Grammar Effects in Judgments of Sentence Acceptability'. *Perceptual and Motor Skills* **79**(3), 1171–1182.

Cowart, W.: 1997, *Experimental Syntax: Applying Objective Methods to Sentence Judgments*. Thousand Oaks, CA: Sage Publications.

Greenbaum, S.: 1976, 'Syntactic Frequency and Acceptability'. *Lingua* **40**(2/3), 99–113.

Greenbaum, S.: 1977, 'Judgments of Syntactic Acceptability and Frequency'. *Studia Linguistica* **31**(2), 83–105.

Lasnik, H. and M. Saito: 1984, 'On the Nature of Proper Government'. *Linguistic Inquiry* **15**(2), 235–289.

Lodge, M.: 1981, *Magnitude Scaling: Quantitative Measurement of Opinions*. Beverley Hills, CA: Sage Publications.

Nagata, H.: 1989a, 'Effect of Repetition on Grammaticality Judgments under Objective and Subjective Self-Awareness Conditions'. *Journal of Psycholinguistic Research* **18**(3), 255–269.

Nagata, H.: 1989b, 'Judgments of Sentence Grammaticality and Field-Dependence of Subjects'. *Perceptual and Motor Skills* **69**(3), 739–747.

Nagata, H.: 1989c, 'Repetition Effects in Judgments of Grammaticality of Sentences: Examination with Ungrammatical Sentences'. *Perceptual and Motor Skills* **68**(1), 275–282.

Nagata, H.: 1992, 'Anchoring Effects in Judging Grammaticality of Sentences'. *Perceptual and Motor Skills* **75**(1), 159–164.

Sorace, A.: 1996, 'The Use of Acceptability Judgments in Second Language Acquisition Research'. In: W. C. Ritchie and T. K. Bhatia (eds.): *Handbook of Second Language Acquisition*. San Diego, CA: Academic Press, pp. 375–409.

Stevens, S. S.: 1975, *Psychophysics: Introduction to its Perceptual, Neural, and Social Prospects*. New York: John Wiley.