Grammaticality Judgements and Linguistic Methodology

Carson Theodore Robert Schütze

A paper submitted in conformity with the requirements
for the Degree of Master of Arts in the
Department of Linguistics, University of Toronto

Supervisors: Prof. Peter Reich, Prof. Graeme Hirst

Abstract

My goal is to argue that the absence of a methodology of grammaticality judgements in linguistics constitutes a serious obstacle to meaningful research, and to begin to propose a suitable remedy. Since at least the beginning of the generative paradigm in linguistics, judgements of the grammaticality/acceptability of sentences have been the major source of evidence in constructing grammars, leading some to suggest that theoretical linguists are in fact constructing grammars of linguistic intuitions, which need not be identical with the competence underlying production or comprehension. Also, in this pseudo-experimental procedure of judgement elicitation, there is typically no attempt to impose any of the standard experimental control techniques, and often the only subject is the theorist himself or herself. We provide a survey of how grammaticality judgements are currently used in theoretical syntax, and argue that such uses, in combination with the problems of intuition and experimental design, demand a careful examination of judgements, not as pure sources of data, but as instances of metalinguistic performance.

Several important issues arise when this view of grammaticality judgements is taken, including what tasks one can use to elicit them, how people might go about giving them, and what they might tell us about linguistic competence. Our central hypothesis is that grammaticality judgements result from interactions between primary language faculties of the mind and general cognitive properties, and crucially do not involve special components dedicated to linguistic intuition. We review the psycholinguistic research that has examined ways in which the judgement process can vary with differences between subjects and with experimental manipulations. Parallels with other cognitive behaviours that our hypothesis predicts are pointed out. We then integrate the substantive and methodological findings in the form of a model of linguistic knowledge that reflects what is known about linguistic intuitions, and a proposed methodology for collecting grammaticality judgements while avoiding the pitfalls of previous work and taking account of the conditions that have been shown to influence them. Finally, we discuss how mainstream linguistic theory might be affected by the growing body of research in this area.

It is simultaneously the greatest virtue and failing of linguistic theory that sequence acceptability judgments are used as the basic data.

(Bever 1970b)

Acknowledgements

Just as the Navajo weavers purposely make one error in a rug, to let the soul out, so I cannily craft errors into all of my papers.

(Bever 1970a)

This paper has benefitted enormously in both content and style from the contributions of several people. None of them bears any responsibility for its remaining flaws, cannily crafted or otherwise (they are all the fault of that little man that runs around inside my Macintosh making it work). First and foremost, I would like to thank my supervisors, Professors Peter Reich and Graeme Hirst, without whose comments and criticisms a far inferior product would have resulted. Peter has enthusiastically supported my academic work in general for several years, and this project in particular since August 1990, when we both discovered to our surprise and delight that there was a literature on the topic of grammaticality judgements. At the same time, he has encouraged my non-academic pursuits, for which I am truly grateful. Graeme has been invaluable in pointing out relevant work in fields I was totally unaware of, in tracking down current unpublished research, and in his meticulous scrutiny of my prose. He has been most generous with his time and energies, despite innumerable other priorities. Both Graeme and Peter have provided ongoing support of my attempts to do Cognitive Science at an institution that was not designed for the purpose; I hope this paper shows that it can be done.

Much of the groundwork for this paper was laid in the course of the M.A. Forum class, and so I owe a vote of thanks to my fellow participants. Professor Elan Dresher, our Forum supervisor, supplied encouragement and skepticism in just the right doses to keep us working steadily. He also read drafts of several chapters of the paper, providing a perspective that would otherwise have been lacking. His open-mindedness and sense of humour have been a boon to us all. To my fellow Forum students, Amy Green, Päivi Koskinen and Ana Palma dos Santos, I would like to say thanks for their comments on my work, and more importantly for the camaraderie they provided as we faced this awesome task together.

Several others have contributed in important ways to the present opus. I thank Professor Elizabeth Cowper for useful discussions at the beginning of this project, and for her advice on portions that deal with current syntactic theory. Professor Susanne Carroll brought to my attention one of the most important sources on this topic, Birdsong 1989, that I might otherwise not have come across. I should also thank Charles Houpt at Cornell, whom I have never met, but who, through the miracle of modern computer networking technology, shared his thoughts (and course papers) on this topic and encouraged me to pursue it as a Forum paper. I would also like to acknowledge various USENET readers who contributed pointers to the literature or participated in my pilot survey on *that*-trace effects.

Working on this paper has been a long, tough haul, so I would be remiss if I did not express my gratitude to those who helped me get through it, whether they realized they were doing so or not. I thank the members of the Department of Linguistics, and in particular Jila Ghomeshi, for helping me keep the work and my life in perspective, thereby allowing this paper to be completed in the prescribed time with some modicum of my sanity intact. My major source of relaxation/therapy over the past year has been music, so I must thank those who gave me the opportunity to play with them and who put up with my distraction and wavering commitment: the Skule™ Stage Band, the Skule™ Nite gang, and most especially my musical soul brother, Brent Klassen, whose friendship has been a priceless commodity that I come to appreciate more with each passing day. Above all in this regard, I am so grateful to my parents for their moral and financial support of my academic pursuits, for providing an ideal working environment, and for not complaining when school work took priority over doing the dishes, mowing the lawn, etc., etc.

My Master's research has been financially supported by a Postgraduate Scholarship from the Natural Sciences and Engineering Research Council of Canada, whose support of research in Cognitive Science is hereby gratefully acknowledged.

Contents

Chapter 1: Preliminaries 1

1. Introduction 1
2. Definitions and Historical Background 3
3. The Uses of Judgement Data in Linguistic Theory 14
   - 3.1 Introduction 14
   - 3.2 The Dangers of Unsystematic Data Collection 16
   - 3.3 A Case Study in the Use of Subtle Judgements 19
   - 3.4 The Interpretation of the Annotations and Degrees of Badness 21
   - 3.5 Summary 25
4. Summary and Motivation 26
5. Scope and Organization 27

Chapter 2: Judging Grammaticality: The Nature of a Metalinguistic Performance Process 30

1. Introduction 30
2. Tasks that Elicit Judgements of Grammaticality 31
3. The Graded Nature of Judgements 36
4. The Judgement Process 49
5. The Interpretation of Judgements with Respect to Competence 57
6. A Hypothesis 64
7. Conclusion 65

Chapter 3: Between-Subject Factors in Grammaticality Judgements 67

1. Introduction 67
2. Individual Differences: Three Representative Studies 68
3. Organismic Factors 75
   - 3.1 Field Dependence 75
   - 3.2 Handedness 77
   - 3.3 Other Organismic Factors 79
4. Experiential Factors 80
   - 4.1 Linguistic Training 80
   - 4.2 Literacy and Education 87
   - 4.3 Other Experiential Factors 90
5. Conclusion 92

Chapter 4: Stimulus and Procedural Factors in Grammaticality Judgements 93

1. Introduction 93
2. Procedural Factors 95
   - 2.1 Instructions 95
   - 2.2 Order of Presentation 97
   - 2.3 Repetition 98
   - 2.4 Mental State 102
   - 2.5 Judgement Strategy 107
   - 2.6 Modality of Presentation and Register 109
   - 2.7 Speed of Judgement 110

3. Stimulus Factors 111
3.1 Context 111
3.2 Meaning 118
3.3 Frequency 121
3.4 Lexical Content 123
3.5 Morphology and Spelling 124
4. Conclusion 125

Chapter 5: Theoretical and Methodological Implications **126**

1. Introduction 126
2. Modelling Grammaticality Judgements 127
2.1 Previous Work 127
2.2 A Preliminary Model 128
2.3 Applications of the Model 135
3. Methodological Proposals 138
3.1 Materials 138
3.2 Procedure 140
3.3 Analysis and Interpretation of Results 147
4. Conclusion 152

Chapter 6: Looking Back and Looking Ahead **154**

1. Introduction 154
2. Directions for Future Research 156
3. The Future in Linguistics 157

References **161**

Chapter 1

Preliminaries

Linguists have not formulated a "methodology of sentence judgements."

(van Riemsdijk & Williams 1986)

1. Introduction

The goal of this paper is to argue that the truth of the statement above constitutes a serious obstacle to meaningful linguistic research, and to begin to propose a suitable remedy. Since at least the beginning of the generative paradigm in linguistics, judgements of the grammaticality/acceptability[1] of sentences have been a major source of evidence in constructing grammars. A priori, it is not obvious why a description of people's competence in understanding and producing language should be based on behaviour in situations where they are arguably doing neither, but rather are rendering intuitions. There are three key reasons. First, by eliciting judgements on sentences provided by the researcher, we can examine reactions to sentence types that might occur only very rarely in spontaneous speech or recorded corpora. This is one reason for performing experiments in social science—observational study simply does not always provide a high enough concentration of the phenomena we are most interested in.[2] A second, related purpose is to obtain a form of information that scarcely exists within normal language use at all, namely negative information, in the form of strings that are not part of the language. The third reason for using judgements is that when one is merely observing

---

1 These terms will be defined and distinguished in §2.

2 In principle, the conclusion does not automatically follow. One could theoretically do experiments on the production and comprehension of sentences chosen by the researcher, without recourse to judgements. In practice, however, this is problematic. On the production side, it is extremely difficult to induce a subject to produce precisely the sentence one wishes to study without actually exposing the subject to the sentence. On the comprehension side, it is hard to discover anything about the nature, or even the success or failure, of the comprehension process without eliciting some additional reaction, e.g. a judgement.

speech it is difficult to distinguish reliably slips, unfinished utterances, etc., from normal production.[3]

While such justifications seem sensible enough, perhaps even unavoidable, it must be acknowledged that soliciting linguistic judgements is problematic in a number of respects. Not only is the elicitation situation artificial, with the standard issues of ecological validity, but the subject is being asked for an entirely different sort of behaviour than in everyday conversation. This has led some to suggest that theoretical linguists are in fact constructing grammars of linguistic intuitions or judgements, which need not be identical with the competence underlying production or comprehension (e.g. Bever 1970a, Birdsong 1989, Gleitman & Gleitman 1979). In addition to these problems, which often come up in psychology as well, there are important shortcomings that arise because linguistic elicitation does not follow the procedures of psychological experimentation. In the vast majority of cases in linguistics, there is not the slightest attempt to impose any of the standard experimental control techniques, such as random sampling of subjects and stimulus materials, counterbalancing for order effects, etc. Perhaps worst of all, often the only subject in these pseudo-experiments is none other than the theorist himself or herself (Newmeyer 1983; Bradac, Martin, Elliott & Tardy 1980). In 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. Not a few observers of linguistics have agreed with Labov's "painfully obvious conclusion . . . that linguists cannot continue to produce theory and data at the same time" (Labov 1972a, p. 199). What is to stop linguists from (knowingly or unknowingly) manipulating the introspection process to substantiate their own theories?

An additional rationalization for the use of grammaticality judgement data in some cases seems to have been related to Chomsky's competence-performance distinction:[4] since actual speech production and comprehension are fraught with errors of all kinds, false starts, etc., and subject to human memory limitations, these so-called performance variables serve to obscure the underlying competence. But what if we could relieve subjects of the "cognitive burden" of actual production or comprehension and pre-

---

3 See Grandy 1981 for these and other standard arguments in favour of judgements; see Newmeyer 1983, pp. 62–63 for additional arguments against alternative data sources; see Carden 1976 for a review of some moderately successful non-judgement tasks, and Greenbaum and Quirk 1970 for extensive use of such tasks in conjunction with grammaticality judgements.

4 See §2 for a detailed discussion of this distinction.

sent them with ready-made sentences such that the only task would be to judge their grammaticality? Would this not get us much closer to people's true competence?[5] Unfortunately, there is ample evidence that the answer is "No." While, we claim, grammaticality judgements offer a *different* access path to competence than language use, they are themselves just another sort of performance (Birdsong 1989; Levelt, van Gent, Haans & Meijers 1977; Bever 1970, 1974; Bever & Langendoen 1971; Grandy 1981), and as such are subject to at least as many confounding factors as production, and likely even more.

The purpose of this chapter is to motivate the search for resolutions to the issues raised above. Its format is as follows. §2 is devoted to a brief history of the issues surrounding the notion of grammaticality, the associated terminology, and diverse views on its role in linguistic theory. In §3 we shall survey the varied ways in which grammaticality judgements are being used in the linguistic literature, considering the types of data collected and the manner in which these are employed to argue theoretical points. In §4 we use these and other reasons to motivate the goals and approach of the remainder of the paper: before intuitions (or any other behaviour) can really begin to tell us something about competence, we need to at least be aware of, and ideally understand the effects of, the component psychological processes that intervene between the two. It is proposed that this should be achievable in principle if we set out to construct a comprehensive model of the process; this would allow the extensive research already conducted by psycholinguists to be unified and integrated, and contradictory results scrutinized. At the very least, a well-supported model of this type should raise the awareness of linguists to the vast complexities underlying the apparently simple task of deciding, "Is this a good sentence?" Finally, §5 sets out the scope and structure of the remainder of the paper.

2. Definitions and Historical Background

An investigation into the nature of grammaticality judgements demands a description of precisely what is intended by the term *grammaticality* and thus, what could consti-

---

5 It is difficult to find explicit examples of this reasoning in the theoretical linguistic literature, but the belief seems to have been very widely held; Birdsong cites numerous instances where Lasnik, Chomsky and others attempt to curb this view, e.g. Lasnik 1981, p. 20: "Grammaticality judgments are often *incorrectly* considered as direct reflections of competence" (emphasis added). Certainly many authors have accused Chomsky of claiming that people have a consistent ability to assess grammaticality (e.g. Nagata (1988)), which is certainly not true of any of his works written after the mid-1960s. The view might have stemmed in part from confusion of Chomsky's terms "intuition" and "judgment," a matter that we take up in §2.

tute a judgement of it. Since our eventual goal is to scrutinize the use of these judgements in generative grammar, we adopt the assumptions of that framework without further comment, although much of our investigation has theory-independent implications. Thus, for the relevant definitions we turn to Chomsky, and in particular to the familiar competence-performance distinction. Chomsky's basic claim is that we must distinguish what a speaker of a language knows (subconsciously) about the structure of the language from his actual use of the language. The goal of linguistic theory, under this view, is to describe the knowledge, independent of (and logically prior to) any attempt to describe the role that this knowledge plays in the production or understanding of language.[6] Whether a sentence is *grammatical* is a question about competence, i.e., is the sentence generated by the speaker's grammar, is it part of the language as delineated by his competence? We will assume for the purpose of discussion that the answer to this question is determinate in all cases, i.e. that whatever form the competence takes in the mind, it implicitly ascribes either grammaticality or (perhaps some degree of) ungrammaticality to each string of words.[7] On the other hand, whether a sentence is *acceptable* is a question about performance, i.e. does a speaker consciously accept the sentence as part of his language upon hearing it?

Given that linguistic competence is only one contributing factor in any observable behaviour of a speaker, it is reasonable to ask whether in principle there can be any operational test for the grammaticality of a sentence.[8] In his early work, Chomsky seems to have thought that the answer could be "Yes." In *The Logical Structure of Linguistic Theory* (Chomsky [1955] 1985, hereafter *LSLT*)[9] and *Syntactic Structures* (Chomsky 1957, hereafter *SS*) we find the following remarks:[10]

---

6 There has been considerable criticism of this view; see Greenbaum 1976b for a list of dissenting opinions.

7 It is conceivable, however, that competence in this sense of statically-represented knowledge does not exist. It could be that the status of a given string is only computed when necessary, and that the demands of the particular situation determine how the computation is carried out, e.g. by some sort of comparison to prototypical sentence structures stored in memory. Since such a scenario would demand a major re-thinking of the goals of the field of linguistics, we will not deal with it further.

8 See Oller, Sales & Harrington 1970 for insightful commentary on the nature of the competence-performance distinction and the empirical status of generative grammars.

9 All page numbers refer to the 1985 printing.

10 We cite many of Chomsky's passages verbatim, because the wording is often subtly nuanced and easily misinterpreted or mis-paraphrased. The reader is then free to disagree with our interpretation.

> One way to test the adequacy of a grammar proposed for L is to determine whether or not the sentences that it generates are actually grammatical, i.e., acceptable to a native speaker, etc. We can take certain steps towards providing a behavioral criterion for grammaticalness so that this test of adequacy can be carried out. (*SS*, p. 13)

> [A speaker of a language] can also distinguish a certain set of "grammatical" utterances, among utterances that he has never heard and might never produce. (*LSLT*, p. 61)

He goes on to propose some behavioural correlates of ungrammaticality. The number of passages cited below demonstrates that there was nothing idle or casual about the idea (although we have emphasized some apparent hedges as well):

> Yet (1) [*Colorless green ideas sleep furiously*], though nonsensical, is grammatical, while (2) [*Furiously sleep ideas green colorless*] is not. Presented with these sentences, a speaker of English will read (1) with a normal sentence intonation, but he will read (2) with a falling intonation on each word; in fact, with just the intonation pattern given to any sequence of unrelated words. He treats each word in (2) as a separate phrase. Similarly, he will be able to recall (1) much more easily than (2), to learn it much more quickly, etc. (*SS*, p. 16)[11]

> Such sentences with conjunction crossing constituent boundaries are also, in general, marked by special phonemic features such as extra long pauses . . ., contrastive stress and intonation, failure to reduce vowels and drop final consonants in rapid speech, etc. Such features *normally* mark the reading of non-grammatical strings.[12] (*SS*, pp. 35–36, fn. 2)

> We know that a speaker of the language can select, among sequences that he has never heard, *certain* grammatical sentences, and that he will do this in much the same way as other speakers. We might test this by a direct determination of some sort of "bizarreness reaction," or in various indirect ways. (*LSLT*, p. 95)

At the very same time, however, it was apparent that behavioral criteria were not always the last word: the theory could also dictate that some sentences *must* be grammatical, regardless of how speakers might react to them. With regard to sentences containing embedded *if-then* and *either-or* pairs, Chomsky states, "Note that many of the sentences . . . will be quite strange and unusual . . . But they are all grammatical sentences, formed by processes of sentence construction so simple and elementary that even the most rudi-

---

11 Regarding this and the following passage, Chomsky (1961) states that he was careful not to suggest *general* criteria for grammaticalness. If that truly was not his intent at the time, I believe he could have chosen his words more appropriately.

12 The emphasis in all quoted passages in this chapter is my own, unless otherwise indicated.

mentary English grammar would contain them" (*SS*, p. 23). The reasoning seems to be this: given that certain sentences are uncontroversially part of the language, our intuitions about what grammars can look like tell us that certain other sentences must also be part of it, although our judgements of these latter sentences are not so clear-cut. Thus, as our concept of what the theory is about changes, the status of any given sentence can change from grammatical to ungrammatical, depending on what criteria "count" towards grammaticality. For instance, *Colorless green ideas sleep furiously* was considered "grammatical" in *SS*, but "deviant" in Chomsky (1965), because it violated selectional rules (p. 149). (See Newmeyer 1983, p. 58 for more discussion of this sentence.) What this means for eliciting judgements is that we somehow must get across to the naïve subject what counts, which of course requires that we as linguists must have an explicit understanding of that ourselves. That this is not the case is well demonstrated by the numerous instances where Ney (1975) disagrees with published judgements of sentences. In many (perhaps most) cases, the disagreement is likely over whether various types of anomalies are in the domain of "grammaticality" versus some other dimension.

Thus, although in principle the theory is subject to empirical disconfirmation, in practice it is only the indisputable judgements that will be accepted as falsifying evidence, as the following passages state explicitly:

> A grammar . . . is to be confirmed or disconfirmed in terms of empirical evidence drawn, ultimately, from investigation of the linguistic intuitions of the language-user (which might, *in principle*, be analyzed in terms of operational tests . . .). (Chomsky 1973, p. 37)[13]
>
> Clearly the sequences generated by the grammar as grammatical sentences must be acceptable, *in some sense*, to the native speaker. (*LSLT*, p. 101)
>
> Our purpose is to construct an integrated and systematic theory, which, when applied rigorously to linguistic material, gives the correct analysis *for the cases where intuition* (or experiment, under more desirable circumstances) *makes a clear decision*. (*LSLT*, p. 415)

Note that Chomsky generally does *not* use the terms "intuition" and "judgement" interchangeably; it is my best understanding that the latter is a product of performance, the former is part of competence. When he says that "the speaker has an 'intuitive sense of grammaticalness'" (*LSLT*, p. 95), this does not translate into the ability to *judge* grammat-

[13] We have included passages from Chomsky 1973 as representative of his early work, since it provides background to *LSLT*, of which it is the introduction. Of course, the intervening years might have brought a change in perspective.

icalness. But the very close semantics are probably responsible for the misapprehension alluded to in §1, whereby the two are equated and thus judgements are seen as directly reflecting competence, since competence consists of intuitions. As stated above, this was not Chomsky's intent (see Chomsky 1965, p. 21), although passages like the following could easily serve to mislead the unwary reader: ". . . the theory is refuted if the *judgments* are not in accord with the predictions of the grammar" (Chomsky 1973, p. 36). In the accompanying footnote, however, we find a qualification: "Note that there is a further idealization here, in that we abstract away from other factors that may interact with knowledge of language to determine judgments."

If there was some question at the time as to the possibility of judging grammaticality, there was no question that acceptability was something speakers knew about. But if the following passage from the introduction to *LSLT* truly reflects Chomsky's thinking at the time that work was written, the concept was broader than we have so far assumed: "Sentences are acceptable (or perhaps acceptable under particular circumstances) if they are suitable, appropriate, adequate to the purpose at hand, etc. The competence grammar contributes to determining acceptability, but the latter concept involves many other factors" (Chomsky 1973, p. 8). The notion thus defined seems to belong to pragmatics rather than syntax. Perhaps the distinction was originally not a performance versus competence one, but rather reflected judgements of two different sorts: syntactic well-formedness versus pragmatic appropriateness. These would both be concepts defined by our linguistic knowledge (competence), and people could render judgements about either or both of them, which might or might not always jibe with the "pure" competence.

But by the time of *Aspects of the Theory of Syntax* (Chomsky 1965, hereafter *ATS*), Chomsky's opinion concerning the possibility of empirical tests for grammaticality, and his definition of acceptability, had shifted quite sharply, as the following well-known passage shows.[14]

> For the purposes of this discussion, let us use the term "acceptable" to refer to utterances that are perfectly natural and immediately comprehensible without paper-and-pencil analysis, and in no way bizarre or outlandish. Obviously, acceptability will be a matter of degree, along various dimensions. One could go on to propose various operational tests to specify the notion more precisely (for example, rapidity, correctness, and uniformity of recall and recognition, normalcy of intonation) . . . The more acceptable sentences are those that are

[14] As late as Chomsky 1961, he still spoke of "a battery of tests that may ultimately succeed in giving a characterization of grammaticalness" (p. 229, fn. 20).

> more likely to be produced, more easily understood, less clumsy, and in some sense more natural. The unacceptable sentences one would tend to avoid and replace by more acceptable variants, wherever possible, in actual discourse.
>
> The notion "acceptable" is not to be confused with "grammatical." Acceptability is a concept that belongs to the study of performance, whereas grammaticalness belongs to the study of competence . . . Like acceptability, grammaticalness is, no doubt, a matter of degree . . . but the scales of grammaticalness and acceptability do not coincide. Grammaticalness is only one of many factors that interact to determine acceptability. Correspondingly, although one might propose various operational tests for acceptability, *it is unlikely that a necessary and sufficient operational criterion might be invented for the much more abstract and far more important notion of grammaticalness*. (pp. 10–11)

The intended meaning of acceptability here does not extend beyond the syntactic or structural sense proposed at the beginning of this section; the concept seems to have been used in that narrow sense most of the time in *LSLT* and *SS*, and that is the meaning we will continue to assume. At any rate, it is clear that Chomsky's belief in the possibility of finding tasks that would directly reflect grammaticality had been diminished. Reich (1969) suggests that this is attributable at least in part to experiments such as those of Miller (1962), which found that the criteria proposed by Chomsky (e.g. intonation) did not correspond to what he believed must be true about the grammar of English. As Reich puts it, "when confronted by adverse data, Chomsky retreated from his empirical position of 1957, to a theory that he himself admits cannot be tested empirically."[15] That is, although acceptability is an empirically-defined concept[16] (in fact, defined in terms of many of the previously-proposed operational criteria for grammaticality),[17] grammaticality is not, and the former does not provide direct evidence concerning the latter; in fact, there

---

15 It is true that this shift coincides with the larger change in focus from E-language (the external view of language as a set of sentences) to I-language (the internal grammar) in Chomsky's work (see Chomsky 1986, pp. 24ff.), but the causes of the larger move might be the same.

16 Strangely, Reich's proposal for a definition of acceptability is more problematic than Chomsky's: "A sentence is acceptable to me if my estimate of the probability of occurrence of a sentence of like construction in a natural language text is greater than zero. I exclude from natural language text sentences dreamed up by linguists, psychologists, English teachers, and poets" (Reich 1969, p. 832, fn. 7). The proposed decision procedure is based on one person's subjective estimate of probability, without even specifying what sorts of information or experience should be the basis of the estimate. A similar alternative definition is suggested by R. Lakoff (1977): "the probability of such a sentence being uttered, or the number of conceivable real-world circumstances or the normality of the real-world circumstances in which this sentence is apt to be used" (p. 75). Again, applying this definition requires someone to make estimates of probability and conceive of possible scenarios where sentences might appear, again lacking objectivity.

17 This point is made by Scott and Mills (1973) as well.

are no empirical criteria for grammaticality. (Marks (1967) comments on what he perceives as the incoherence of this position.) It does not make any sense to speak of "grammaticality judgements" given these definitions, because people are incapable of judging grammaticality—it is not accessible to their intuitions (Newmeyer 1983, p. 51); linguists may construct arguments about the grammaticality of a sentence, but all that a linguistically naïve subject can do is judge its acceptability. (Nevertheless, in the remainder of this paper I will follow the existing literature in treating "grammaticality judgement" and "acceptability judgement" as synonyms,[18] with the understanding that the former is unquestionably a misnomer, and only the latter is a sensible notion. I will continue to follow Chomsky's definitions in other contexts when the distinction is important, e.g. "acceptable sentence" versus "grammatical sentence.")

Given that grammaticality is what Chomsky seeks to investigate, it would not be surprising if he saw no useful purpose in the systematization of linguistic data collection: in the end, no single empirical fact can be crucial to the issues at hand. And yet, a very few pages after the above passage from *ATS*, after elaborating this point once again, Chomsky does suggest that there might be room at some future time for a methodology more systematic than reliance on everyday commonsense:

> There are, in other words, very few reliable experimental or data-processing procedures for obtaining significant information concerning the linguistic intuition of the native speaker. It is important to bear in mind that when an operational procedure is proposed, it must be tested for adequacy . . . by measuring it against the standard provided by the tacit knowledge that it attempts to specify and describe . . . If operational procedures were available that met this test, we might be justified in relying on their results in unclear and difficult cases. This remains a hope for the future rather than a present reality, however . . . there is no reason to expect that reliable operational criteria for the deeper and more important theoretical notions of linguistics (such as "grammaticalness" and "paraphrase") will ever be forthcoming . . . The critical problem for grammatical theory today is not a paucity of evidence but rather the inadequacy of present theories of language to account for masses of evidence that are hardly open to serious question . . . it seems to me that sharpening of the data by more objective tests is a matter of small importance for the problems at hand . . . *Perhaps the day will come when the kinds of data that we now can obtain in abundance will be insufficient to resolve deeper questions concerning the structure of language.* (pp. 19–21)

---

18 It is possible that researchers who have defined grammaticality and/or acceptability in other ways might make a principled distinction between two types of judgements, but since I assume Chomsky's definitions I will collapse the terms unless otherwise noted.

I would like to argue that this day has come, 26 years later. I will devote §3 to demonstrating that the questions linguists are now addressing rely crucially on facts that are "open to serious question."

Note also in this second passage that Chomsky assumes there is a core of "unquestionable data concerning the linguistic intuition of the native speaker," which would presumably include judgements of some sort, and that these "obvious" facts would keep linguists busy for a long time, thus postponing the need for reliable tests applicable to "less obvious" cases.[19] That is, for some sentences, acceptability judgements provide transparent evidence about grammaticality, while still not constituting grammaticality judgements in the literal sense. The problem, of course, is that each investigator is free to pick and choose these "unquestionable" cases to suit the theory.[20] (McCawley (1985) argues that, by Chomsky's own definition of grammaticality, *all* sentences must be considered unclear cases, because we never have direct information about grammaticality.) For instance, in *LSLT* Chomsky examines the "naturalness" of particle movement as a function of the complexity of the intervening NP and concludes, "This is systematic behavior, and we might expect that a grammar should be able to state it" (p. 477). But in *ATS* he says of the same sentences (and, more celebratedly, of multiply centre-embedded ones), "it would be quite impossible to characterize the unacceptable sentences in grammatical terms. For example, we cannot formulate particular rules of the grammar in such a way as to exclude them" (pp. 11–12). If people reject sentences that the grammar cannot exclude, Chomsky is forced to say that their rejection is not an "unquestionable" reflection of their intuition. As Reich (1969) points out, the result can only be circular argumentation: the grammar is supposed to account for facts about language, but what counts as a fact about language is determined by the grammar. (Oller, Sales, and Harrington (1970)

---

19 An argument against this position is that the large masses of unquestionable data, if indeed they exist, might still be of insufficient quality for linguistic theory, if they do not bear on the crucial issues that it must address (Botha 1973; Labov 1972a).

20 That there is no standard way to make this decision is argued in detail by Botha (1973): "the level of rationality at which grammatical inquiry and general-linguistic inquiry are conducted would be raised if it were clear . . . under what circumstances an intuitive evidential statement may be properly regarded as being evident" (p. 188); "transformational grammar lacks a set of conditions . . . governing the evidentness or obviousness of intuitive evidential statements" (p. 193). Furthermore, even at an intuitive level, clear-case judgements are not necessarily windows into competence: "In terms of the notion 'clear case' spurious linguistic intuitions could, despite their spuriousness, qualify for membership of the evidential corpus; in terms of the notation 'unclear case' linguistic intuitions which were both genuine and correct could, despite their genuineness and correctness, be denied membership of this corpus" (p. 206). Botha defines a spurious judgements as one that has been influenced by extra-linguistic factors.

construct a similar argument for circularity in the definitions of "grammar" and "grammaticality.") Thus, the most we can ask of a grammar is that it account for the facts that (its author believes) it is capable of accounting for. Such a system is unfalsifiable *in principle*, as well as in practice.[21] (See Labov 1975 for more discussion of Chomsky's hedges with regard to the use of clear cases, including instances where he admits the data are unclear but proceeds to construct the theory on the basis of his own intuitions anyway.)

In general, there seem to be three "lines of defense" by which a theory is protected from potentially falsifying data, before any change can be incited. The present work is intended to help objectify two of these three procedures. Specifically, when faced with a sentence that apparently contradicts the theory, the first line of defense would be to argue that the data are invalid, i.e. the sentence is not really acceptable as claimed. While it is trivial to make this statement for one's own intuitions, such arguments ought to be supported by empirical investigations of others as well. To the extent that we can standardize this process, we can eliminate data disputes. The second defense is to claim that the data are not relevant to the theoretical issue at hand, i.e. the sentence is good (or bad) because it is (dis)allowed by some other part of the grammar and is not "under the jurisdiction" of the relevant constructs. This approach generally relies on logical reasoning and may be subject to differing opinion but not to factual dispute. The third choice, typically indicative of the least understanding on the part of the theory's proponent, is to say that the sentence *is* generated by the grammar, but non-grammatical factors are causing judgements not to reflect this fact. Until we have an explicit understanding of such factors, such a claim is unfalsifiable. (For a much more detailed examination of the roles of argumentation and evidence in generative theory, see Botha 1973).

As I see it, this is precisely why we *should* strive for a better understanding of acceptability judgements. 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. Until then, as Birdsong (1989) states, if we do

---

21 In Chomsky 1981 we find the same circularity with regard to what Universal Grammar is intended to account for: while the division between core grammar and marked periphery is theoretically subject to empirical criteria (the periphery being someone else's problem), "such evidence is, for the time being, insufficient" and "we are therefore compelled to rely heavily on *grammar-internal considerations* and comparative evidence, that is, on the possibilities for constructing a *reasonable* theory of UG and considering its explanatory power in a variety of language types" (p. 9).

not agree on what our data represent, we cannot hope to agree on an analysis. If we can understand what factors intervene between the grammar and performance, we can circumscribe the cases where these factors might cause (un)acceptability not to reflect (un)grammaticality, and exclude these as evidence. In fact, Chomsky himself suggests this as a potentially fruitful approach (Chomsky 1986, pp. 36–37); the proposal dates back at least to Maclay and Sleator (1960). If we can go that (huge) step further and deduce a reverse mapping from acceptability to grammaticality, we *can* derive operational tests that bear directly on grammaticality by determining which unacceptable sentences are grammatical and which acceptable sentences are ungrammatical,[22] and thus objectify the range of facts over which grammars must have scope. "We require a science of linguistic introspection to provide a theoretical and empirical basis for including some acceptability judgments as syntactically relevant and excluding others" (Bever 1974, p. 195). Bever goes on to make some preliminary suggestions about sentential properties that will likely affect acceptability but that are outside the realm of the grammar: these include sentence length, absurdity, difficulty of comprehension and difficulty of pronunciation. And of course, the paradigm example in this category would be the short-term memory limitations that are said to result in the unacceptability of multiply centre-embedded sentences. But few of these are uncontroversial:[23] for instance, Reich (1969) and Spencer (1973) question why limitations on centre embeddings should not be taken to reflect the grammar, and the role of meaning vis-à-vis the grammar has been a point of great debate over the history of generative linguistics. Under Bever's approach, only those unacceptable sentences whose badness cannot be explained by any known aspect of speech *behaviour* are ungrammatical (Bever's example of such a case is *I hope it for to be stopping raining when I am having leaving*).

This idea of factoring grammaticality out of acceptability judgements has been proposed before in various camps (e.g. Birdsong 1989; Carroll, Bever & Pollack 1981; Botha 1973). Among the more striking are the following comments from Grimshaw and Rosen (1990), who argue that, contrary to first appearances, children's linguistic be-

---

22 That such mismatches could exist was considered a "frightening spectre" by Ney (1975). The former case (grammatical but unacceptable) is certainly the more familiar, since it is easier to argue for, but see Langendoen & Bever 1973 and Bever 1974 for some supposed instances of ungrammatical acceptable sentences. Of course, as theories change, such sentences might no longer be considered ungrammatical.

23 Newmeyer (1983) believes that discrepancies on this point account for many of what appear to be data disagreements among theorists.

haviour does tell us something about their grammars, namely that they include Principle B of the Binding Theory.

> Performance in an experiment, including performance on the standard linguistic task of making grammaticality judgements, cannot be equated with grammatical knowledge. To determine properties of the underlying knowledge system requires inferential reasoning, sometimes of a highly abstract sort. (p. 188)

> The inevitable screening effects of processing demands and other performance factors do not prevent us from establishing the character of linguistic knowledge; they just make it more challenging . . . an analysis of these performance factors makes it possible to see, if only dimly, through the performance filter. (p. 217)

The paper is somewhat unusual in that it represents work by theoreticians where a major goal is the explanation of the connection between behaviour on judgement tasks and linguistic knowledge. While a naïve view of the facts contradicts their claim, they argue that once psychological factors such as response bias and experimental demand characteristics are taken account of, the results support their theory. One may still dispute their conclusions, but the efforts are certainly in the right direction.

It is interesting to examine what other theoretical linguists today believe about the types of evidence that are available to them. The following unusually explicit passage (whence the opening epigraph of this chapter is drawn) confirms that judgement data are still the primary source, and thus underscores the importance of studying their properties. It also states that we continue to lack a principled criterion for choosing data.

> No kind of data is excluded in principle, only as a matter of practice—judicious practice, we think, but not irrefutable . . . grammarians use data like "such and such a string of words is a sentence in such and such a language" or "such and such a string of words means such and such," where such facts are determined by native speakers of the languages in question. Data of this kind vary enormously in quality—ranging from the clear fact that *He are sick* is not grammatical in English to the rather subtle judgments involved in determining whether *John* and *his* can refer to the same person in *His mother likes John.* Despite this variation in quality and despite the fact that linguists have not formulated a "methodology of sentence judgments," such data remain the principal source of information about grammar, again, not as a matter of principle, but because they have so far provided successful insights.

> Thus, the study of grammar is not the study of sentence judgments; rather, sentence judgments are our best current avenue to the study of grammar. In other words, the grammar is a real thing, not an artifact erected on top of an arbitrarily demarcated set of facts or types of facts. Therefore, it is often difficult to determine whether a given fact bears on grammar or not; this is not an

> arbitrary decision, but ultimately an empirical question about how the world divides up." (van Riemsdijk & Williams 1986, p. 2)

It is at least possible a priori that the reason judgements seem to work well for linguists is that they can be manipulated, distorted, etc., to suit the purpose of the analysis. In this connection, Birdsong (1989, p. 82) suggests that "linguistics is a potentially fraudulent enterprise when elicitation data can be manipulated to substantiate pet theoretical analyses. It would be hard to imagine a more powerful argument for understanding the psychology of metalinguistic performance." Some authors profess ignorance of this choice, for example Baker:

> We focus on those linguistic behaviors which *for some reason* are most likely to reveal the mental structures in their true light. The situation can be likened to the physicist who tries to determine the force of gravity . . . unfortunately there is every indication that much of the linguistic behavior we have record of is like the autumn leaf—complicated by many other external factors . . . I do not claim to have the wisdom to reliably discern which linguistic behaviors are like autumn leaves and which are like steel ball bearings." (Baker 1988, p. 29)

Despite his lack of "wisdom," Baker implicitly chooses to continue the tradition of making primary use of judgement data.

Levelt et al. (1977) take the more cynical view; one could scarcely hope to summarize this section and our position more succinctly and eloquently than they do in the following passage:

> Linguistic intuitions became the royal way into an understanding of the competence which underlies all linguistic performance. However, if such a linguistic competence exists at all, i.e., some relatively autonomous mental capacity for language, linguistic intuitions seem to be the least obvious data on which to base the study of its structure. They are very derived and rather artificial psycholinguistic phenomena which develop late in language acquisition . . . and are very dependent on explicit teaching and instruction. They cannot be compared with primary language use such as speaking and listening. The empirical domain of Chomskian linguistics is linguistic intuitions. The relation between these intuitions and man's capacity for language, however, is highly obscure. (pp. 88–89)

3. The Uses of Judgement Data in Linguistic Theory

3.1 Introduction

The purpose of this section is twofold: first, we wish to demonstrate the claim, made in §2, that current issues in linguistic theory require "non-obvious" data for their

resolution; second, and relatedly, we illustrate that the use of judgements in theoretical work has moved far beyond good versus bad, or even graded goodness and badness decisions. Once again, the situation is characterized most poignantly by Levelt:

> In the early years of the transformational grammar [the low reliability of absolute grammaticality judgments] was not an important issue, since the 'clear cases,' i.e., the highly uncontroversial cases of grammaticality and ungrammaticality, were sufficient for constructing and testing linguistic theory. It was expected that, in its turn, the theory constructed in such a way would decide on the 'unclear cases.' This hope has vanished. (Levelt et al. 1977, p. 88)

> It has slowly but surely become clear that it is not possible, on the basis of incontrovertible, directly evident data, to construct a theory so extensive that all less obvious cases can be decided upon by the grammar itself. It is becoming more and more apparent that decisions on very important areas of theory are dependent on very unreliable observations . . . There is a tendency toward preoccupation with extremely subtle distinctions, not the importance, but rather the direct observability of which can seriously be called into question. (Levelt 1974, vol. 2, p. 6)

The same complaint has been made throughout much of the history of generative grammar, e.g. by Bever (1970a, p. 348), Labov (1972a, p. 191), and Birdsong (1989, p. 81). It has come to be generally acknowledged that not all speakers of "the same language" may have the same competence, but that does not justify basing the theory only on sentences for which there is universal agreement, and extrapolating by some means to dictate the status of the remainder. In cases where people disagree, that fact cannot be ignored; the theory must be able to describe *every* speaker's competence, and thus must allow for variation wherever it occurs. This is why establishing the extent of inter-speaker agreement is important: theories are now being based on sentences whose status turns out not to be unanimous, as we will see in §3.2. Coppieters (1987) summarizes the argument as follows: "All that is left are idiolects; we assume that these share many features within a given language community, but that they also show a certain degree of independence from each other. When dealing with straightforward non-controversial aspects of a language, we can maintain the fiction of a standardized object named English or French, characterized by standard conventions of usage and standard intuition reports concerning the elements which belong to this object. However, such an approach becomes completely inappropriate at a higher level of complexity" (p. 548). See Chapter 5, §3.3 for further discussion of the implications of individual differences in grammaticality judgements and linguistic competence.

What follows is surely not a random sample of the theoretical syntax literature, but it includes some very influential and widely-cited papers. Our particular interest will be not just the types of judgement data that are employed, and hence the judgement abilities attributed to native speakers, but also the importance of these judgements to the theoretical arguments, i.e. to what extent the arguments would be weakened if the fine-grained judgements were unavailable. In many cases we will not mention the details of the theoretical issues, since they are irrelevant to our purpose here. We preserve the original example numbers in quoted passages and data, so that the interested reader may consult the original sources.

3.2 The Dangers of Unsystematic Data Collection

Let us begin with an important case where inter-speaker variation in judgements has been *ignored*, to the detriment of the theory, before examining the ways in which judgements are *used* by theoreticians. The belief is widely-held among theoreticians that the majority of the data on which their theories are based are indisputable. But one cannot assume that what is a clear-cut judgement for oneself applies to all, or even a large majority of, speakers. A case in point is the celebrated article by Lasnik and Saito (1984). One of the major proposals in this work is a substantial revision of the mechanisms of Proper Government to allow sentences like their (99):

(99) Why do you think that he left?

The authors assume that such sentences are ambiguous, i.e. *why* can be taken as questioning the reason for the thinking or the reason for the leaving. In general, they assume that adjunct *wh*-words do not show *that*-trace effects, so that all sentences of this form should also be ambiguous with *why* replaced by *where*, *when* or *how*. On the basis of these assumptions, they propose various complications to the operation of the Empty Category Principle (ECP) and a process of *that*-deletion at Logical Form so that the sentences will not violate the ECP, as they did in earlier theories. But are the crucial readings grammatical?

In a subsequent paper, Aoun, Hornstein, Lightfoot, and Weinberg (1987) propose an alternative theory in the same domain, this time with the goal of accounting for the *ungrammaticality* of some of the very same sentences that Lasnik and Saito went out of their way to include in the grammar, namely those containing *why* and *how* as the *wh*-

words. Anticipating reaction to the apparent data disagreement, they make the following comments:

> Some speakers claim to get a lower-clause interpretation for *why* in (51a) [*Why did she say that there are men outside*] even if a complementizer is present. However, we have found that when asked to repeat the sentence, those speakers omit *that*, as if it were not perceived.
>
> . . . English speakers who accept (51a) may be able to use *why* referentially, in the sense of 'for what reason'. But the *acceptability* of (51a) for such speakers does not seem to us to indicate *grammaticality*, unless they also accept (26)–(28) [e.g. **Who remembers what we bought why?*] and the like; rather, an analogical process is involved. (pp. 563–564; emphasis in original)

Aoun et al. seem to be proposing two different explanations. On the one hand, they suggest that the judgement data are inaccurate, i.e. people really cannot get the relevant reading of these sentences. On the other hand, they propose that this reading *is* acceptable for some speakers, and then attempt to argue on theory-internal grounds that it still must not be generated by their grammars, since they would then expect certain other sentences to be acceptable as well.

While time constraints have prevented me from carrying out a carefully-controlled investigation of judgements on the crucial sentences, an informal preliminary survey indicates that the sentences *are* acceptable for about two-thirds of the population. This area provides a striking demonstration of why linguists *must* improve their data-gathering techniques. In the first case, Lasnik and Saito show no evidence of being aware that the only sentences that prompt their major revision of the theory are not universally accepted. If their proposal had been adopted, it would have constituted a major step in the wrong direction, in the absence of any proposal for why a large minority of speakers should find the sentences bad. In the second case, Aoun et al. conclude on the basis of a less-than-rigorous survey that only a small number of speakers claim the sentences to be acceptable, and that some or all of these judgements are incorrect, i.e. they did not consider the crucial presence of *that*. This also turns out not to be the case, again calling the analysis into question. Given the extent to which judgements are divided, I suspect that syntacticians would not want to base any conclusions about Universal Grammar (UG) on these

sentences.[24] But until the detailed judgement facts were known, there was no way to assess the situation accurately.[25]

A similar case is made by Sobin (1987) with regard to *wh*-extraction across *that* versus *whether*. He points out that most theories assume these two kinds of extractions, exemplified in his sentences (1) and (4) below, are equally bad—categorically ungrammatical.

(1) *Who did you say that kissed Harriet?

(4) *Who did you ask whether loves Mary?

But the results of his questionnaire survey (corroborated by various anecdotal observations by others) present a distinctly different picture. He asked 42 nonlinguists to classify sentences into one of three groups, representing active acceptance, passive acceptance, and rejection.[26] Pooling the first two classes, he found the average rejection rate for sentences like (1) to be 17.5%, whereas for sentences like (4) it was 97.6%. The active acceptance rate for (1) was 45.2%. These differences certainly imply that we cannot rely on identical grammatical constraints to rule out both sentence types; while the *whether* sentences are almost unanimously rejected, the *that* sentences are quite widely accepted. Once again, several theories had been constructed on the assumption that (1) was bad for everyone, and furthermore these analyses predicted that it should be just as bad as (4), since it violated precisely the same constraint. As before, these are core data for the formulation of the ECP and associated constraints. Sobin proposes that structures like (4) be ruled out universally, whereas a parametrized rule could determine whether or not (1)

---

24 To do so, one would have to accept Aoun et al.'s 'analogical processing' explanation as applying to the majority of speakers, or propose some alternative.

25 Newmeyer (1983) attempts to play down the significance of a similar situation from the early 1970s, describing it as a case of "letting the theory decide" on a marginal case. In the present situation, however, both camps went out of their way to account for the judgements they perceived, which were not predicted by existing versions of the theory.

26 His descriptions to subjects were worded as follows:

> (a) it sounds like a sentence that you (the informant) might say in the right context or situation;
> (b) it sounds like a possible English sentence, one that even if you don't say it that way, you would not be particularly surprised to hear someone else say it to you that way or to see it written;
> (c) it sounds odd, so that you doubt that people say it that way.

One might raise some questions about these instructions; for instance, they require subjects to have some sense of what a possible English sentence is, and to be aware of how other people might speak.

would be allowed. Whatever the eventual analysis, it is clear that ignoring variation led the theory astray in this case too.

3.3 A Case Study in the Use of Subtle Judgements

We go on now to consider various ways in which judgements are used in theoretical argumentation, beginning with Belletti and Rizzi's (1988) influential work on psych-verbs in Italian. They wish to argue that (some) Experiencer subjects are underlyingly internal arguments, so they rely heavily on the ability to diagnose derived subjects. One criterion they use is their inability to bind anaphors, which generally holds for the Experiencers in question. However, they admit that with non-clitic anaphors the resulting sentences are not entirely bad; they rate them as "*?" or "(?)", where by the latter they seem to mean 'very close to fully acceptable.' They clearly consider the lack of total badness an important problem, since they propose an analysis to explain it. Parallel constructions in English, they point out, "are judged deviant to some extent," which they apparently consider to be support for the Italian data by implicitly assuming the same explanation in both cases. Another correlate of derived subjecthood is the impossibility of arbitrary PRO, but again there are generic contexts where the contrast is "weaker, but still detectable" as compared to specific event contexts, the predicted bad sentences being marked "?" or "??". (No definition is given for "??" in relation to "*?", but in general, people seem to assume that any rating containing a star is worse than one containing only question marks.) Here again, an explanation is proposed by Belletti and Rizzi. In both of these cases, in the absence of an explanation the marginality facts would undermine main arguments for their analysis. On the very next page, however, a "??" sentence is treated as bad with no further comment; ditto for "(*)" later on.[27] Why is it that some instances of marginality demand comment whereas others do not? Most likely because the authors have no explanation for the latter, but know that readers will not be upset if certain marginal data are left unaccounted for, as long as they are not systematic across a whole paradigm. But what constitutes a paradigm, and hence what constitutes a systematic pattern, is determined by their theory. Thus, we have another case of selective, theory-driven use of judgement data.

---

27 Belletti and Rizzi do not exhaust the possible annotations. In addition to the five already mentioned, they employ the standard "*" and show grammatical sentences as unmarked, for a total of seven, but there are others in the literature, for instance occasional examples of "**" to mean 'much worse than a sentence which is already pretty bad.'

Later we find an instance where (citing Burzio 1981) they equate two uses of question mark, claiming that the marginal status of one sentence is unchanged by the application of passive, i.e., they are equally marginal:

(69) a. ?John gave pictures of each other to the kids.
b. ?Pictures of each other were given to the kids.

Their argument is that binding requirements may be satisfied at D-structure, before passive movement, so no change in grammaticality is predicted. However, in the corresponding Italian cases the apparent surface binding violations (parallel to (69b)) are "slightly more awkward," "but the contrast is much weaker than cases involving violations of the Binding Theory" (p. 316). In the abstract, their argument takes the following form:

> The differences in grammaticality between sentences A and B are significantly less than between C and D.
>
> D constitutes a binding violation, but A and C are fine.
>
> Therefore, B does not constitute a binding violation because it is not bad enough.

The assumption is that all (Principle A) binding violations cause exactly the same change in grammaticality rating, independent of any properties of the sentences themselves.

It is interesting to look at the prose descriptions that Belletti and Rizzi use to accompany the various annotations of sentences. Their sentence (75), marked "?", is "more or less acceptable" but such sentences with one question mark "still produce a weak violation of the chain condition." That is, the sentence is bad enough that it must violate something, and just bad enough to be a violation of this condition, but must not be violating anything else or it would have to be worse. Examples labelled "??" are variously described as "quite strange" and "weakly deviant"; does this mean that the notation underdifferentiates, or is the descriptive prose merely being stylistically varied?

Despite their use of no less than seven degrees of grammaticality distinction, Belletti and Rizzi remark about some more sentences that "these judgments are extremely subtle, and the usual OK vs. * notation is perhaps not appropriate for characterizing such contrasts. In fact, examples like (79a)–(80b) are already quite marked; still, there seems to be a detectable systematic difference in the indicated direction" (p. 322). Now, for some reason, any "detectable" pattern warrants an explanation, although these differences

are supposedly subtler than the ones ignored earlier, since the notation can no longer capture the distinctions. After proposing an account of the difference, the authors then claim "independent support" for it by suggesting that another "subtle" contrast seems "exactly on a par with" the previous ones. That is, we must allow for judging strict equality, as well as inequality, in contrasting pairs of judgements.

In this paper we also find the paradigmatic case of comparison of degrees of badness, ECP versus Subjacency violations: "The relatively mild ill-formedness of (94b) [which they mark "??"] suggests that the empty category left after extraction is properly governed within NP, otherwise these examples would violate the ECP, and a stronger unacceptability should result" (p. 328, fn. 22). We conclude that people supposedly have an absolute sense of how bad ECP violations are, and this is not bad enough to be one. At issue for the authors are relative "amounts" of Subjacency violation, so the absence of the much worse ECP violation is crucial to their argumentation. But there is no general theory of which principles *should* cause worse violations, i.e. the theory makes no prediction in the case of, say, θ-Criterion versus Case Filter violations. The whole notion of relative badness is ad hoc, and used in just those cases where it is convenient.

3.4 The Interpretation of the Annotations and Degrees of Badness

Let us now turn to a different kind of problem with judgement data. There seem to be two distinct uses of marginality markings, chiefly question mark, in the literature. One use denotes variable inter-speaker ratings, i.e. the sentence is good for some people, bad for others. The second meaning is that (most) individuals rate the sentence as marginal.[28] One could imagine the conjunction of these situations as well. The same is true of disjunctive notations like "/" or "{ }": are both alternatives acceptable to all speakers, or are there two groups, each of which only accepts one? This situation is at best a notational inaccuracy that could be easily corrected by adopting new symbols. Unfortunately, there are cases where the surrounding prose description does not make clear which meaning is intended. An example of the second kind of ambiguity appears in verb agreement with nominative objects in Icelandic. Thráinsson (1979) gives the following datum and description:

---

[28] One does occasionally find "%" used to mark acceptance by some speakers but not others.

(3) Mér líkar/líka þessir bílar.
Me likes (3rd sg.)/like (3rd pl.) these cars (N pl.) (D)

"There are some idiolectal differences as to the preference of verb forms, but the fact that some speakers prefer the 3rd sg. form here indicates that the nominative NP is not perceived as the subject" (p. 466). One interpretation of these comments might be that there is between-speaker variation across degrees of preference for each form.[29]

Andrews (1990) examined this and a number of other subtle agreement phenomena in Icelandic. His work is unusual for a theoretician in that he actually reports the results of grammaticality questionnaires he administered. Although he was not interested in the nature of judgements per se, it is interesting that his results are very similar to those of Ross (1979), to be discussed in Chapter 3, §2, who did have that focus. He elicited ratings on a 6-point scale, characterized as follows (p. 203; this is one of the rare instances where an explicit meaning is given for the symbols):[30]

√: Completely acceptable and natural
?: Acceptable, but perhaps somewhat unnatural
??: Doubtful, but perhaps acceptable
?*: Worse, but not totally unacceptable
*: Thoroughly unacceptable
**: Horrible

He reports results on 20 sentences, with between 12 and 17 subjects responding. Of these sentences, only three were rated uniformly "√"; none were rated uniformly "*" or uniformly "**," and only two were rated as either "*" or "**" by everyone. (The overall patterns clearly match Ross's finding that judgements on good sentences are less variable.) But despite having access to such detailed information, Andrews fails to clarify the status of Thráinsson's variability, stating, "Either of the above [agreement variants] seems to be acceptable (on the basis of questionnaires returned by seven informants)" (p.

---

29 Newmeyer (1983) suggests that there is an even more basic inconsistency in the use of stars, question marks, etc., namely their indication of ungrammaticality versus unacceptability. My impression is that the authors reviewed here, like most current authors, intend the latter interpretation. McCawley (1985) explicitly states that he uses asterisks to mark "whatever kind of oddity of a sentence . . . that I am at the moment concerned with; thus I use it to report data, not conclusions as to 'grammaticality'" (p. 673). We note that he rejects the notion of grammaticality, as opposed to acceptability, anyway, feeling that there is nothing to be gained by classifying sources of unacceptability as being inside or outside the grammar.

30 Labov (1972b) gives the following definitions: ? = questionable, ?* = questionably ungrammatical, * = ungrammatical, ** = outstandingly ungrammatical.

212), which to me at least is still ambiguous. The implications of inter-speaker differences versus intra-speaker marginality should be clear. The former, if not reflective of extra-grammatical factors, demands different grammars for the two groups, whereas the latter demands a single grammar with a less severe constraint.

Another use of "?" is illustrated by Pollock (1989) in his widely-cited paper on the structure of IP, with regard to French sentences like the following:

(20) b. ?Je pensais ne pouvoir pas dormir dans cette chambre.

He says, "The question mark is meant here as an indication that (20b, d, f) have a very literary ring to them, not that they are unacceptable" (p. 375). Such data are in serious danger of being misinterpreted out of context, especially since on the very next page "?" is used to indicate marginal acceptability. On this next page we also find "(?)" indicating that some speakers find another sentence better than the question-marked one, although Pollock admits to having found some who hold the opposite opinion. One could see this either as again confounding marginality with inter-speaker variability, or as an indication that ratings may reflect arbitrarily-chosen subgroups of speakers. Fortunately, the contrast of "?" and "(?)" is not part of his arguments, but the difference between these marginal sentences and certain starred ones is crucial, in fact, to several arguments in Pollock's paper. Later on, "(?)" is used not to mean 'slightly better than ?' in a relative sense, but "perfect, with at worst a slightly literary ring," which might or might not correspond in an absolute sense to the prior usage; indeed, literariness and marginality might be separate dimensions of ratings altogether, which cannot be meaningfully compared on the same ordinal scale, much as height and weight as integers cannot.

Finally, we consider a conference paper by Browning (1987), wherein more than half of the example sentences are marked with some number of question marks or stars. Since the paper is concerned (among other things) with the definition of Subjacency, one of the few Government-Binding (G-B) principles that has graded behaviour in its very definition, it is not surprising to find extensive reliance on relative judgements. One proposal of Browning's is to account for the marginality of parasitic gaps such as her examples (1) and (2), cited below, by the same mechanism as paradigm Subjacency violations such as her sentences in (40):

(1) ?Which paper did you read before filing

(2) ?an artist that close friends of admire

(40) a. Which car is it time for John to wash
b. Who did John buy a suit to impress
c. What did John wonder how to fix
d. Who did they leave before meeting

> Consider first the degree of ungrammaticality which results from one barrier intervening between two points in a chain. Several examples are given above in (40). I have been assuming the standard judgement for parasitic gaps such as (1) and (2), namely, a mild marginality. If this marginality is due to the intervening barrier, then the severity of the violation is clearly in the ball park represented by (40). (pp. 68–69)

Two things are noteworthy here. First, as we have seen before, there is a "standard" rating for constructions of a particular type, independent of the sentences themselves. This is surely a huge idealization—there are experiments showing that identical structural violations are given different grammaticality ratings depending on their particular lexical content (see Chapter 4, §§3.4–3.5). To the extent that *linguists* give uniform ratings for all such sentences, it is much more likely to be because they recognize them (perhaps subconsciously) as instances of parasitic gaps than from any pre-theoretic goodness rating.[31] That is, they are judging conformity to structural patterns or sentence templates and then reporting the "standard" rating for that structure.[32] Second, we have another example of equating the badness of sentences and taking this as support for a common violation. Has it never occurred to anyone that two different principles could yield the same degree of ungrammaticality when violated? That is, just because two sentences are equally bad does not mean they violate exactly the same constraint(s), especially since most of G-B's constraints have yet to receive a "seriousness" rating. Furthermore, the aforementioned problem with standard ratings applies here too: could it not be a structural commonality shared by parasitic gaps and long-distance *wh*-movement that linguists are identifying, rather than an assessment of overall goodness? Browning goes on to draw support for her arguments from "facts" including comparisons among sentences with no question marks or stars whatsoever.

---

31 This hypothesis is supported by the fact that people's absolute judgements are highly unstable, as we will see in later chapters; identical ratings of the kind linguists claim to have could not arise on the basis of acceptability judgement alone.

32 I am assuming that sentence processing itself does *not* work primarily by templates.

3.5 Summary

Let us summarize what we as naïve linguists might conclude are legitimate uses of judgement data in linguistic work, on the basis of the articles just surveyed. We can appeal to at least six levels of grammaticality. We can use the same symbol for sentences where there is claimed to be a grammatical violation and ones where there is not. We can use the same symbols to represent variation along several dimensions, including between-speaker differences, marginality, and "literary ring." Where the theory predicts a sentence to be strictly good or strictly bad but judgements place it somewhere in between, we can ignore that. We can choose to represent notationally only the judgements of a subgroup of speakers who fit our predictions. We can judge sentences as bad enough or not bad enough to constitute violations of particular grammatical constraints, but the theory does not tell us which violations should be better or worse, and we can ignore any such differences if it suits our purpose. We can claim that all sentences containing a particular violation are equally bad, although we know this is not true and suspect that what we are really doing is identifying sentence patterns. We can appeal to equality and inequality of differences in grammaticality of pairs of sentences, e.g. A versus B is a greater contrast or is exactly the same contrast as C versus D. It is hardly surprising that researchers in other disciplines do not think linguists do good science—it is true!

In general, it is clear that subtle judgement data have become important to theoretical argumentation; if they were not crucial, surely they would be ignored—clear-cut data make a much more impressive case. We have identified three kinds of problems in the use of these data, one of which will be our main focus for the remainder of this paper. The first is that they are not systematically reported or notationally identified. The second is that they are used or discarded as it suits the linguist's fancy. The third is that their use attributes various sophisticated abilities to native speakers without any evidence that they are actually capable of reliably making the discriminations in question, and without any attempt to systematically control the process of obtaining these judgements. The first two problems are more properly examined under the rubric of philosophy of science, and are likely traceable to a lack of understanding of and appreciation for the complexity of the judgement process. The third falls in the domain of psychology, which has the means to determine what people can actually do and provide a method for collecting data when they do it. Subsequent chapters will address these goals.

4. Summary and Motivation

The aim of the preceding sections has been to build up the motivations for an in-depth investigation of the nature of the process of forming grammaticality judgements. We have argued that an understanding of this process would provide the basis for an objective way to establish which judgement data bear directly on the grammar, and perhaps how to "extract" grammatical information from judgements confounded by other factors. We have presented several examples showing that the days when linguistics had more than enough to worry about with uncontroversial, common-place judgement data are over, and that the sophisticated and complex judgements now in use by theoreticians assume much about human abilities that remains unproven, even unscrutinized. We simply do not know whether the questions we are asking people are meaningful and can be answered in any principled way. We have shown that there is much to be gained by employing the experimental methodology of social science to the gathering of grammaticality judgements, and that in the absence of such practices our data might well be suspect. But eliminating or controlling for confounding factors generally requires us to have some idea of what those factors might be; such an understanding can only be gained by systematic study of the judgement process. Finally, we have argued that by studying inter-speaker variation rather than ignoring it (by treating only the majority dialect or one's own idiolect), interesting facts come to light.

This general approach is not a new proposal; Levelt et al. and Bever have articulated the general direction with great foresight:

> Where do grammaticality intuitions come from? It makes no sense to assume a priori that the domain of linguistic intuition is a relatively closed one, as many linguists appear to do. Such intuitions are highly dependent on our knowledge of the world and on the structure of our inferential capacities. (Levelt et al. 1977, p. 89)

> *What is the Science of Linguistics a Science of?* Linguistic intuitions do not necessarily directly reflect the structure of a language, yet such intuitions are the basic data the linguist uses to verify his grammar. This fact could raise serious doubts as to whether linguistic science is about anything at all, since the nature of the source of its data is so obscure. However, this obscurity is characteristic of every exploration of human behavior. Rather than rejecting linguistic study, we should pursue the course typical of most psychological sciences; give up the belief in an "absolute" intuition about sentences and study the laws of the intuitional process itself. (Bever 1970a, p. 346; emphasis in original)

Elliot, Legum, and Thompson (1969) make the case for studying variation: "there are facts both about linguistic theory and about the grammars of particular languages whose existence will be obscured unless variation is taken into account" (p. 52); "at least some variation is not completely mysterious and seems amenable to statement in terms within the realm of linguistic theory. At the same time, linguists have a responsibility to determine what kinds of variation exist rather than ignoring variation by basing syntactic descriptions on trivially small numbers of informants" (p. 58). These authors go on to show that variability on theoretically important points such as the *do so* construction and reflexive anaphors falls into implicational hierarchies of acceptability.

Thus, the approach to be pursued in this paper is the examination and modelling of the process of judging grammaticality, including the role of the grammar and its relation to the other relevant mental components. Many of the reasons for this endeavour should now be apparent, but there are others too. In addition to the basic interest of modelling an intriguing form of behaviour, one that has been almost entirely overlooked in favour of production and comprehension modelling, we hope to integrate the existing research findings in this area by sorting out the facts from the specific theories proposed in each study, assess their consistency, understand where they fit into the bigger picture, establish which methodologies get the best results in terms of reliability, validity, and informativeness, and propose new experiments to fill gaps in our knowledge. While the psychology of grammaticality judgements might hold as many complexities and mysteries as language itself, that is no reason for despair or dismissal—it is all the more reason for us to begin the task of unravelling them.

5. Scope and Organization

I will conclude this chapter with an outline of the scope of the present investigation and the organization of the remaining chapters. Of necessity, given the time constraints involved, we cannot endeavour to treat the area of grammaticality judgements in its entirety; the boundaries I have drawn on two dimensions are somewhat arbitrary, but fairly sensible, in that a detailed and reasonably complete picture of one sub-area, arguably the most important, will be given. First, with regard to what the grammaticality judgements are judgements *of*, I will look only at the acceptability (and grammaticality) of word strings, i.e. syntactic as opposed to phonological well-formedness, although in a broad sense acceptability/grammaticality often includes conformity to the phonology as well, and outside generative grammar, even to other linguistic domains. Second, while

several sorts of experiments are potentially relevant to the area, I will systematically exclude a number of subject populations: there will be little mention of the judgements of second-language learners and non-native speakers in other situations; only a passing glance at the development of metalinguistic awareness, which has virtually become a field unto itself; and no data from aphasics or others with language impairments. Putting it positively, we will focus mostly on the syntactic grammaticality judgements of normal adult native speakers.

Since the present work is not the first to include a survey of this literature, it is worth a moment to acknowledge my debt to those who have gone before, and to point out the important ways in which the present work differs from theirs. Three of these differences are fairly obvious: since the current work post-dates the most recent of forerunners by two years, it encompasses more recent research; since it is considerably longer than the relevant portions of the other works, most of the material is presented in greater detail; and none of the others has included a psychological-modelling approach. Newmeyer (1983) devotes a chapter of his book to the data base of linguistic theory, but his goal is to defend, rather than to (constructively) criticize, the generative modus operandi, so we will end up disagreeing with many of his conclusions, despite citing many of the same sources. Chaudron (1983) deals only with psycholinguistic experimental work, but provides a useful chart-form summary of many of the studies we will discuss, including many procedural details that we omit;[33,34] however, at least half of his paper is devoted to studies of second-language learners. Labov (1975) takes a position quite sympathetic with our own, but is concerned mostly with sociolinguistic variation; while much of the experimental work he discusses is not directly relevant here, his methodological proposals have heavily influenced our own. Lastly, Birdsong's (1989) review of the literature, which occupies two of his chapters, overlaps considerably with our own, but lacks the sort of principled overall organization that I have attempted to provide; his orientation, like Chaudron's, is that of applying discoveries about grammaticality judgements to issues in second-language learning and teaching research. Nonetheless, many of his

---

33 To compare the results of studies on the basis of Chaudron's chart would be misleading, however; the experiments differed in ways too subtle and too complex for his categorizations to capture.

34 It will become apparent that our reports of experimental work are often concerned with two particular features of elicitation: the instructions that were given to subjects, and the evaluation scheme (rating scale, categories, ranking procedure, or whatever) that was used. The importance of these two factors will be discussed in detail in Chapter 2, §3 and Chapter 4, §2.1, respectively. For now, suffice it to say that they are perhaps the biggest reasons why virtually no two studies in this field are directly comparable.

methodological proposals have also been incorporated here. Thus, none of the major extant works have taken the position of one whose basic goals are those of generative grammar, but whose specific aim is to propose major changes in its treatment of judgement data. That is the gap that the present opus endeavours to fill.

The paper is organized as follows. In Chapter 1 we have summarized the history of the concepts grammaticality and acceptability, focusing on the ways in which grammaticality judgements are used by syntactic theorists today and arguing that such uses demand a careful examination of them, not as pure sources of data but as instances of metalinguistic performance. Chapter 2 is a discussion of several important issues that arise when this view of grammaticality judgements is taken: tasks one can use to elicit them, scales one can use to report them, how people might go about giving them, and how and what they might tell us about linguistic competence. It concludes with the presentation of our central hypothesis, which ties the very broad properties of the judgement process discussed in Chapter 2 to the more specific ones discussed in Chapters 3 and 4: we propose that the entire behaviour of grammaticality judgements is the result of interactions between primary language faculties of the mind and general cognitive properties, and crucially does *not* involve special components dedicated to linguistic intuition. Chapters 3 and 4 cover the major body of psycholinguistic research that has been devoted to discovering ways in which the judgement process can vary systematically with differences between subjects (Chapter 3) and experimental manipulations (Chapter 4). Chapter 3 covers individual differences in two major categories: endogenous or organismic, and exogenous or experiential. Chapter 4 covers treatment factors in two major categories: stimulus materials or what is to be judged, and procedural methods or how it is to be judged. In reviewing the literature in these two chapters, we attempt wherever possible to point to the parallels with other cognitive behaviours that our hypothesis predicts. Chapter 5 represents the integration of the substantive and methodological findings and discussions of Chapters 2–4, in the form of a model of linguistic knowledge that reflects what is known about linguistic intuitions, and a proposed methodology for collecting grammaticality judgements while avoiding the pitfalls of previous work and taking account of the conditions that have been shown to influence them. Chapter 6 summarizes what remains to be done, that is, directions that our work suggests could be pursued to advantage in future studies, and how mainstream linguistic theory might be affected by the growing body of research in this area.

Chapter 2

Judging Grammaticality: The Nature of a Metalinguistic Performance Process

Metalinguistic data are like 25-cent hot dogs: they contain meat, but a lot of other ingredients, too. Some of these ingredients resist ready identification.

(Birdsong 1989)

1. Introduction

The purpose of this chapter is to explore some of the basic qualities of grammaticality judgements and some current thinking on their theoretical status, as a prelude to examining detailed studies of their behaviour under various experimental manipulations in the subsequent two chapters. We will cite some experimental work, but also a fair amount of theoretical discussion.

The structure of the chapter is as follows. We begin by asking how it is that we can get people to judge grammaticality: what tasks have been invented for this purpose, and what their relative merits are (§2). Next, we consider what has been a most important and controversial feature of the judgements that result from these tasks, and one which distinguishes them from production and comprehension, namely that people seem to judge grammaticality in a graded rather than a dichotomous fashion (§3). This is perhaps the most widely-studied topic in the literature on grammaticality judgements; a major issue is how to get at these scalar judgements *reliably*. We proceed with some speculation on how the intuitions behind our judgements might arise, how a sentence is processed for judgement, the extent to which the hypothesized process could reflect the grammar, and how we might make it do so more directly (§4). Then we tackle more directly the suggestion, which has become almost unanimous among psycholinguists, that no privileged status can be accorded to judgement data over any other sort of performance data, therefore we cannot draw direct conclusions about the grammar from them. Numerous leading authors have made this argument in various ways; we review the major contributions (§5). We then refine this proposal to produce a more specific hypothesis concerning the

interaction of extra-grammatical factors in the judgement process and their relation to cognition in general (§6). This hypothesis will constitute a recurring theme to be tested throughout the remainder of this work. Finally, we conclude by relating the high-level properties of grammaticality judgements discussed in this chapter to the low-level properties to be examined in the next two chapters.

2. Tasks that Elicit Judgements of Grammaticality

In this section we look at some of the ways researchers have used to get subjects to express their opinion on the grammaticality of sentences. This list does not attempt to be exhaustive (see Labov 1972b, p. 106, and Bialystok & Ryan 1985 for other types of intuitional judgements), and we will concentrate on those methods that require the least amount of inference on the part of the experimenter: for instance, we will not be concerned here with inferring grammaticality on the basis of spontaneous conversations or texts, because the inference is problematic and because they do not involve judgement, which is the focus of our examination, in any explicit sense. We will, however, extend the term "judgement" considerably beyond the paradigm case of asking the subject, "Do you think this is a good sentence?" A number of other tasks have been used to get at these opinions in other ways, or gain additional information; many of these will be represented in the studies we review in subsequent chapters.

The first extension we can make is to supplement judgements in various ways. For instance, we can ask subjects to explain them. In the case of sentences judged bad, this can involve asking subjects why they feel the sentence is bad, and/or where in the sentence the problem is. While there is potentially a lot of information to be gained by this, there are problems as well. For instance, it is not clear that subjects will be able to answer such questions in all cases: as Birdsong (1989, p. 110) puts it, the response "ungrammatical" can result from a "rather vague, gestalt-like impression": It just sounds bad, while at other times, one can detect something specific that is deviant about the sentence. In would be an interesting study to examine under what conditions the two feelings tend to arise, assuming subjects can reliably report the difference; the question appears not to have been studied. There is another methodological problem as well, which is how to balance this task for the sentences that are considered good, to avoid a biased

procedure; it seems to make no sense to ask, "Why is this sentence good?"[1] To the extent that experimenters have worried about this, they seem to have used a paraphrase task instead, which is useful in the sense that it helps to ensure that the subject actually thought about the sentence and took the intended reading. Another type of extension of the judgement task, particularly useful in marginal cases, is to ask under what conditions the sentence could be grammatical, if any. This could refer to a number of different aspects: the context of the utterance (e.g. "Only in a cartoon world where toasters can think"), prosodic features (e.g. "It's OK with heavy stress on *dog*"), restrictions on referents (e.g. "It's good as long as *they* refers to something animate"), novel lexical items ("Then *of* would have to be a noun"), etc. A parallel task for the ungrammatical cases might be correction, i.e. asking the subject how to fix the sentence (e.g., what words to add or delete); generally, we are interested in the *minimal* necessary changes. If the initial judgement was scalar rather than binary, it may make sense to ask both kinds of questions about the same sentence.

We can also go beyond explicitly asking for grammaticality assessments and look to other "metalinguistic"[2] tasks in collecting this information. One simple variation is to request rank-orderings of sentences by grammaticality, a procedure we take up further in §3. One might also ask for a comparison of the *type* of violation in bad sentences, i.e. asking whether they are bad in the same way. Another interesting method makes use of ambiguity. If we have a sentence that is uncontroversially good under one reading, but questionable under another, we can ask subjects whether it is ambiguous, and then verify their answers by eliciting paraphrases of the readings they find. In fact, the latter task without the former can provide some of this information without putting subjects in a judging mode at all, which (it will be argued in §4) may be important. But the most widely-cited non-judgement tasks, which were very popular in the 1960s and early 1970s, are the so-called compliance tests; Quirk and Svartvik (1966) are often cited as their originators. The task in such cases is to transform a stimulus sentence in some way, e.g. convert it from a statement to a question, make it negative, switch the pronouns, etc. The

---

1 A linguist might respond to such a question by demonstrating that it can be generated from the available mechanisms, or that it satisfies all the relevant well-formedness constraints, depending on the theory, but naïve subjects cannot be expected to attempt this.

2 Even Birdsong (1989), whose entire book is devoted to metalinguistic performance, believes this term requires a "rather vague interpretation." Its most important feature seems to be the objectification of language, i.e. attention to linguistic form rather than content. He also suggests that it describes "language-related activities typically not associated with the casual conversation and listening of non-linguists" (p. 62).

experimenter is actually not interested in these operations at all, but in whether the subject changes the remaining portion of the sentence while converting it. For instance, in investigating the grammaticality of a bare adjective complement of *regarded*, the task might be to convert the sentence *He is regarded insane* into a question. The dependent measure is the number of "relevant noncompliances" (RNCs), subjects who change the relevant part of the sentence, in this case the complement (e.g. by the insertion of *as*), which constitutes noncompliance with the instructions, since subjects are told to make only the change that the experimenter requests (Greenbaum & Quirk 1970). An RNC suggests that the subject considers the original form ungrammatical, although we must be careful about potential interfering effects such as forgetting the exact syntax of the original. In their book, Greenbaum and Quirk describe in great detail several orally-administered "batteries" of such tests, as well as relative and absolute judgement tasks. The obvious question was whether compliance versus judgement tests gave the same results for particular sentences; unfortunately, their presentation does not allow a concise overall summary of the results,[3] so we are limited to somewhat vague generalities. There was a large degree of agreement, and the authors attempt to provide very detailed explanations of the minor systematic discrepancies. For instance, lexical co-occurrence restrictions are judged bad but not changed, whereas certain word order errors are changed but not judged bad. Tottie (1977) rightly cautions, in response to performance tests like these (although her comments could apply equally well to judgement tests), that we should ask ourselves

> whether we are actually justified, from a psychological as well as from a linguistic point of view, in asking subjects to substitute one word for another or to produce negative or interrogative counterparts of affirmative sentences. Obviously, the sentences produced in that way cannot *a priori* be assumed to be equivalent to spontaneously produced linguistic structures of the same type . . . However, we need to know a good deal more about the psychology of speech production before we can arrive at anything more than a very tentative evaluation of such tests. (p. 209)

It is a major goal of the present endeavour to address this issue, at least for explicit judgements.

A more recent technique for assessing grammaticality, which is just beginning to show its full potential, is the measurement of event-related brain potentials (ERPs).

---

3 Chaudron agrees with this assessment and leaves their studies out of his summary chart of experimental results.

These are patterns in the electrical activity of the brain as measured by scalp recordings during the presentation of stimuli, in this case, sequential visual presentation of words. Electroencephalogram readings are broken down into component waveforms and categorized by the direction of change of the potential, positive (P) or negative (N), and the latency in milliseconds (e.g. 300, 400, etc.) from the stimulus onset. Three ERP components have been identified with well- or ill-formedness of sentences: N400, P300 and P600. The N400 occurs when a semantically anomalous word appears in an otherwise coherent sentence, while the P600 is triggered by syntactic anomaly and P300 seems to be connected with well-formed completion of a sentence. Kutas and Hillyard (1983) teased out the triggers of N400 using three types of stimulus sentence: syntactically well-formed and coherent; well-formed but containing one semantically anomalous content word; and ill-formed but semantically coherent, containing mismatches of tense or plural morphemes. In all cases, the primary task for the subjects was reading for content, although they were warned that errors might appear. The experimenters found that semantic anomalies did elicit N400s, but grammatical errors did not, although they point out that the latter were considerably more subtle than the former. In a later study, Van Petten and Kutas (1991) compared syntactically well-formed anomalous sentences with random word strings and found that the ERPs elicited by the final word of each string differed significantly. They interpret the reaction to the well-formed string as belonging to the P300 class of ERPs, which occurs in a wide variety of tasks, but here seems to be associated with syntactic closure, the realization that a sentence is complete. P600 appears to indicate temporary parsing failures, whether or not they are subsequently resolved. In particular, it occurs in garden-path sentences at the point where the initial parsing choice fails, e.g. at *to* in (1a), but not at the same word in the superficially-similar non–garden-path (1b). A P600 has also been found to occur upon presentation of the word *was* in (2a), but not in (2b) (Mike Tanenhaus, personal communication).

(1) a. The stockbroker persuaded to sell the stock . . .
b. The stockbroker hoped to sell the stock . . .

(2) a. The lawyer charged the defendant was lying.
b. The lawyer charged that the defendant was lying.

Ideally, one would also hope to find a reliable ERP correlate of actual, not just temporary, grammatical violations, so that overt judgements could be avoided, but in practice this will not be so straightforward. Not only are ERP experiments costly and difficult to conduct, requiring very carefully controlled and unnatural reading conditions, but the inter-

pretation of the results is not straightforward. Still, we may hope that someday ERP studies will at least allow us to disentangle various sources of ungrammatical judgements.

There are obviously many important questions about the relationships among metalinguistic task performance, regular linguistic performance, and competence, many of which will be considered in §§4 and 5 below. (Many of the authors we will cite take the area of metalinguistic performance even more broadly than we have in this section, including tasks that do not bear on grammaticality, but we believe that their discussion applies equally well to our subset.) It is not even clear whether we should speak of metalinguistic indicators as a whole, because there is considerable debate as to whether metalinguistic skill is a unitary phenomenon: from a developmental perspective, Hakes (1980) argues that it is, whereas from a cross-cultural perspective, Scribner and Cole (1981) argue that it is not, because people who do well on one task often do poorly on another (see §5, and also Chapter 3, §4.2 for the latter). Birdsong (1989) views metalinguistic performance as a collection of skills, arguing that it shows three typical features of skilled behaviour: there are differences in the number and kind of skills that individuals exhibit; there are differences in the degree to which they exhibit a given skill; and the skills tend to improve with practice or training. The reader is referred to Birdsong 1989, pp. 51ff., for detailed evidence on each of these points, which are somewhat controversial (not all of his evidence comes from judgements); within the present paper, Scribner and Cole's work bears on the first two points, and linguist-nonlinguist differences (Chapter 3, §4.1) bear on the third. If Birdsong's view is more or less correct, then these skills can be expected to make their own contributions to grammaticality judgement results, separate from those of linguistic competence. Intertwined with these issues is the deeper question of whether we are really interested in what forms people actually use, as opposed to what they claim they use, or what they passively accept. We know that use of a construction does not imply acceptance and conversely (Greenbaum 1976a), and sociolinguists have long known that speakers may deny using forms that they actually use frequently in everyday speech; in fact, as Labov (1975) demonstrates, this phenomenon is not limited to socially significant linguistic variables. (Hindle and Sag (1975) give an anecdotal report of this denial phenomenon on a (presumably) non-social variable: acceptance of *anymore* with and without a negative-polarity context.) The various metalinguistic tasks will reflect the three sets of sentences—those actually used, those claimed to be used, and

those accepted—to different degrees.[4] Which set is most relevant may depend on whether one believes in linguistic competence as separate from production and comprehension mechanisms, a question we take up again in Chapter 5.

3. The Graded Nature of Judgements

The following passage, from R. Lakoff 1977, probably reflects the beliefs of most newcomers to linguistics regarding the possible grammatical status of sentences, although it is doubtful whether this view was ever widely-held among linguists themselves: "It was tempting to believe that linguistic markers, like other animals, came in pairs, and it was therefore natural to assume that grammaticality was an either-or question . . . this seemed to us the way things ought to be in a well-ordered universe, and we were still capable of believing, with our endearing childlike faith, that the linguistic universe was well-ordered" (p. 73). At the risk of disillusioning the reader, we can state quite uncontroversially that a dichotomous view of grammaticality or acceptability will exclude a huge amount, perhaps most, of the interesting facts about linguistic well-formedness. The idea that grammaticality is in some sense a matter of degree rather than of binary choice is not a new realization in linguistic theory; it dates back at least to *LSLT*, wherein Chomsky asserts, "there is little doubt that speakers can fairly consistently order new utterances, never previously heard, with respect to their degree of 'belongingness' to the language" (p. 132). So many studies have illustrated this point with regard to particular syntactic questions that it would be impossible even to list them here; we will begin with two examples that demonstrate the range of phenomena that can be elucidated by a non-dichotomous approach, then go on to examine in more detail two of the major research areas of this type, namely Chomsky's three levels of violation, and rearrangements of surface word order.

In our first example study, Marks (1968) looked at those most mystical of linguistic beasts, multiply self-embedded sentences. He instructed subjects to judge their grammatical structure, not their length, complexity, difficulty of comprehension, or frequency of usage. For sentences with up to five self-embeddings, his results showed a power-law correlation between degree of embedding and subjects' rating; that is, accept-

[4] For instance, Hindle and Sag (1975) present anecdotal evidence suggesting that the task of judging shows a bias towards incorrect rejections as opposed to incorrect acceptances, i.e. towards lower-than-deserved ratings, as measured by speakers' actual usage.

ability grew as a function of the number of embeddings to a constant exponent (about 0.25–0.30).[5] In the second study, Danks and Glucksberg (1971) looked at the effects of violating adjective ordering constraints (e.g. *Swiss red big tables* versus *big red Swiss tables*) using a ranking test with the six possible permutations of three pre-nominal adjectives. The results showed that the position of the adjective that was most closely related to an intrinsic property of the noun was the primary determinant of acceptability: the closer it was to the noun, the higher the sentence was ranked.

A huge amount of research in the area of degrees of (un)grammaticality was generated by Chomsky's proposal in *ATS* (pp. 148ff.) that the grammar predicts at least three levels of increasing deviance, corresponding to the violation of selectional restrictions, subcategorization, and lexical category requirements. As was the fashion at the time, this claim spurred a flurry of experiments designed to test these predictions on judgement ratings. In retrospect, this goal seems somewhat misguided, since Chomsky himself never claimed that these degrees of *grammaticality* would translate intro degrees of *acceptability*; in fact, as quoted in Chapter 1, §2, he explicitly stated that the two did *not* coincide;[6,7] predictably, the ensuing results have often been contradictory. Nevertheless, we can learn quite a bit about the nature of scalar judgement from these experiments. Our review here is necessarily selective; see Moore 1972 and works cited therein for further references.

Downey and Hakes (1968) studied the effects of the three aforementioned levels on acceptability ratings, paraphrasing and free-recall. The ratings were on a scale from 0 ("completely acceptable") to 3 ("completely unacceptable"); the authors state that subjects were given two examples of how a sentence could be unacceptable, but they do not elaborate. The order among subjects' mean acceptability ratings was as predicted, although the difference between subcategorization and phrase structure violations was not

---

5 Marks does not state what the upper bound of the rating scale was, or whether it was open-ended. A rating of 0 was to be assigned to completely grammatical sentences. Interestingly, he also points out that many of his subjects did not realize that the crucial sentences were self-embedded, as determined by a post-rating questionnaire, and suggests the results may have come out differently if this had been pointed out to them.

6 This is another area where there has been much selective interpretation of acceptability as bearing on grammaticality: if the results go the right way, they are taken as evidence; if not, they are dismissed as 'performance artifacts.'

7 It is rather ironic, then, that current work in G-B regularly makes such heavy use of relative degrees of badness.

significant. However, the recall scores showed a reversal of this pattern, with sentences containing selectional violations being harder to learn than those with subcategorization errors.[8] Moore (1972) set out to test a hypothesis somewhat more general than Chomsky's statement, which applied only to verbal features, asking whether there is an acceptability hierarchy created by the three types of violations, regardless of where in the sentence they occur. He also sought corroboration for the hierarchy from sources other than judgements, in particular, reaction time (RT) for subjects to make them. His prediction was that a severely ungrammatical sentence should be processed faster than a marginally ungrammatical one, because more thorough processing would be required to detect the subtler error. The first experiment in this program used a paradigm that recurred in many subsequent studies. Subjects were shown a sentence with a blank where a missing word would go (e.g. *Sincerity may ___ the boy*), then shown the word that filled the gap on a separate screen and asked to decide as quickly as possible whether the sentence would be "appropriately completed" by that word.[9] The incomplete sentences were designed so that there was no way of assessing their grammaticality until the missing word was seen. The sentences in (3), (4) and (5) below illustrate blanks in verb, subject, and object positions, respectively (shown by the underline under the subsequently-presented target word); in each case, the (a) sentence contains a lexical category violation; (3b) violates strict subcategorization, while (4b) and (5b) violate selectional restrictions between the verb and the noun phrase; (3c) violates a selectional restriction of the verb, while (4c) and (5c) violate selectional restrictions between the noun and its modifying adjective.[10]

---

8 Results from the paraphrase task were not quantitatively analyzable; the authors merely discuss what they believe the subjects' strategies were.

9 Moore apparently wanted to ensure that subjects took selectional restrictions into account in their decision. He says, "The [experimenter] explained to [the subject] that terms such as 'appropriate' and 'acceptable' were deliberately being used, instead of 'grammatical,' because of the fact that the inappropriate sentences were inappropriate for varying reasons, some more syntactic than semantic. Inasmuch as 'ungrammatical' is frequently employed as being synonymous with 'syntactically deviant,' such instructions attempted to preclude any such dichotomy being set up by [the subject]." Since his subjects were not linguistics students, however, one might suspect that all this terminology only left them more confused.

10 Moore did not consider strict subcategorization to be a property of nouns, and therefore could not make all (b)-level violations the same type. He seems to assume that Chomsky's theory predicts all selectional restrictions to be equally ungrammatical, so (4 b and c) and (5 b and c) should be equivalent. However, since we are not particularly concerned with the theoretical implications of the study, that need not concern us.

(3) a. Smart voters uncle honest politicians.
b. Noisy dogs growl night animals.
c. Catchy slogans believe unwary citizens.

(4) a. Modern wanders improve factory efficiency.
b. Sensible ideas distrust public officials.
c. Nosey ditches annoy suburb dwellers.

(5) a. Large factories utilize efficient hesitates.
b. Big corporations appoint many machines.
c. Factory foremen appreciate eager tools. (Moore 1972, p. 553)

The main effect of level of violation seemed to support Chomsky's theory: RTs increased from (a) to (b) to (c) sentences. Interestingly, the mean RT for filler sentences that were grammatical was between those for (a) and (b) sentences. However, there were several mitigating interactions. In particular, there was no difference between RTs for sentences like (3b) versus (3c), while (4c) and (5c) did show longer decision time than their (b) counterparts (but Chomsky's theory may not have predicted the latter difference). Moore takes this as evidence that checking grammaticality takes place in two passes: first the major relations between subject, verb, and object are checked, and then relationships within the NP constituents are examined. Under this view, both verbal subcategorization and selectional restrictions are examined in the first pass and have no differential status, as reflected in the RT data. Several results also point to the importance of the verb in determining requirements for the rest of the sentence: for instance, although (3c) and (4b) constitute exactly the same type of violation, (3c) took significantly longer to reject.

A second experiment examined whether grammaticality ratings of the same sentences on a 20-point scale would conform to Chomsky's hierarchy. A new group of subjects was told that a sentence was "acceptable" if it "could occur in normal, everyday usage";[11] if so, they were to rate it 1, whereas scores of 2 to 20 represented increasing ungrammaticality.[12] Once again, the main effect of level of violation was as predicted: mean ratings for (a), (b), and (c) sentences were 13.5, 11.0, and 9.2, respectively, but the

11 Moore does not explain why the definition was changed from that of the first experiment.

12 One good feature of the instructions was that they explicitly encouraged subjects to look over a few of the (practice) sentences, to get an idea of the range within which they were working. They were told to make use of the full range of the scale.

latter two did not differ significantly for verbal blanks, again contradicting the theory and several previous studies.[13] The general trends were also confirmed in a replication of ratings by 12-year-olds (Moore 1975), although they differed in rating sentences like (3b) better than (3c) and failing to rate lexical category violations worse than verbal selectional restrictions. Moore and Biederman (1979) attempted to distinguish various possible serial and parallel models that could account for subject-verb-object relations being checked faster than noun-adjective ones, using the same blank paradigm as Moore's first experiment but with sentences that contained *two* kinds of violation, e.g. *Old houses quarrel valuable relics*. If both kinds of violation are searched for in parallel, one would expect an average judgement speed gain on such sentences as compared to either of the two violations by itself (assuming the search for ungrammaticality is self-terminating), but no such significant gain was found. On the other hand, no significant slow-down was found, suggesting that the search does terminate when one violation is encountered. The authors take this as support for a serial model where subject-verb-object relations are checked before internal NP relations. A follow-up rating task showed that double violations did decrease the grammaticality of sentences as compared to single violations, so that this rating process, unlike the speeded binary decision, does not terminate on encountering the first violation.[14]

The three levels of violation are not the only theoretical proposals that have spawned experimental work on levels of grammaticality. A study by Marks (1967) was inspired by Chomsky's informal statement that some ungrammatical sentences obviously have more structure than others. Marks's hypothesis was that there is an additional feature in people's grammaticality judgements when it comes to bad sentences, aside from their status as described by the grammar, namely the serial position of the violation in the sentence. Since sentences are processed left-to-right, earlier errors should interfere more with processing, because early words prepare the processor for later ones, set up expectations and restrictions, etc. He constructed stimulus materials by taking simplex sentences and ones with infinitival clauses and reversing the order of two adjacent words in various positions, producing a paradigm like (6):

---

13 Moore suggests that other studies failed to control for the location of the violation, and hence would not have seen the crucial interaction.

14 This experiment used an unbelievable 100-point rating scale. The authors give no justification for it.

(6) a. The boy hit the ball.
b. Boy the hit the ball.
c. The boy hit ball the.
d. The hit boy the ball.
e. The boy the hit ball.

Sentences were presented in groups with order randomized and subjects had to rank them from best to worst English. As predicted, noun-determiner inversion was judged less acceptable if it occurred earlier in the sentence; also, sentences like (6d) were judged worse than (6e), although Marks points out that the two types of inversion are not the same, and serial position may not be the important factor here. But at least in the former case, it is hard to see how any traditional grammar would distinguish the grammaticality of the sentences, since such grammars treat all noun phrases as equivalent. Serial position thus constitutes a reasonable candidate for an extra-grammatical factor that contributes to acceptability.

Scott performed a series of experiments (Scott 1969; Scott & Mills 1973) along similar lines, except that he used a single basic sentence order (subject-verb-object-qualifier) and rearranged whole constituents rather than words. His subjects rated each permutation as "acceptably grammatical" or "not grammatical."[15] The percentage of subjects who accepted various permutations ranged from 100% to 0%; Scott takes the results to show that there are at least five degrees of grammaticality among these sentences, but this number seems to be taken out of the air. We should also keep in mind that, unlike Marks's subjects, Scott's were only giving good-bad judgements, so the gradations appeared only in the pooled results and do not bear on individuals. Scott tries to account for the numbers of acceptances on the basis of how many constituents were moved and in how many places the canonical constituent order was split. This index does not yield a perfect correlation with judged "grammaticality," so Scott and Mills looked for other factors determining the outcome, in particular, meaningfulness.[16] This turned out to show no significant effect, but a useful outcome of the experiment was that when the permutations were not presented all together with their canonical form, grammaticality was rated

15 There was a third choice, viz. grammatical but with a different meaning from the unpermuted sentence; we shall not be concerned with this possibility here.

16 Scott and Mills cite various psychological sources for their definition of meaningfulness as "the association value of a single written verbal unit," for which they use frequency of occurrence as a metric. This does not correspond to what other authors have meant by the meaningfulness of a sentence.

much lower, suggesting that people accept a sentence more often if they can see that it is a rearrangement of a grammatical sentence.

More recently, Crain and Fodor (1987) looked at the effects of different kinds of ungrammaticality on a sentence matching task, where the subject must decide whether two simultaneously-displayed sentences are identical. The basic finding was that number agreement and quantifier placement errors (shown in (7) and (8), respectively) increase matching times, while Subjacency and (certain) ECP violations (shown in (9) and (10)) do not.

(7) *Mary were writing a letter to her husband.

(8) *Lesley's parents are chemical engineers both.

(9) *Who do the police believe the claim that John shot?

(10) *Who did the duchess sell Turner's portrait of?

While previous work had attributed this difference to different levels of ungrammaticality, Crain and Fodor argued that it was due instead to the correctability of the error: the first two types of error are easy to correct automatically, while for the other two there is no obvious correction that can convert them directly into grammatical sentences. The claim is that if a correction is made, it must be *undone* in order to perform the matching task, since the subject must decide if the sentences are literally identical; in cases like (7) and (8) no correction is possible, hence the bad sentence can be compared directly. Forster and Stevenson (1987) question this interpretation, suggesting that the correlation with correctability is epiphenomenal and cannot be the cause of the observed time differences. Both sets of authors acknowledge that other factors are at work as well, but the possibility that the correctability of ungrammaticality could be a factor in relative ratings of acceptability should not be dismissed; whether it bears any relation to theories of grammaticality is a matter of theoretical debate.

Now that we have seen how graded judgements are used in linguistics, we must consider their theoretical and cognitive status. It is important to recognize that Chomsky's theory, like many others, assumes that absolute grammaticality exists, i.e. sentences that violate no constraints of the grammar are uniformly grammatical. If a sentence is less that absolutely grammatical, it must violate some constraint(s) of the grammar. Thus there are no degrees of grammaticality, but there are degrees of ungram-

maticality. (See Levelt 1974, vol. 3 and below for some alternative proposals.) In terms of string sets, then, we have a primary dichotomy of good versus bad, with no distinctions among the good sentences but graded distinctions among the bad. It is reasonable to ask whether there is any psychological evidence for this theoretical distinction reflecting cognitive reality: even though acceptability is affected by other factors, one might expect this dichotomy to "show through" them, if the other factors were relatively orthogonal to grammaticality versus ungrammaticality. I am not aware of any clear evidence of this sort. Ross (1979, reported in Chapter 3, §2) did make a distinction between good, marginal, and bad sentences (on the basis of questionnaire data) and found that judgements on the first class showed the least inter-speaker and intra-speaker variation, but his study was so methodologically naïve that this result cannot be taken as anything more than suggestive until further experimentation is done.

In our view a much more likely scenario is that grammaticality rating behaves in much the same way as conceptual classification ratings of the sort elicited by Rosch (1975): just as we can ask "How good an example of a bird is a robin/ostrich/butterfly/chair?", we can ask "How good an example of a grammatical sentence is *X*?", for any string *X*. The responses will likely spread along a *continuum* with no indication of a clear-cut break of the sort discussed above, provided they are not biased by a lop-sided rating scale. Kess and Hoppe (1983, p. 47) concur: "apparently shared linguistic abilities operate on the same type of a graded continuum scale that cognitive abilities of a more general sort do." We must be cautious in extrapolating from such a result (if it is found) to the nature of the grammar, however. For instance, Levelt et al. propose the application of fuzzy set theory to account for this. (See Chapter 5, §2 for another attempt to formalize judgement gradience.) But prototypicality effects do not necessarily imply the absence of an underlying discrete system. As G. Lakoff (1987) reminds us, Rosch herself never proposed that graded classification effects reflect degrees of category membership or representation in terms of prototypical features or exemplars; in fact, there have been empirical demonstrations to the contrary. Armstrong, Gleitman, and Gleitman (1983) applied Rosch's original experimental paradigms to uncontroversially discrete concepts such as even number and female: subjects had to rate the extent to which exemplars represented the meaning of the category, and were timed on their responses to true/false categorial questions; the discrete concepts brought out the same pattern of results as Rosch's original materials. Specifically, the goodness ratings for various exemplars were graded quite uniformly across subjects, and reaction times for deciding membership in the category were longer for the worse exemplars, again with as much cross-subject consistency

as for the taxonomic concepts that Rosch studied. Despite being able to grade exemplars consistently in this way, the subjects demonstrably knew that membership in categories such as even number was an either-or proposition. Which behaviour reflects their true cognitive representations of the concepts? Armstrong et al. do not see these results as contradictory, because there are two different tasks involved: judging exemplariness versus deciding membership. They discuss various possible theoretical accounts of this difference, assuming that the real concepts are discrete and suggesting possible origins of the gradations, e.g. as stemming from a quick, heuristic identification procedure. Lakoff argues against this last idea, proposing that prototype effects reflect a mismatch between potentially-discrete conceptual knowledge and the real world; for example, in the real world not all unmarried men are eligible to be married, and hence cannot be rated as bachelors to the full extent. However, there are still concepts for which there appears to be no discrete decision criterion, e.g. richness, and these also exhibit prototypicality effects. Thus, it appears that graded structure on prototype tasks tells us nothing about the nature of the underlying mental representations. Is the same true about graded structure in grammaticality judgements and its bearing on mental grammars?

Barsalou (1987) suggests that graded structure may be a universal property of categories, and that the properties of an exemplar that determine its goodness as an instance of some category can vary depending on the situation; these may include, but are not limited to, similarity to the central tendency, similarity to the ideals of the category, frequency of occurrence, context, etc. He summarizes as follows:

> The graded structures within categories do not remain stable across situations. Instead a category's graded structure can shift substantially with changes in context. This suggests that graded structures do not reflect invariant properties of categories but instead are highly dependent on constraints inherent in specific situations. (p. 107)

As we will see, particularly in Chapter 4, this view jibes well with the findings on grammaticality: judgements are not invariant, and any of a large number of factors can come into play. Barsalou also looked at intra- and inter-subject reliability across a wide variety of conceptual types. When people order exemplars by typicality, the average between-subject correlation is about .45; for the same subject judging the same stimuli on two occasions one month apart, it is roughly .75; in both cases it is the moderate exemplars (neither very good instances nor very good noninstances) that are the most unstable. (This again jibes with Ross's findings.) Barsalou goes on to argue that there simply are no invariant representations of categories in the human cognitive system: these are merely

"analytic fictions" created by psychologists; perhaps linguists should be added to the list of culprits. Nonetheless, he suggests that judging typicality may not use the same representations as judging set membership: the former may use probable properties, the latter discriminative ones.

However, it seems that the nature of the particular tasks used by prototype theorists (and linguists) inherently induces graded behaviour, independent of the nature of the underlying knowledge, so the status of that knowledge as discrete or continuous must be demonstrated by other means. Does this question have any empirical content, however? How could we ever know whether the grammar, if it exists independent of performance mechanisms, classifies sentences dichotomously? If performance mechanisms induce graded structure by themselves, and if (as we have been arguing) they can never be circumvented because competence is not directly accessible, then how the grammar itself classifies sentences may not be empirically determinable. There are many possible combinations of mental structures that could yield graded acceptability judgements. For instance, Fillmore et al. (1979b) argue that judgement ratings could reflect the interaction of discretely-varying elementary components that only give the appearance of continua; we would add that not all of these components need be part of the grammar per se. Carroll (1979) suggests that graded acceptability can result from a discrete grammar plus performance rules of some sort. In either case, neither grades of grammaticality nor grades of ungrammaticality would be part of the grammar. It could be that while fully grammatical sentences can be judged as such without much reference to their meaningfulness, interpretability becomes the major factor in judging ungrammatical sentences, that is, how close can we come to figuring out what the sentence was supposed to mean. (See Fowler 1970 for essentially this argument: he insists that "an ungrammatical sentence is an ungrammatical sentence is an ungrammatical sentence," regardless of how it might be interpretable on the basis of extra-grammatical information; others have claimed there is an identifiable class of semi-grammatical sentences.) We cannot even be certain that any *linguistic* component of the mind places sentences along a continuum, in light of the findings reviewed above. Such questions about the nature of the concept "grammatical sentence" may eventually be answerable, but for now we must leave them open and move on to a related question that likely *is* answerable, viz. can we obtain useful judgements of degree of acceptability from subjects?

It is a fundamental assumption throughout the present work that empirical facts are useful (and interesting) if they are systematic, because they must tell us something

about the minds of the subjects who produce them; it remains a matter of analytical interpretation to decide *what* they tell us. Thus, we must first determine whether graded judgements are systematic, and the results mentioned throughout this section strongly suggest that they are. The next thing one might ask is just how many meaningful distinctions of levels of acceptability are available, which would in turn start to determine a procedure for eliciting them.[17] Chaudron (1983) cites several psychometric studies showing that in general ratings scales increase in reliability with increasing numbers of levels up to 20.[18] Presumably this can be shown by giving subjects different sizes of scale on which to rate the same stimuli: if you have too few levels, people merge them arbitrarily, whereas if you have too many, people split them arbitrarily.[19] Thus the "true" number of distinctions will show the greatest consistency within (and perhaps also between) subjects. It follows that studies that choose inappropriate numbers of levels add spurious variation to their results, possibly concealing the effects they were looking for. However, as far as I am aware, such a psychometric investigation has never been done with specific regard to grammaticality judgements. It is at least possible in principle that different tasks reflecting different kinds of underlying concepts display different amounts of differentiation, so we suggest that such an investigation should be a high priority.

Even if we can find the "optimum" size of rating scale, there will still be problems with this measure of grammaticality judgements. One major problem is how to quantify inter- and intra-subject consistency, which is an important part of much work in this field: if we use a 20-point scale, should we require two subjects to give exactly the same rating of a sentence in order to consider them consistent? Would plus or minus one position be sufficient? What if two subjects show exactly the same distances between ratings of multiple sentences, but their absolute ratings are offset by some constant—can we merely say that one is biased towards more conservatism or more liberalism, and consider them fully consistent? Depending on the size of the offset constant, that may not seem appro-

---

17 One does not have to look very hard to find evidence that such procedural guidance is necessary. For instance, Greenberg and Jenkins (1964) asked subjects to make certain phonological judgements using *any* system of numbers: "Just choose some numbers that you feel comfortable with" was their instruction! They termed this "free magnitude estimation," and went on to explore how it correlates with fixed scales.

18 Snow (1975) points out the apparently contradictory finding that psychologists measuring attitudes have shown subjects find scales with more than seven points hard to use.

19 Since the number of levels used in grammaticality experiments has ranged from 2 to (at least) 100, the problem must be fairly widespread.

priate, but neither would a conclusion of total inconsistency. If we standardize means and standard deviations, can we be sure we are not throwing away real differences?[20] Similar problems arise if some subjects simply fail to use the whole range of the scale, which can easily happen unintentionally if they have no idea what range of sentence types they will see. (For this reason alone, practice trials with representative anchor sentences are a good idea.) If we are attempting to compare consistency of subjects between studies that used different rating scales, the consistency measure will have to be scaled accordingly. Such problems have prompted many researchers to consider whether, instead of asking for absolute ratings of sentences, we should instead require subjects simply to rank order them from most to least acceptable. This approach does have certain advantages. For one thing, psychometric research indicates that people are much more reliable on comparative, as opposed to independent, ratings (Mohan 1977). Rank orders also solve the problem of different baselines on a rating scale, and there are non-parametric statistical tests for assessing the consistency or correlation between sets of rank orders. They are not without problems, however. One is efficiency (Maclay & Sleator 1960): the amount of information one can extract from a given number of judgements is much less than for absolute ratings. While it does not require exhaustive pair-wise comparisons to come up with an ordering of a set of sentences, there is surely a limit to how many sentences subjects can handle in one group; then one must somehow elicit inter-group orders.

There is a further problem of interpretation because pair-wise differential acceptability may not be transitive: that is, a subject who judges sentence A better than sentence B, and also judges B better than C when considering them two at a time, does not necessarily judge A better than C when they are examined side-by-side.[21] Hindle and Sag (1975) cite an instance of something like this with regard to the sentences in (11) that contain *anymore*, although they only present group data.

(11) a. They've scared us out of eating fish anymore.
b. It's dangerous to eat fish anymore.
c. All we eat anymore is fish.

---

20 In doing so, we would be implicitly adopting the theoretical position that any such differences simply are not part of what we are studying. For instance, the fact that Speaker A could be consistently more conservative in grammaticality judgements than Speaker B does not tell us anything about their grammars. I do not think we are in position to say this with any degree of certainty.

21 A hybrid solution that solves this and some other problems is to *elicit* absolute ratings but *convert* them to rankings. Then circularity can never arise.

Twenty-two such sentences were presented to 36 subjects, who were asked to compare them and then give each a grammaticality rating on a 5-point scale.[22] Then it was determined for each subject which of a given pair of sentences he or she had rated more grammatical, or if they had been rated equal, and subjects were tallied on this basis. These researchers found that while more subjects preferred (11a) over (11b) than the reverse, and more preferred (11b) over (11c), more preferred (11c) over (11a).[23] They eventually concluded that these comparison data are spurious, essentially because they involve an apples-and-oranges comparison: the sentences are too structurally diverse and hence their grammaticality is not subject to the same determining factors. When Danks and Glucksberg (1971) encountered similar circular triads on an individual level, they took them as a measure of a subject's inconsistency. While a detailed examination of this issue would take us too deeply into psychometric theory, our purpose has been merely to point out that such methodological problems will have to be dealt with if this paradigm is followed.

It is also an open question what to make of discrepancies between absolute ratings and rank-orderings given by the same subjects, as have been found by Snow and Meijer (1977) (see Chapter 3, §2), for instance. Even if we can establish that the discrepancies are due to context or contrast effects from neighbouring sentences, this does not determine which kind of judgement is closer to "the truth." Greenbaum and Quirk (1970) also examined the question of intra- and inter-subject consistency, and this rating-ranking contrast in particular, using "evaluation" versus "performance" tests. The former involved a rating on a 3-point scale: "perfectly natural and normal," "wholly unnatural and abnormal," or "somewhere between." The latter involved the presentation of multiple variants of a sentence together, and required subjects to rank them as well as rating each. Again, our summary is necessarily imprecise, since these researchers describe in great detail numerous experiments with minor variations. They typically used groups of 20–30 subjects and found that cross-group consistency was quite high, with very few significant differences on judgements (and other kinds of metalinguistic tasks). Also, their design allowed for several sentences to be judged a second time: most sentences showed 90–95% consistency (measured as the number of subjects giving the same judgement both

---

22 Obviously these data are not quite equivalent to ranked comparisons, since a maximum of five distinctions could be made.

23 It must be acknowledged that the differences involved are quite small, with many subjects rating the pairs as equal, which is not surprising given the small size of the rating scale compared to the number of sentences.

times), but some were as low as 54%.[24] A very few sentences were both rated and ranked: the two measures generally correlated, but sometimes sentences that were rated equal were ranked differently, even though tied rankings were allowed. We could interpret this to mean that the 3-point scale was too limiting, not allowing enough "room" for the distinctions subjects wanted to make.

Yet another study comparing rating and ranking was conducted by Mohan (1977). Ratings were on a scale of 1 ("completely well-formed") to 10 ("completely ill-formed") that was anchored by an example sentence for each of the extremes (probably a very good idea). There were 11 sentences to be ranked; procedure was a within-subjects variable, the two tasks being separated by a 2-week interval. Unfortunately, the instructions seem a bit too usage-oriented: "Consider each of the sentences and decide if it would be possible that you would say this in conversation." The study was actually concerned in part with establishing whether ordinal scaling of sentences is part of the competence of individual speakers, as opposed to dichotomous judgements with different thresholds across speakers and arbitrary rankings among the good and bad sentences. Nonparametric statistical analysis showed that the cross-speaker agreement in rankings was much higher than would be expected under the latter interpretation. Mohan also found some evidence for a yea-saying factor, i.e. a tendency to favour accepting sentences regardless of their grammatical status, by correlating the number accepted by each individual on two unrelated sets of sentences; there was a small but significant positive correlation. As for rating versus ranking, correlating number of acceptances again gave a highly significant result,[25] although the correlation itself was modest (.57).

4. The Judgement Process

In this section we will consider what people might actually be doing when judging the grammaticality of a sentence. Just about the only fact we know for sure is that we do not know what they are doing. At all. What follows is some thoughtful speculation. Many researchers want to relate this question to another unanswered question, viz. What happens in the ordinary processing of a sentence that one performs during the course of a conversation or while reading? There are two extreme positions one can take on the rela-

---

24 See Chapter 4, §2.3 for the use of repeated judgement as a systematically-imposed manipulation.

25 Sentences rated 1–5 were treated as acceptances; subjects drew a threshold line in their rank orderings, which allowed the comparison.

tion between these two processes: they might be identical, or they might be totally different. In the first case, some might argue (this is perhaps the null hypothesis) that the only difference between processing for judgement and processing for conversation is that in the former case the reply consists of a "yes" or "no" (or a numeric rating or whatever), instead of a pragmatically-related utterance. Obviously, the decision between the possible judgements has to come from somewhere, but on this view the processing of the sentence itself is identical; the differences come in deciding a rating versus deciding what to say next, both of which are separate from the parsing, semantic analysis, etc. that go into decoding the incoming utterance. At the other extreme, one might say that judging is nothing at all like understanding and involves none of the same cognitive mechanisms. If you are told you will have to judge a sentence, you route it to the "sentence judging" processor in the mind, rather than the "sentence comprehending" processor; these two modules are entirely separate and may differ in arbitrary ways. (If this were put forward as a serious proposal, one would have to address the question of how and why such a mechanism would come to exist in the mind.) As with most interesting psychological questions, many researchers suspect that the answer lies somewhere in the middle: we hope reality is not like the second position, but fear it is not like the first either. Let us consider what positions the major researchers in this field have espoused, the extent to which these positions have empirical support, and what their implications are for getting at the grammar. Our own proposal on this issue will be presented in detail in Chapter 5, §2.

Among critics of generative grammar, much has been made of its heavy reliance on introspective judgements and their non-equivalence to production and comprehension, specifically the fact that introspection is a slippery and unpredictable beast. Indeed, many of the problems that led to the downfall of introspectionism in late 19th century psychology seem to apply to the work of linguists as well (Levelt 1974, vol. 3). (See Levelt 1972 for some more general background on the historical relationship between psychology and linguistics.) In the introspectionist paradigm, established by Wundt in 1879 in the first experimental psychology laboratory, trained subjects were asked to describe their impressions of a wide variety of physical objects and experiences (Grusec, Lockhart & Walters 1990). The idea was to describe internal experience in terms of "elementary sensation"; that is, rather than saying that one saw a book, one should relate the colours, shapes, etc. that were perceived. There were several problems with this approach. One was that the elements of most experiences simply cannot be discerned just by reflecting on them, just as one cannot discern the elements of water by looking at it (Dellarosa 1988). Another was the fact that Wundt's subjects were far from naïve with regard to the

experimental procedure: they had to undergo at least 10,000 supervised practice trials before they could be used in an experiment, during which time they were taught special terminology in which to describe their sensations. One cannot help but suspect that Wundt's own ideas on what experience was like took their effect during this "training" period, although at the time it was thought that subjects were merely unlearning "bad" perception habits. Each of these problems is probably applicable to the linguist's situation to some degree, but perhaps the most significant drawback, which led to the demise of structuralist psychology, is strikingly evident among linguists today. Dellarosa (1988) describes it as follows:

> Despite the careful training that observers received, agreement among introspective reports was the exception rather than the rule. It was not unusual to obtain markedly different reports from two observers who were exposed to the same stimulus. Such disagreements could not be settled in any scientific fashion owing to the inherently private nature of internal events. In more technical terms, introspection failed as a bona fide scientific method because it violated a fundamental rule concerning scientific investigation: that of independent access to both causes and effects. Although the cause (i.e., stimulus) was open to public observation, the effect (i.e., internal sensation) was not. Without such independent observation of the internal sensation, it was impossible to tell which of two conflicting introspective reports was the correct one. The conflicting reports could have arisen because (a) Subject A was truly experiencing a different sensation than Subject B, or (b) Subject A was experiencing the same sensation as Subject B but was misreporting it, or (c) Subject A was simply lying . . . There was no scientific way to determine which of these three conditions was true. (p. 5)

(Carden and Dieterich (1981) cite Ringen (1975) as espousing this view of linguistics.) We might hope that ERPs could help to sort out these possibilities. Even staunch supporters of the generative enterprise such as Newmeyer (1983) admit that the theory may well be skewed by artifacts of the introspection process, but resign themselves to this as part of the "early stages" of the field of linguistic investigation. We have suggested that a more useful approach would be to try to learn more about the creature rather than simply accepting its influence over us. If indeed metalinguistic behaviour can tell us anything at all about normal language use, how can we extract that information?

Graeme Hirst (personal communication) has suggested the following analogy to another type of judgement, namely food tasting. If someone asks you what you remember about last night's dinner, chances are you will not have much to say: unless there was something strikingly good or strikingly awful about the food, you will likely have only a general impression that it was OK, if no particular attention was drawn to it at the time you ate it. On the other hand, if someone offers you some food and asks you for your

impressions *before* you taste it, you will pay particular attention to the flavours, textures, aromas, etc., perhaps chew more slowly, and may be able to give much more detailed comments, e.g. concerning particular herbs you detect, how tender the meat is, and so on. Your host could ask you more detailed questions, too, like whether you think there is too much garlic in the tomato sauce. Intuitively, it seems at least plausible that the taste stimuli are being processed in a different way, or to a different degree, than if no attention were being drawn to them, and the same might go for sentence tasting.[26] Hirst also suggests a third, hybrid scenario, namely soliciting the opinion immediately *after* the tasting. If the question was unexpected, then tasting will have proceeded as usual (as in the first situation), but since no time has elapsed we may have access to information that will later be lost or forgotten, impressions that were induced by the stimulus but ignored because they were irrelevant. To the extent that processing for pre-warned judgement differs from regular processing, this last scenario could provide the best of both worlds: regular processing, but access to additional information, which Hirst refers to as the traces of processing. (See also the discussion of speed of judgement in Chapter 4, §2.7.)

If the reader has not wandered off to the fridge by now, let us apply these ideas more directly to linguistic judgements. (See also Birdsong 1989, pp. 202–203.) In the worst case, we could imagine that expected judgement causes people to revert to conscious reasoning *about* sentences, rather than processing *of* them. Consciously-known rules could be applied in this way to decide grammaticality, but the only rules about language that most nonlinguists have conscious access to are those learned in grade school, which tend to be of the prescriptive variety. Thus, subjects may reason that a sentence is ungrammatical because it ends with a preposition, since they remember a rule that states that this is a Bad Thing. Because prescriptive grammar does not necessarily have any relation to descriptive reality, such judgements are of no use to us. But what if we avoid these generally well-circumscribed cases; can we not then expect to avoid conscious processing? Hirst argues in the negative. In general, people seem to be able to invent spurious rules or principles as post hoc rationalizations of behaviour; why not in language?[27]

---

26 This argument has been made for other metalinguistic tasks as well. For instance, Kess and Hoppe (1983) suggest that in an ambiguity detection situation, looking for ambiguity puts people in a different mode than a paraphrase tasks where the stimulus just happens to be ambiguous, so different results can be expected.

27 A related point, made by Coppieters (1987), is that we cannot distinguish intuitions that come from the syntax versus semantics versus pragmatics portions of our grammars: "Such attributions . . . represent essentially post-hoc attempts to organize and make sense of the data. As with most mental activity, we become conscious of some of the products of our subconscious linguistic mechanism (not necessarily in a direct and unbiased fashion); but the system itself which gave rise to these intuitions remains sub-

In Chapter 4 we will report on studies where respondents who had to justify their grammaticality choices gave (by linguistic standards) quite outrageous answers. On the other hand, even if people sense their true "intuitions" about a sentence, they may not express them if they cannot fabricate a justification. Thus, conscious reasoning/parsing is a most undesirable situation. But does it ever happen? As drastic and ad hoc as it seemingly must be, would it not result in judgements so far from what we know about actual usage that the discrepancies would be strikingly obvious? Again, Hirst argues negatively. Perhaps there really *is* a huge shift, comparable to the difference between written and spoken language, which also may not be obvious until systematically studied. This is all the more likely, given that judgements typically involve such rarely-occurring forms anyway: the usage data with which to compare them are extremely sparse.

There is suggestive experimental evidence for this position. Nagata (1990) wanted to examine the extent to which ungrammaticality affects our initial parsing of a sentence, as opposed to our post-hoc evaluation of it. To do this, he measured RT for people to judge the grammaticality of sentences on a good-bad scale, and plotted it against their grammaticality rating on a scale from 1 to 7 (where 1 = grammatical and 2–7 represent increasing degrees of ungrammaticality), which was elicited after the timed judgement. To control for the length of the sentences involved, each sentence was presented in two parts, with the subject pressing a button to expose the second part and start the timed trial. Stimulus sentences were paired such that the identical target strings could be used as the second parts of two different sentences. The sentences were designed so that the target string completed one sentence grammatically, but the other ungrammatically, thus matching the length of the timed portion exactly. Nagata's initial hypothesis was that highly ungrammatical sentences would be reacted to more slowly than mildly bad ones, because minor violations could go unnoticed whereas major ones would disrupt parsing. His findings showed something quite different, however. When RT is plotted against mean grammaticality rating, the result is an inverted U-shaped curve, i.e. sentences of intermediate ungrammaticality took more time to judge than very good or very bad ones. (Incidentally, we note that this differs from the data that Moore (1972) reported: he found RT inversely related to severity of violation. We must keep in mind, however, that he was looking only at three very specific types of violation, which may not have encompassed the full range of possible severity: thus, his results may all come from

---

conscious" (p. 548). Nonetheless, a substantial amount of linguistic theory is built upon linguists' intuitions that marginality or unacceptability is due to a particular component of linguistic knowledge.

the higher end of Nagata's spectrum, with which they are consistent. Nagata's data confirm Moore's finding that completely grammatical sentences take somewhat longer to judge than the worst violations, presumably because the latter do not require the whole sentence to be read.)

There are many possible interpretations to such a general finding, but here are some speculations. Judging perfectly good sentences and very bad sentences is quick because their status as good or bad is immediately obvious. Marginal sentences, on the other hand, do not fall clearly into one class or the other, hence more time is required to decide for them. Severe ungrammaticality did not slow down parsing because subjects were not trying to analyze or comprehend the sentence in any normal manner: as soon as a violation was detected, the decision could be made, perhaps without even considering the remainder of the string. If this interpretation is correct, then it shows that the study really did not get at the "on-line" nature of grammaticality as it impacts on "normal" parsing: since subjects knew they would be timed on judgments, they went into "quick judging mode," which may be quite different from normal parsing for comprehension. Nagata's purpose would be better served by not eliciting judgements at all, but rather by assessing processing speed when subjects were engaged in reading for comprehension, as proposed below.

Obviously one possible course of action at this stage would be to look for more evidence of drastic differences between judgements and actual use, by employing corpus-based analysis for instance (but see Hirst 1981, p. 55 for problems with real-world texts, e.g. the fact that one can generate bad sentences in writing without realizing it). But if such differences are found, we will not be any closer to a general method for getting at the linguistic knowledge that underlies regular performance. So we will move on to how we *would* like the judgement process to work. In the abstract, it would be nice if the language processor could run as usual, but a homunculus could be allowed to inspect the process and then report back on what he has seen. He could then observe not only the fact that, say, the parser had failed to parse a sentence, but exactly where in the sentence this occurred and why. Unfortunately, there is no evidence to suggest that people can introspect on the language mechanism in this way. If that is not possible, then at least the homunculus should be allowed to inspect the state of the processor when it is finished, although being a rather robust device, it may have managed to get through the sentence somehow and left little trace of a problem. This latter method, the interpretation of the "trace of execution," is what we might hope to achieve through post-presentation testing.

Obviously, speed will be of the essence, because much research in psycholinguistics has shown that our memory for the form of an utterance decays extremely quickly as compared to its content (e.g. Sachs 1967 and many subsequent studies). Others have reasoned along similar lines:

> There is very little evidence in the literature that people *are* conscious of many of their own mental processes. Awareness seems to be restricted to the outcome or results of such processes, and if people do report on processes, this is—[Nisbett and Decamp Wilson (1977)] contend—usually a logical reconstruction of how such a result might have come about (often in the form of a motivation) rather than a memory trace of the process itself. (Levelt, Sinclair & Jarvella 1978, p. 7)

Let us go on to consider how this method could conceivably be used in linguistic judgement. The obvious problem is going to be that before too long, the subject will be "on to us," i.e. will realize that we are going to ask for judgements, and so may revert to "judging mode" on any sentence after the first. (It is not difficult to imagine bogus tasks that would keep the subject in the dark until after the first sentence was presented.) This seems to require interspersing judgement trials at a low concentration among nonjudgement trials, or at the very least making the distractor task sufficiently engaging or realistic that the subject does not have an opportunity to reflect on what is going on. For instance, we might present sentences in the guise of a text that has been translated from a foreign language (say, a play or television show), and ask the subject to point out places where the translation is bad English (or whatever language) while keeping track of the plot for a later recall test, which forces them to keep processing for content.[28] The alternative is to try to deduce judgements from some nonintrusive measure, so that the subject is never aware that grammaticality is at issue. For instance, we can simply ask subjects to read a passage for content while taking some standard measure of reading speed and location (e.g. eye tracking or word-by-word button pressing). It is reasonable to predict that when unexpected ungrammaticality is encountered, a delay in reading will result; we may even learn something about where in the course of processing the error was detected. (Kutas and Hillyard (1983) review some other on-line measures along these lines.) We could also use ERP measurements in this way, or look for people to do a "double take" (Newmeyer 1983). Of course, there are other variables that affect reading speed (and ERPs) that will have to be factored out, the sensitivity of these measures to structural violations may not be terribly high, and the concentration of ungrammatical sentences

[28] This idea arose from a suggestion by Bill Poser (personal communication).

still should be kept reasonably low. Any of these methods is probably best used as corroboration for data derived by other means.

We conclude this section by briefly describing the work of Bialystok and Ryan (1985), who have proposed a high-level model of language skill that attempts to unify the cognitive requirements of various metalinguistic (and linguistic) tasks we have discussed in terms of the demands they place on two fundamental dimensions of language proficiency. The first, which they dub analyzed knowledge, consists of explicit, structured knowledge about language that is accessible to conscious reasoning and can be manipulated in solving problems, e.g. explaining errors in bad sentences. While regular language production and comprehension make relatively little use of analyzed knowledge, metalinguistic tasks like judging grammaticality require considerably more.[29] It is this type of knowledge we have alluded to above. Their second dimension is labelled cognitive control. This is a skill required for focusing one's attention on particular information and attending simultaneously to multiple facets of a stimulus, e.g. its form, meaning, and context, coordinating them within time constraints imposed by the task. Behaviours that have become automatic, e.g. attending to the meaning of a conversational utterance, require very little cognitive control, whereas moving one's focus away from meaning and onto form, as in judgement, requires considerably more. This may be a large part of what happens when we go into "judging mode." Thus, on both counts metalinguistic tasks are more demanding than conversation. Also, since the two dimensions are theoretically orthogonal (although in practice there is a correlation across the tasks people actually perform), we might expect that people's proficiency can vary along them and each could be subject to improvement through training or experience. (For instance, as will be argued in Chapter 3, schooling and literacy may contribute to such improvement;[30] experience as a newspaper editor will increase one's ability to detect errors in written text; linguistic training might also be expected to improve one's abilities, but the matter is not nearly so simple—see Chapter 3, §4.1.) Also, particular tasks and particular stimuli within those tasks will vary in the demands they make on the two dimensions. (For example, more salient errors require less cognitive control in order to be detected.) The authors propose that grammaticality tasks can be ordered by increasing amount of analyzed knowledge

---

29 It is not entirely clear from their terminology whether they believe that metalinguistic judgements always involve *consciously-accessed* knowledge.

30 Bialystok (1986, cited in Birdsong 1989) claims that schooling contributes most to development of the control dimension, whereas literacy increases analyzed linguistic knowledge.

required, as follows: grammaticality judgement, locating ungrammaticality, correcting ungrammaticality, explaining ungrammaticality. The sort of evidence they use to demonstrate such claims is that second language learners differ significantly on their ability to perform the various tasks even when precisely the same grammatical phenomena are involved in all of them. If it is true that various types of skills are involved in judgements, we must ask what the nature of the interface between the metalinguistic behaviours and the competence grammar is.

5. The Interpretation of Judgements with Respect to Competence

Many researchers have been convinced that there must be some differences between linguistic knowledge as revealed by judgements and that which underlies language use (e.g. Carden & Dieterich 1981). The question for those who accept this assumption then becomes whether and to what extent judgement data can be used as evidence of competence: if they are not "pure" reflections of that competence (as argued eloquently by Levelt (1974, vol. 3, pp. 5–7)), if they have "no special epistemological status" (Levelt, Sinclair & Jarvella 1978) vis-à-vis the grammar, then how can the impurities be removed? In this section we look at the views of a number of researchers who have made the further argument, in various ways, that judgements are somehow special or abnormal, unique among language behaviours and built on a different competence base. We will examine whether they have any evidence to support these claims, and whether they lead to any substantive proposals on how to make the best use of judgement data.

Bever has been the most widely cited proponent of the view that many of the properties that linguists attribute to the grammar, i.e. to a process-independent competence, really do not belong there at all, being in actuality properties of the *particular* behavioural process through which the data were obtained, be they intuitive judgements, production, or whatever:

> Even if our linguistic intuitions are consistent, there is no reason to believe that they are *direct* behavioral reflections of linguistic knowledge. The behavior of having linguistic intuitions may introduce its own properties . . . a linguistic grammar may have formal properties that reflect the study of selected subparts of speech behavior (for example, having intuitions about sentences), but which are not reflected in *any* other kind of speech behavior. (Bever 1970a, pp. 343–344)

A major proposal of the current investigation is that, if such properties in fact are not part of linguistic competence, they might be part of more general nonlinguistic cognitive sys-

tems, in which case we could expect them to show up in other tasks besides evaluating sentences. (This proposal will be made more explicit in the next section.) Bever goes on to make the distinction between properties of the linguistic processing mechanism and properties of the introspective process, neither of which should be reflected in the grammar, but both of which have played a role in grammar construction in actual practice. Thus, we may actually be constructing *two* different "contaminated" grammars.

> The relationship between linguistic grammar based on intuition and that based on the description of other kinds of explicit language performance may not just be "abstract" . . . but may be *nonexistent* in some cases. First, apparently "linguistic" intuitions about the relative acceptability of sequences may themselves be functions of one of the systems of speech behavior (for instance, perception) rather than of the system of structurally relevant intuitions. Second, the behavior of producing linguistically relevant intuitions may introduce some properties which are *sui generis* and which appear in *no* other kind of language behavior. (Bever 1970a, p. 345)

Thus, one of the general goals in this area should be to sort out which properties are attributable to which performance procedures, so that we can treat data from each source most appropriately, rather than trying to identify *general* performance artifacts that might actually not apply across the board.

Let us now look at some specific properties that judgement data have been proposed to exhibit, in contrast with usage data. Several suggestions come from work by the Gleitmans and their colleagues:

> We take judgments about language to be manifestations of an executive, or metalinguistic, skill that has psychological interest in its own right. The metalinguistic capacity shows more individual and population differences than the linguistic capacity; it appears relatively late in development; and it is sensitive to linguistic levels. Specifically, the more "surface" aspects of language are more difficult to access for the sake of giving judgments than are the "deeper" or more meaningful aspects. This distinction in performance may reflect differences in decay rates for less and more highly processed linguistic material. (Hirsh-Pasek, Gleitman & Gleitman 1978, p. 99)

> [Generative] grammars reflect the judgmental ("metalinguistic") aspects of language knowledge more directly than they do knowledge of language itself . . . Whatever differences exist between these organizations may derive from the fact that the "executive" thinking capacities have properties of their own, which enter into the form of the grammars they construct . . . Differences in tacit knowledge are small in comparison to differences in the ability to make such knowledge explicit. (Gleitman & Gleitman 1979, p. 121)

Let us consider the suggested properties one at a time.[31] First, it is claimed that metalinguistic abilities exhibit more individual differences than other linguistic abilities, and that different people's grammars are more similar than our externalizations about them would suggest. In fact, Gleitman and Gleitman go on to claim that people whose linguistic performance is *the same* differ in their metalinguistic performance. But how could we possibly demonstrate identical performance? Not surprisingly, the evidence provided to support this claim is rather nonspecific. For instance, they argue that there is more variation in learning to read than in learning to talk; but this may not be entirely a function of metalinguistic ability. They also argue that since even retarded individuals and most 4-year-olds achieve "adequate syntactic form" in their speech, we can conclude that there are few individual differences in syntactic usage. This is an absurd non sequitur. We suspect that this impression arises because metalinguistic tasks are typically used to probe areas of linguistic knowledge that rarely occur in regular speech, and that therefore likely do exhibit more inter-speaker variation than the most common sentence structures, but it remains to show whether differing judgements result from different grammars or from differences in the intuitional mechanism—this must be empirically determined. Another of their claims is that low-level properties are harder to intuit about than high-level (i.e. meaning-related) properties and that the latter intuitions show less variability. They paraphrase this by saying that "fully processed" forms of language are easier to judge than only partially-processed ones, e.g. syntactic forms without their semantics. Now, it is certainly true that meaning is the property of language we deal with and use most frequently, and so one might expect that meaning-related tasks such as paraphrase or ambiguity judgement would come more naturally than structural well-formedness judgements. But the actual evidence provided by Hirsh-Pasek et al. is not general enough to warrant their conclusion, since it all comes from the phonological domain: they show that children's word detection abilities are superior to their syllable identification, which in turn is better than their segment differentiation. Since the authors consider the word level to be "deeper" (more basic) than the syllable and segment levels, they draw the more general conclusion, but in fact meaning versus form is not the relevant contrast. At another point, Gleitman and Gleitman argue that giving judgements is more difficult than participating in conversation, by virtue of requiring "self-consciousness," i.e., taking a prior cognitive process as the object of a higher process. Fillmore et al. (1979b) concur that metalinguistic performance requires more skills than regular language use. But while the

[31] We defer discussion of their developmental argument until we have reviewed some of the relevant experiments below.

"objectification of cognitive processing" view has a certain analytic appeal, and may even seem intuitively right, we have no solid evidence that anything of the kind is actually going on; our impressions may be epiphenomenal. In all these cases, then, the proposals are unsupported; this is not to say that they are false, but one can envisage much more direct experimental ways of verifying or falsifying them, a worthwhile undertaking.

These authors have also gone on to make specific suggestions as to where we should look for the source of properties that are special to metalinguistic behaviour: they propose viewing it as an instance of the class of metacognitive behaviours.

> There need be no formal resemblance between metacognition and the cognitive processes it sometimes guides and organizes. Rather, one might expect to find resemblances among the higher-order processes themselves. On this view, judgments (and therefore grammars) have little direct relevance to speech and comprehension, but rather to reasoning. Whatever resemblance exists between language processing strategies and grammars may derive from the fact that the human builds his grammar out of his observation of regularities in his speech and comprehension. Whatever differences exist between these organizations may derive from the fact that the reflective capacities have properties of their own, which enter into the form of the grammars they construct. (Hirsh-Pasek, Gleitman & Gleitman 1978, p. 128)

(Bialystok and Ryan derive the same prediction from their model.) For the hypothesis that there are resemblances among higher-order processes to be of any use, we must find some other metacognitive tasks to compare grammaticality judgements to; unfortunately, very few have been discussed. In the domain of memory, recollection, i.e. knowing that you remember something, could be considered metamemory, and as a special case, the tip-of-the-tongue phenomenon involves metamemory in the absence of memory itself (Gleitman, Gleitman & Shipley 1972). The authors also propose that intentional learning, through deliberate memorization or other means, constitutes another type of metacognitive activity. But no one has yet illustrated how comparisons with such processes can shed any light on the nature of linguistic intuitions.

The arguments that we have seen so far for the secondary nature of grammatical intuitions have been based on comparisons between fully-developed linguistic versus metalinguistic abilities in adults. Another major set of arguments about the nature of linguistic intuition come from developmental work on the acquisition of metalinguistic abilities. This is a huge area in its own right, which we cannot hope to do justice here; see Chaudron 1983 and Birdsong 1989 for literature reviews. Instead, we will concentrate mainly on one research project that makes a particularly provocative suggestion about how metalinguistic abilities develop, viz. the work of Hakes (1980). His thesis,

with a Piagetian backdrop, starts from the observation that judgements and explanations of syntactic well-formedness emerge developmentally at about the same time as the ability to explain judgements of space and number and develop intentional memorization strategies, which is considerably later than corresponding production and comprehension abilities (which seem to appear in the preoperational period). He suggests that these are all forms of concrete operational thought, since they all involve controlled processes, whereas sentence comprehension and casual memory are automatic processes.

To test this idea, his experiments involved giving children ages 4 to 8 various metalinguistic tasks (comprehension, judgements of synonymy and acceptability, phonemic segmentation), as well as other cognitive tasks (e.g. conservation tests). His finding was not only that performance on these tasks shows improvements strongly correlated with age, but also that the nature of the improvements was similar, towards objectifying or "decentering," i.e. using controlled processing to stand back from and evaluate a situation, a process that Piaget attributed to the period of middle childhood. Hakes thus argues that a *general* metalinguistic ability underlies successful performance in all these tasks. For instance, synonymy judgements were based on superficial form in the youngest children, but on meaning and form together at a later stage. Acceptability for the youngest children was determined by whether they understood the sentence.[32] At a later stage it was based on the truth or desirability of the situation described in the sentence, its moral correctness, etc.,[33] while the older children generally used linguistic form, although even some 8-year-olds labelled sentences unacceptable due to falsehood of content. Another general trend was that fewer bad sentences were judged good as the child grew older, which Hakes interprets as indicating that more grammatical rules were being learned.[34] The claim that controlled processing is a crucial factor is supported by the fact that children seem to have the necessary skills to perform concrete operations earlier than they actually emerge: they have been known to display metalinguistic behaviour spontaneously in conversation, and can make use of deliberate memorization strategies when so

---

32 There is obviously a great deal of variation in the ages at which particular abilities emerge. Certainly it would be incorrect to say that children under 4 cannot assess grammaticality: Gleitman, Gleitman, and Shipley (1972) showed that 2 1/2-year-olds can detect and correct grammatical errors in simple imperatives, and that 6-to-8-year olds could correctly explain a wide variety of grammatical errors.

33 Such factors are not entirely abandoned in adulthood; see several studies reported in Chapter 4, notably Hill 1961 and Vetter, Volovecky & Howell 1979.

34 However, Hakes's procedure could have been subject to a response bias, since he asked the children to explain their reasons for rejection but not for acceptance (see Chapter 5, §3.2).

instructed, but until a certain stage do not seem to be able to choose the appropriate routines to fit the situation. (Hakes also provides an interesting discussion of the methodological problems in getting linguistic judgements from young children, which must be much harder again than getting them from adults.) To follow-up a point deferred earlier, Hirsh-Pasek et al. take data such as Hakes's to argue that metalinguistic ability emerges late in development, and therefore must differ in important ways from language use. They also report that children have been known to judge bad sentences as grammatical, even though they have demonstrably mastered the relevant grammatical form in their own speech. But if Hakes is on the right track in pointing to objectification as the crucial skill that must be added to comprehension processing to allow judgement, then it does not necessarily follow that this objectification distorts the data that it examines, and we certainly cannot conclude that there is a *separate* knowledge base underlying intuitions on this basis.

Besides adult native speakers and children, data from a third group, adult second-language learners, have been used in exploring the relationship between judgements and competence. Coppieters (1987) attempted an experimental demonstration that syntactic intuitions do not improve as speaking ability in the second language increases. Specifically, he wanted to show that native and near-native speakers could have identical linguistic performance but radically different intuitions, and then take this to support the indirectness of the link between language use and linguistic intuitions, i.e. as "a particularly striking illustration of the relatively independent status of two linguistic planes: language use and language form." His procedure began by finding non-native speakers of French who could not be clearly distinguished from native speakers in interviews that he conducted and who were considered to have native proficiency by their colleagues or friends; many of these subjects were linguists. He also interviewed a group of native speakers in the same way. He then proceeded with informal interview elicitations of judgements on a number of sentence types (subtleties of French syntax such as adjective placement, choice of past tenses, etc., usually with two alternatives) that were rated as "correct or good," "uncertain or problematic" or "incorrect or bad." Subjects were also asked to explain meaning differences between pairs of minimally-distinct sentences. The average ratings of the native speakers were used as a norm against which to evaluate individuals from both groups. Native speakers differed from their norm on 5–16% of the sentences, while near-natives disagreed on 23–49% of the judgements. Qualitatively, Coppieters reports that near-natives had strikingly different feelings about how sentences differed and the contexts where they could be used, and showed lots of variation in their

explanations, whereas the native speakers where quite homogeneous in their answers. But do these results really show intuitional differences in the face of identical performance? The fact is that the two groups were never compared on their *use* of the crucial constructions tested in the judgements (e.g. by injecting them surreptitiously in casual conversation), so it is equally possible (and seemingly more likely) that the same differences would show up in performance as well;[35] there might be no differential effect of judgement whatsoever in this case. (A related experiment by Snow and Meijer (1977) will be reported in conjunction with their other experiments in Chapter 3, §4.3.)

Yet another set of authors who have followed the same approach to argue that the degree of individual differences in metalinguistic ability implies that it relies on skills beyond those of language use are Masny and d'Anglejan (1985). In trying to pin down these skills, they looked at advanced ESL students for statistical relationships between second language (L2) grammaticality judgements and corrections and selected cognitive and linguistic variables: L2 proficiency, L1 reading competence, reasoning (non-verbal intelligence), field (in)dependence (a measure of cognitive style—see Chapter 3, §3.1) and others. Using multiple regression analysis they found that the best predictor of L2 metalinguistic ability was L2 proficiency; in apparent contradiction of Gleitman and Gleitman's claim, they found no correlation between metalinguistic ability and reasoning ability or cognitive style. Birdsong (1989) tries to make the same case on the basis of Scribner and Cole's (1981) study of Vai speakers in Liberia. Among these people, literacy and/or schooling seems to be a prerequisite for the ability to *explain* grammaticality judgements, but not for the ability to *make* them. We will discuss their findings further in Chapter 3, §4.2. Regardless of the questionable effectiveness of this particular line of argumentation, it is hard to dispute Birdsong's general conclusions about metalinguistic behaviour as a reflection of linguistic competence; in fact, much of the remainder of this paper will lend credence to it. Birdsong concludes, "Inasmuch as metalinguistic performance reflects idiosyncratic skill parameters, which vary across tasks and across individuals, it cannot, in any rough-and-ready manner, reflect the grammar or linguistic competence presumably possessed by all speakers of a language" (p. 61); "the inference of grammatical competence from linguistic and metalinguistic performance requires convergent evidence from a variety of validated sources, as well as a profound understanding of the variables that determine the form of the evidence" (p. 44).

---

35 Coppieters himself admits this as a likely possibility.

6. A Hypothesis

In this section we set out our own basic hypothesis regarding the interaction of metalinguistic performance factors and the grammar in determining grammaticality judgements. The proposal is formulated in terms of a generic effect on the judgement process, by which we mean any variable other than the grammar that can be shown to affect judgements. Chapters 3 and 4 are devoted to documenting these effects in detail, but we have seen hints of many of them already: linguistic training, pragmatic context, experimental instructions, literacy, etc. Our hypothesis, then, is that for any effect E_L on a language (judgement) task, there is an analogous effect E_C on a similar nonlinguistic cognitive (judgement) task. We have parenthesized the word "judgement" to indicate that we suspect that the truth of this hypothesis would extend beyond judgements to other metalinguistic tasks, although that will not be our concern here. In other words, our strong claim is that none of the variables that confound metalinguistic data are peculiar to judgements about language, but can be shown to operate in some other domain in a similar way. In practice it is not so easy to find convincing instances of such domains, however, since many cognitive processes may be mediated by linguistic representations, and we wish to claim that the properties in question are more global. The ideal candidates would be judgements in another sensory modality, such as taste, smell, or vision, which at least at a low level do not likely involve the language facilities of the brain. We cite just two arbitrary examples. In the visual domain, shape recognition and judgements of size, numerosity, etc. are potential candidates for parallels. Bergum and Bergum (1979a, 1979b) have found that in judging visually ambiguous figures (e.g. Necker cubes) certain individuals notice reversibility much more easily than the average; we might predict that these people also detect linguistic ambiguity more easily than average. In the perfume industry, experts are employed to smell products that are to be marketed and test for certain properties that nonexperts in this field have never heard of; they may differ from naïve perfume smellers in the same ways as linguists differ from naïve sentence judges. Note that while we are essentially claiming that different kinds of behaviour are the same in some ways, we can search for direct supporting evidence, because the hypothesis predicts the *presence*, rather than absence, of certain measurable effects. Such findings could greatly assist us in factoring out these effects from our grammatical judgement data, bringing us closer to the true representation of linguistic knowledge.

Thus, wherever possible in the following two chapters, we shall try to draw parallels between experimental results in psycholinguistics and known effects in other fields,

or propose a search for such effects. I would argue that this hypothesis is what one would expect to be the null hypothesis about the relation between language and other behaviours, and thus would not be a particularly surprising result. It is a natural position to take, and not without precedent. For instance, some have attempted to reduce all of our linguistic knowledge to general cognitive principles; Bever (1970a) takes the more moderate position that "certain aspects of linguistic structure are direct reflections in language of our general cognitive structure and its development" (p. 281). On the flip side, however, there have been countless studies that have concluded, on the basis of the manipulability of linguistic judgements (or their gradability or other properties), that the grammar itself must have these properties, or that they must be part of the language-specific component of the brain; I feel that such conclusions are not justified. However, we should note that if our hypothesis is supported, it still does not determine how cognitive principles and linguistic knowledge come to interact in the mind to produce linguistic judgements. There are (at least) two extremes of interpretation possible. On the one hand, it could be that these properties (e.g. context-dependence, susceptibility to training effects, etc.) belong to separate modules of the mind that are implicated in judgement behaviour but not in other forms of behaviour, e.g. a decision-reporting component. On the other hand, it could be that these properties are inherent in the cognitive substrate on which language and all other higher cognitive functions are built. Both outcomes have important implications that go far beyond our work here; my intuition is that each is probably true of some properties, but it will not be possible to settle the issue for any of them in the current work. In principle the two explanations are empirically distinguishable, however, since the modular theory predicts that there could be behaviours that circumvent the modules in question and do not show the relevant effects, whereas the substrate theory predicts that they are everywhere and inescapable. (These arguments are of course drastically oversimplified.) And if we should find that for a given effect there seems to be no parallel elsewhere in human cognition, then and only then would we have the beginnings of an argument for the special nature and encoding of language among human knowledge systems.

7. Conclusion

We began this chapter by considering various ways to elicit subjects' impressions regarding the grammaticality of sentences, considering the pros and cons of various methods. These should be kept in mind when examining the studies that are reviewed in Chapters 3 and 4. Along the way, we considered how judgements fit into the larger class

of metalinguistic behaviours, a theme to which we will return in Chapter 5 when we attempt to model the judgement process. We looked in detail at one heavily-explored property of judgements, namely their scalability and the methodological problems it raises. The next issue was the much broader question of what really goes on during the judgement process and how we might manipulate that process to keep it more in line with language use. We then reviewed several kinds of supposed evidence for just how far judgements seem to be from competence, our major determination being that they were inconclusive. Finally, we proposed that nongrammatical variability in grammaticality judgements has the same properties as variability in other cognitive domains, and suggested how we will explore that hypothesis in subsequent chapters. A derivative hypothesis is that the high-level features of judgements described in this chapter are the result of such lower-level features, which will be the focus of Chapters 3 and 4.

Chapter 3

Between-Subject Factors in Grammaticality Judgements

Speakers perversely disagree among themselves about what is grammatical in their language; some of the principal sources of suffering and dispute within generative linguistics have been over ways of coming to terms with such realities.

(Fillmore 1979)

1. Introduction

Despite their common genetic make-up, humans exhibit individual differences in virtually every aspect of behaviour; it should not be surprising to find that linguistic intuitions are no exception. The central question we wish to address in this chapter is the extent to which these differences are systematically attributable to differences either in properties of the organism or in its life experiences. In both cases, we will see that there are some features on which people differ that contribute rather transparently to their grammaticality judgements, and to linguistic behaviour generally, whereas in other cases the connection is surprising and still poorly understood. Throughout the chapter, consistency will be a major theme: the extent to which the same subject gives a sentence the same rating on different occasions, or different subjects give a sentence the same rating. In the former case, inconsistencies are liable to be the result of factors having nothing to do with subjects' linguistic representations, e.g. whether they are fresh or fatigued, uncooperative, attentive or distracted, etc. (Bradac et al. 1980). In the latter case, inter-speaker differences may be attributable to differences in deeper properties of the minds of the people in question: in their grammars or in some other module that affects grammaticality judgements. The implications of these various possibilities will be taken up in Chapter 5.

Our approach here will be to begin with three important studies that have looked quantitatively at individual differences in grammaticality judgements. The amount of variation found there will then motivate us to search for systematic factors that might account for some of it. In §3 we examine organismic factors in this regard. Two have been

studied extensively: field dependence, a concept from the personality literature (§3.1), and handedness, which seems to be an important indicator of linguistic structures in the brain (§3.2). Some others, such as age, sex, and general cognitive endowment, seem like obvious candidates but have been given little or no attention in the literature, so we consider them briefly in §3.3. In §4 we turn our attention to features of the person's experience. The most controversial and most discussed of these has to be linguistic training; innumerable critics of the linguistic enterprise have made their case on the basis of linguists being their own informants. We will look at several studies that have tried to establish whether linguists are suitable sources of grammaticality judgement data (§4.1). A less studied but very intriguing source of variation in judgement abilities may be the amount of literacy training and general schooling someone has received; investigations with remote cultures are the major source on this topic (§4.2). We conclude the section once again with a grab-bag of miscellaneous experiential factors, such as the amount of exposure one has had to a language (for instance, as a near-native speaker versus a native speaker) and accumulated world knowledge (§4.3). §5 concludes the chapter by summarizing the findings and using them to motivate the investigations of Chapter 4.

2. Individual Differences: Three Representative Studies

The term most often used for individual differences in language judgements is idiolectal variation, although Heringer (1970) is on the mark when he says, "This term is chosen for want of a better one and is not intended to imply that groups of people do not show the same patterns of variation in acceptability judgements, at least with individual sentence types. To call this dialect variation, however, seems not be appropriate since there do not appear to be geographical or sociological correlates to this variation" (p. 287). The single most widely-studied instance of individual differences comes from the interpretation of quantifier-negative combinations as exemplified in (1a), which might be paraphrased as (1b) or (1c):

(1) a. All the boys didn't leave.
b. Not all the boys left.
c. None of the boys left.

(Note that the spoken intonation pattern of (1a) likely would be very different for the two readings, although no one appears to have studied this issue systematically; see Chapter 4, §2.6.) In an early study, Carden (1970b) claimed speakers fell into three categories

with regard to their interpretation of sentences like (1a): some could only get the meaning (1b), some could only get the meaning (1c), and some found the sentence ambiguous.[1] He, among many other researchers of the day (e.g. Elliot, Legum, and Thompson (1969)), argued that there were important theoretical insights to be gained by examining the full range of dialects, rather than accounting for one and ignoring the others; he was particularly interested in finding implicational relations among dialect differences. In a follow-up study where Heringer attempted to elicit judgements on these sentences, he was faced with "the problem of asking naïve informants to judge the acceptability of ambiguous sentences on specific readings," a problem we have also encountered with regard to adjunct *wh*-movement (see Chapter 1, §3.2): since the sentence is uncontroversially good under one reading, one's initial impression is that it sounds fine; this undoubtedly biases ratings of other readings. Therefore, Heringer constructed a situational context in which only one of the readings was possible, either in the form of a scenario of which the target sentence formed the conclusion, or a prose description of the kind of situation where the sentence might occur. These two types of context are illustrated in (2) and (3), respectively:

(2) All the students didn't pass the test, did they? [Professor Unrat believes he finally has succeeded in making up a midterm which every single one of his students would fail miserably. However, he doesn't know the test results yet, since his poor overworked teaching assistant Stanley has just this moment finished grading them. Unrat asks Stanley this question in order to confirm his belief.]

(3) All the treasure seekers didn't find the chest of gold. [Used in the situation where none of them found it.] (p. 294)

Heringer's instructions then stated that acceptability should only be considered in the context of the material in square brackets. Unfortunately, acceptability was not defined for the subjects (a complaint made by Carden (1970a) as well) and they did not receive any training on practice sentences.

---

1 While these are the major dialects, Carden admits that he also found many subdialects. He also reports anecdotally that some speakers who originally could only get the (1b) reading started accepting both readings after repeated exposure to sentences which forced the (1c) reading.

At any rate, several interesting results came out of this study. One was the ability of context to prompt subjects to see potential acceptability where there otherwise was none, to be discussed in Chapter 4, §3.1. Another interesting finding was that while there were very few speakers who accepted only the (1c) reading, there were many more who accepted neither reading. In Carden's study this pattern had not shown up at all; in general, the results of the two studies differed quite substantially, leading Heringer to speculate on why this should be so. For one thing, the mode of presentation was different: Carden presented sentences orally in interviews, whereas Heringer used a written questionnaire. Another possibility, to be discussed more fully in Chapter 5, §3, is that interviews of the sort Carden conducted are more susceptible to experimenter bias. A third potential problem, mentioned by Carden (1970a), is that Heringer only used one stimulus sentence for each reading in most of the constructions, so it is worth asking whether peculiarities of the sentences chosen could be responsible for some of the results. Nonetheless, Heringer's data apparently refute Newmeyer's (1983) claim that people differ only on their bias of interpretation on these quantifier-negative sentences, i.e. which reading they think of first, but that everyone *can* get both readings. Even when context forced a particular reading, many of Heringer's subjects did not accept it, so subjects seem to differ on something deeper than processing preferences.[2] (See Labov 1972a for a survey of work on quantifier-negative dialects.)

Snow and Meijer (1977) performed three experiments to substantiate their claim about the secondary nature of syntactic intuitions and language data, which corresponds in many respects to our own position as presented in Chapter 2.[3] (The latter two will be discussed in subsequent sections.) Their first experiment used as subjects native speakers of Dutch who were studying linguistics but had not taken any courses in syntactic theory; we might expect them to show somewhat more sophistication than truly naïve subjects. Their materials all involved issues of word order, so multiple arrangements of each set of

---

2 Newmeyer cites a paper by Baltin (1977) to support his claim that everyone can get both readings, but in fact Baltin found nothing of the kind: he found the three dialects that Carden had reported, using question-answering rather than judgements as his primary source. (He also found a significant correlation between subjects' preferences on quantifier-negative constructions and their interpretation of prenominal modifiers as restrictive versus non-restrictive.) However, Labov (1975) does report results along the lines described by Newmeyer, where nonlinguistic tasks were used to force one reading or the other, with almost complete success across subjects.

3 They argue that syntactic intuitions are developmentally secondary, as evidenced by studies like Hakes's (1980), pragmatically secondary, because their function is not communicative, and methodologically secondary, as demonstrated by their experiments reported in this chapter.

words were constructed. There were two conditions: absolute judgements and rank-ordering. In the former, each of 24 sentences appeared on a separate page and the instructions stated, "Will you please read the sentence, then indicate whether you think it is a good Dutch sentence (by 'good' we mean 'acceptable in spoken language' and not 'grammatically correct'). Write + if the sentence is good, – if it isn't good, and ? if it is in-between or if you don't know." In the rank-ordering condition, the sentences were divided across four pages of six sentences each, and the instructions read in part, "Will you please rank these sentences within the groups of six by rewriting them at the bottom of the page with those sentences which are good Dutch, or the best Dutch, at the top and those sentences which are the worst Dutch at the bottom. Sentences which are equally good or bad can be written on one line." Immediately we see a potential confound, since the rank-ordering subjects were not told to rank by spoken acceptability as opposed to grammatical correctness (of course, we do not know whether this terminology was understood in a uniform way by the first group either). Snow and Meijer decided to make this a within-subjects factor, administering the two kinds of tests a week apart to the same subjects, and found no effects of the order of test types, but the instructions could still confound any differences between the two types of task.

The results were first analyzed for within- and between-subjects consistency in the two conditions. The between-subjects consensus on rankings was significant for all sets of sentences, as measured by Kendall's coefficient of concordance, but not extremely high (ranging from .466 to .670 on a potential range of 0–1): the most agreed-upon sentence, which the authors claim is perfectly normal, showed disagreement by 3 of 25 subjects, and all other sentences showed at least 7 disagreements as compared to their mean rank. The absolute ratings similarly showed no total unanimity, although there was one sentence type where 24 of the 25 subjects agreed.[4] On the other hand, five sentence types showed strong disagreements, i.e. at least one subject rated them bad both times while another rated them good both times, and two of these represented almost equal splits. Turning now to within-subject consistency, this rated at 70.8% for the absolute judgements, where two identical ratings for two structurally-identical variants counted as consistent, even if they were both marked "?"; the majority of inconsistencies involved one "?" rating rather than strictly opposed judgements. One subject out of the 25 was consistent on all 10 sentences, while the worst 2 were consistent on only 5. Snow and Meijer

[4] We must keep in mind that in the absolute condition, subjects could indicate that they were unsure, which the 25th subject here did; therefore, this constitutes only a weak disagreement.

correctly advise caution in interpreting this as a good level of consistency, however, because many of their subjects showed strong response biases towards "+" or towards "–": in the extreme case, someone labelling all sentences as good would be 100% consistent.[5] (Since we are not told the normative status of the stimulus sentences, we do not know what an unbiased distribution of responses might look like.) The authors devised a complex scoring system to assess within-subject consistency between rank-orderings and absolute ratings, which ranged on average from perfect consistency to about three out-of-sequence rankings in a set of six sentences. There was no significant correlation between this cross-conditions consistency score for a given subject and his or her consistency within absolute judgements. Even when judgements are pooled across all the subjects, the absolute ratings do not agree entirely with the rank-orderings: there was at least one reversal of position for each set of six sentences. On the basis of these results, it is hard to argue with the authors that "testing even a relatively large group of subjects, all of them relatively intelligent and language-conscious, does not assure internally consistent judgments concerning the relative acceptability of sentences" (p. 172).

Perhaps the most widely-cited study on individual differences in grammaticality judgements is that of Ross (1979). Ross asked 30 subjects to rate the grammaticality of 12 sentences on a scale from 1 to 4,[6] as well as eliciting their perceptions about these judgements. Specifically, the subjects had to state how certain they were of each judgement (pretty sure, middling, or pretty unsure),[7] and how they thought that judgement compared to most speakers (liberal, conservative, or middle of the road). Since we are particularly concerned with the design of instructions for such experiments, we present Ross's description of the rating scale as it appeared on his questionnaire:

---

5 Mohan (1977) cites psychometric work showing that there does not seem to be any general personality trait of 'agreement tendency' or 'yea-saying', but this leaves open the possibility for such biases in specific domains; his own study, reported in Chapter 2, §3, found some evidence for yea-saying on grammaticality judgements.

6 There were actually 13 sentences in his questionnaire, one of which was geared to the semantics of *barely* and *scarcely* and did not yield results comparable to the other sentences.

7 Chaudron (1983) points out that, unlike second language acquisition studies where subjects' self-ratings can be compared to their actual level of competence, here there is no objective way to assess the accuracy of the self-ratings.

> 1: The sentence sounds perfect. You would use it without hesitation.
> 2: The sentence is less than perfect—something in it just doesn't feel comfortable. Maybe lots of people could say it, but you never feel quite comfortable with it.
> 3: Worse than 2, but not completely impossible. Maybe somebody might use the sentence, but certainly not you. The sentence is almost beyond hope.
> 4: The sentence is absolutely out. Impossible to understand, nobody would say it. Un-English. (p. 161)

Note the reference to comprehensibility in item 4; in general, the instructions are quite explicit regarding differentiation of the levels, but give little indication of what counts as a criterion for grammaticality.

By Ross's own admission, this was intended only as a pilot study; as he acknowledges, his presentation of the results shows no knowledge of statistical analysis whatsoever; instead he invents his own numerical measures to assess variability, covariation, etc. and gives numerous large tables of raw data.[8] While these shortcomings make the paper tedious to read and the results hard to interpret, at least someone could use his raw data to do proper statistical analyses. I will report only the more obvious results, with the understanding that none of them should be taken as firm yet. First, we present the sentences employed in the questionnaire, with their mean ratings on the 1–4 scale (Ross did not calculate mean ratings, but computed an overall score by weighting the numbers of subjects who gave each of the four responses, in effect treating the scale as centred about a zero-point. Since his formula is arbitrary and unjustified, we use the standard computation instead. Thus, in his ordered list, the third and forth sentences are transposed):[9]

The doctor is sure that there will be no problems.	1.07	*Core*
Under no circumstances would I accept that offer.	1.23	
We don't believe the claim that Jimson ever had any money.	1.63	*Bog*
That is a frequently talked about proposal.	1.70	
The fact he wasn't in the store shouldn't be forgotten.	1.80	
The idea he wasn't in the store is preposterous.	2.03	
I urge that anything he touch be burned.	2.03	
Nobody is here who I get along with who I want to talk to.	2.60	
All the further we got was to Sudbury.	2.77	
Nobody who I get along with is here who I want to talk to.	2.83	
Such formulas should be writable down.	3.07	*Fringe*
What will the grandfather clock stand between the bed and?	3.30	

8 Another potential problem of interpretation is that 8 of his 30 subjects were non-native speakers of English.

9 The general problem of how to come up with a single rating for a sentence on the basis of multiple judgements on a graded scale has arisen in many other studies as well.

The designations "core," "bog," and "fringe" are used by Ross to refer to the range of good, marginal, and bad sentences, respectively; these divisions are made by eye-balling, not by any formulaic procedure.[10] He found three variables that correlated with this distinction in the order core-fringe-bog, i.e. that changed monotonically such that good sentences were at one extreme and marginal ones at the other: increasing variability between subjects, decreasing confidence in their judgements, and increasing self-rating as conservative. The finding about variability jibes with Barsalou's results reported in Chapter 2, §3 for conceptual typicality judgements. At an intuitive level, these results are not surprising, but the only explanation Ross adduces, namely that "the mind sags in the middle,"[11] does not add much insight. I suspect the same patterns would be found in conceptual classification tasks, e.g. rating the truth of classificatory sentences like *A robin/ostrich/bat is a bird.* While an additional goal of the questionnaire was to assess whether people know where their judgements stand in relation to the rest of the population, the data were not interpretable due to apparent misunderstandings of the liberality scale.[12] Interestingly, Ross found no cases of strongly "polarized" judgements, i.e. sentences that some people rated 1 and the rest rated 4, with no one in between. In all cases, the two most frequent ratings were adjacent on the scale, that is, there were no bimodal distributions. He suggests that this may be an artifact of the particular sentences chosen; if one deliberately chose known dialectal peculiarities, bimodality might still appear. However, as a measure of just how different people are, no 2 of the 30 subjects agreed on their ratings for more than 7 of the 12 sentences on the 4-point scale; in fact, Ross did not try all combinations of sentences, so it may even require less than 7 to differentiate them all. This is the striking result that leads Ross to ask, "Where's English?" (His proposed answer is discussed below.) One experiential factor that contributed to variability among Ross's subjects was that some of his subjects were linguists while others were not. He found systematic differences between the two groups, to be discussed below in §4.1.

---

10 Ross does not commit as to exactly where the divisions should be drawn for the sentences he studied, so we have placed the boundaries arbitrarily within his suggested ranges.

11 Attributed to George Miller.

12 Ross suggests that a better way to get at this information is simply to ask subjects directly what rating they think most other people would give.

Most linguists would acknowledge that no two people will agree on even binary judgements of a large collection of sentences, let alone ordinal rankings;[13] what, if anything, does this tell us about their grammars? Ross's data prompted him to take a very pessimistic view. He proposed in dismay that a language might be defined only as an *n*-dimensional space for some *n* in the thousands, where each point is a sentence and each dimension an implicationally-ordered axis such that acceptance of a sentence as grammatical on a given axis implies acceptance of all sentences closer to the origin along that axis. Then each person's idiolect is an *n*-dimensional vector specifying that person's acceptance threshold for each axis. Most linguists would find this an appallingly messy and uninteresting view of language.[14] We will discuss some alternative positions in Chapter 5. The reader is referred to Fillmore et al. 1979a for a very wide-ranging further discussion of individual differences in language behaviour.

3. Organismic Factors

3.1 Field Dependence

Field dependence/independence is a concept that originated in the personality assessment literature in psychology. It is meant to diagnose how people perceive and think, specifically the extent to which they perform *cognitive differentiation*, the process of distinguishing stimuli along different dimensions. Nagata (1989b) wanted to see whether field (in)dependence would influence grammaticality judgements. The field dependent (FD) person fuses aspects of the world and experiences it globally, whereas a field independent (FI) person is analytical, differentiating information and experiences into components; these are seen as more-or-less permanent traits of individuals (Weiner et al. 1977). There are a number of diagnostic tests for field (in)dependence that have been

---

13 Newmeyer appears to be the exception, claiming that "there is good reason to think that idiosyncratic (i.e., nongeographical and nonsocial) dialects are nothing but artifacts of the now-abandoned view that grammaticality is dependent on context" (1983, p. 57). However, he only cites one case as evidence for this very broad generalization, that of quantifier-negative sentences, and as mentioned in an earlier note, the crucial result is not found in his cited reference, Baltin 1977.

14 It seems to have originated in an earlier proposal of Ross's, the concept of a "squish" (see Hindle & Sag 1975 for a basic discussion). A squish is a two-dimensional matrix where the cells represent judgements; on one axis are forms graded by some property, e.g. increasing volitionality, on the other are environments where the forms might occur, graded by the extent to which they demand that property. One can then make claims about how orderly the implicational pattern in the matrix should be across speakers; unfortunately, it started to look like both hierarchies could vary across speakers, or even that this pattern could be violated by a single speaker through the syntactic analog of statistical interactions: the effect of one dimension on grammaticality depended on the level of the other.

shown by psychologists to be very well correlated. One of these is the tilting-room-tilting-chair test, which involves an apparatus consisting of a small box-shaped room containing a chair, mounted on mechanical devices such that each can be rotated independently in two dimensions. Subjects seated on the chair cannot see outside the room, and are required to judge whether they are seated upright or on a tilt, relative to the outside world. FD individuals tend to believe that they are on a tilt if the orientation of the room makes it appear so, i.e. they have trouble distinguishing the visual cues from the sensory ones, whereas FI's have less trouble. A simpler test to perform, used by Nagata to divide up his subjects, is the embedded figures test: subjects must rapidly pick out simple geometric figures embedded in larger, more complex ones; FD's have more difficulty with this than FI's. A priori, we might expect that these differences in cognitive style could show linguistic side-effects: FI individuals show an impersonal orientation and have well-developed cognitive restructuring skills, while FD individuals show more interpersonal competencies, e.g. they recall social words better than FI's and use them more often in free association tasks. Thus, we could anticipate that FD's would use strategies involving the enrichment of stimulus sentences with context when judging them, while FI's would be more prone to employ structural differentiation. (The nature of these strategies will be described in more detail in Chapter 4, §§2.4 and 2.5, in conjunction with Nagata's other experiments.) However, as reported in Chapter 2, §5, Masny and d'Anglejan (1985) found field dependence had no discernible effect on L2 judgement ability; they also review numerous other studies attempting to relate it to language ability, the results of which were mixed. An additional facet of this distinction is that FD's are more prone to changing their opinions under external influence, since they pay greater heed to others, so we should look for differential reactions to knowledge of other people's judgements.

Nagata's experiment involved repeated presentation of sentences: after rating the grammaticality of a number of sentences (on a scale of 1 to 7), subjects were exposed to each sentence 10 times for 3 seconds per repetition, during which time they were told to think of the grammaticality of the sentence. After the 10th repetition, they rated each sentence a second time. Then they were told that their judgments differed from those of the average college student (which Nagata considered negative reinforcement), and were asked to think about the grammaticality of each sentence again and rate it a third time. Other experiments have shown that for a general population, the repetition treatment makes judgements significantly more stringent (i.e. sentences are rated less grammatical afterwards)—see Chapter 4, §2.3 for details. In this experiment, the judgements of FI's did become more stringent after repetition, but those of FD's showed no significant

change. After the negative reinforcement, both groups' ratings became more lenient (the FD's non-significantly more so). Nagata concludes that FD's approach the task of judging grammaticality differently from FI's, since they resist the usual repetition effect; one might have expected their judgements to become more lenient with repetition, as they considered more potential contexts for the sentences, but this trend was not found either; apparently it is much harder to make sentences get better than to make them get worse (again, see Chapter 4 for more on this topic). The idea that FD's would be more responsive to negative reinforcement was not substantiated. In summary, we can say that field dependence is a factor that induces variability between subjects on grammaticality judgement tasks, just as it does in other domains; for instance, Lefever and Ehri (1976) found a "moderately positive relationship between sentence disambiguation abilities and field independence."

3.2 Handedness

There is already considerable evidence that handedness correlates with differences in language processing, e.g. from Hardyck, Naylor, and Smith (1979). Recently, some preliminary studies have been done on possible correlations between handedness and grammaticality judgement strategies. Work by Bever, Carrithers, and Townsend (1987) was the first to suggest that such differences might be found. The purpose of their study was to show that the assumption that the basic mechanisms of sentence processing are the same for everyone is a severe oversimplification. Specifically, they demonstrated how right-handers from families with at least one closely-related left-hander ('mixed background right-handers') show different processing patterns from right-handers with no familial history of left-handedness ('pure background right-handers'). The former group tend to process in a more structure independent way than the latter, that is, they attend less to syntactic and semantic structures of language and more to conceptual and lexico-pragmatic features. These differences showed up despite the matching of subject groups on several other variables, including age, sex, native language (English), and verbal SAT score. In one study the authors used the classic tone-location paradigm, wherein a subject hears a tone while listening to a sentence and must subsequently identify at which point in the sentence it occurred. They showed that mixed-background subjects did not show a superiority effect for clause boundary location of the tone,[15] that is, they did *not* locate

[15] A guessing bias towards this position was first factored out of the results.

the tone more accurately when the it occurred exactly between two clauses, while pure-background subjects did. A second experiment showed mixed-background subjects to respond more quickly in a word-recognition task (supposedly because they "make more use of the reference of individual words in their processing") and to be insensitive to the position of the target word in the clause, unlike their structure-dependent counterparts, who showed serial order effects. Pure background right-handers also performed more slowly on word-by-word reading tasks. These results support the authors' general conclusion that pure-background people depend more on aspects of sentence *structure*, mixed-backgrounders more on lexical and conceptual knowledge.[16] There is some neurological evidence to corroborate this proposal: familial sinistrality seems to be correlated with a less localized, more widespread language module in the brain, which Bever et al. suggest leads to more contact between language and other kinds of knowledge. Whatever the eventual explanations of these differences, it would not be surprising to find that the different processing strategies are also reflected in different judgement strategies between such groups. In fact, the two types of strategies proposed by Bever et al. are not so dissimilar from those proposed by Nagata for field dependents versus independents; a replication of his procedure with mixed-background subjects could prove fruitful.

Cowart (1989) conducted the first study to look explicitly for familial sinistrality differences to show up in a judgement task. The experiment involved a written questionnaire using a 4-point scale, the extremes of which were designated "OK" and "odd" (since the details of the procedure are not reported, we cannot assess the extent to which subjects were instructed on how to evaluate sentences in terms of these labels). The sentences in question followed the paradigm in (4):

(4) a. What did the scientist criticize Max's proof of?
b. What did the scientist criticize a proof of?
c. What did the scientist criticize the proof of?
d. Why did the scientist criticize Max's proof of the theorem?

(4a) has traditionally been called a violation of the Specified Subject Condition (now subsumed under Binding Theory), while (4b) and (4c) are considered good in some theories and claimed to violate only the lesser constraint of Subjacency by others; (4d) is an uncontroversial control sentence. It was hypothesized that since the violations in (4a–c) are

[16] It is important to note that there were no instances where the two groups showed reverse effects; either they showed the same trend to different degrees, or else one group showed no effect.

all of a purely structural nature, mixed-background subjects would be less sensitive to them and therefore rate them more grammatical than their pure-background counterparts. This prediction was borne out: for cases like (4a–c) the ratings of the former subjects were significantly lower than those of the latter, but no difference was found for grammatical control sentences like (4d).[17] If this insensitivity to structural violations is found throughout the syntax, it would constitute an explanation for a significant amount of inter-subject variation in judgements.

3.3 Other Organismic Factors

In this sub-section we suggest some other organismic factors that we speculate could induce systematic differences in grammaticality judgements. First, let us consider two of the most obvious: age and sex. Ross (1979) suggests that, in general, more contact with a language leads to higher grammaticality ratings for it, an idea inspired by the fact (reported below in §4.1) that linguists rated sentences higher on average than nonlinguists in his questionnaire experiment, which obviously has other possible explanations. If Ross is right, we would expect increasing age to be correlated with increasing tolerance of judgements. His own data do not bear this out, but they were not even based on accurate ages, just his guesses, so there is certainly room for more investigation here. As for sex differences, Chaudron (1983) states in his wide-ranging survey of metalinguistic research that sex has rarely been experimentally analyzed and "does not appear to be a relevant factor," but if the former is true then how do we know the latter for certain? R. Lakoff (1977), while dealing with what she calls "acceptability" differences between men's and women's speech, makes it clear that this is conditioned by situational and social factors, i.e. *when* a particular kind of utterance is appropriate, and not differences in grammars. For instance, she has found no instances of syntactic rules that only one sex possesses, at least not in English. However, Tom Bever (personal communication) *has* found preliminary evidence for gender differences in methods of language processing, which presumably could be reflected in judgements as well. He argues that there is a spectrum of possible ways one can develop abstract representations (linguistic or otherwise), whose extremes are hypothesis refinement versus hypothesis competition and replacement. On tasks such as maze learning and artificial-language learning, females have tended more toward the former end of the scale, males more towards the latter end. Thus,

[17] Another result was that cases like (4b) and (4c) were rated significantly worse than (4d), suggesting that they do indeed constitute Subjacency violations.

this too remains an intriguing area for future investigation, particularly if these gender differences also manifest themselves in the processing of individual sentences.

The second direction we might explore while looking for organismic factors are general cognitive differences that we suspect are implicated in the task of judging grammaticality. For instance, we will see evidence in Chapter 4, §3.4, that part of this process involves imagining a situation to which a sentence could be applied; therefore, the ability to imagine situations, i.e. some form of creativity, is a dimension on which people undoubtedly vary and that could correlate with judgements. Various "perceptual" strategies have been implicated in language processing, and hence also (somewhat controversially) in the generation of judgements; subjects may differ in their ability to use these strategies (Botha 1973). Similarly, a number of "extra-grammatical factors" often implicated in acceptability (as distinct from grammaticality) may be subject to inherent differences, such as working memory capacity, ability to reason by analogy, etc. At a more general level, intelligence and cognitive development may be pertinent, at least up to a certain ceiling. Hakes (1980, reported in Chapter 2, §5) attempts to show that qualitative changes in children's ability to make grammaticality judgements are correlated with Piagetian stages of development, and Masny and d'Anglejan (in the study reported in Chapter 2) looked for correlations between IQ and judgements of second-language learners, although they failed to find any significant patterns. Bialystok and Ryan (1985) propose a model of (meta)linguistic ability as factored into two major dimensions: analyzed knowledge and cognitive control. Each is the product of underlying cognitive abilities on which people may differ: analyzed knowledge is related to intelligence and logical deduction abilities, while cognitive control depends on reflective and impulsive tendencies, as well as field dependence, discussed in §3.1 above. They do not provide specific evidence for these interdependencies, however.

4. Experiential Factors

4.1 Linguistic Training

One of the most frequent criticisms of generative grammar has been the fact that, to paraphrase Labov, the theories that linguists develop are based on data that they themselves create, a situation that constitutes an intolerable conflict of interest and seriously undermines the external validity of the findings. In this subsection we will enumerate some of the specific reasons why it has been suggested that linguists' intuitions differ from those of "naïve native speakers" and thus should not be used as linguistic data; then

we will turn to experimental attempts to establish whether such differences actually exist, of which there have been surprisingly few. It must be kept in mind throughout that finding differences in the way linguists and nonlinguists judge sentences does not inherently count as a strike against using data from the former group: we must examine each difference to see what the potential benefits and drawbacks are for linguistic investigation.

The following passage from Bradac et al. (1980, p. 968) is typical of the views expressed by many outside the generative enterprise: "as a result of their special training, linguists may tend to judge strings differently from nonlinguists. Training in linguistics may produce beliefs or attitudes which are not shared by those who have not received such training. This suggests that the knowledge produced by linguists may become increasingly artifactual; it may fail increasingly to model natural language." While the authors' premise of differing beliefs is almost certainly true, it does not follow that linguists' judgements are artifactual in the sense that they are influenced by factors that are not relevant to the grammars of naïve speakers. A priori it is equally possible that their training allows them to factor out various irrelevant factors that *do* influence naïve judgements, but actually reflect cognitive factors *other than* the grammar that is the object of study (Levelt 1974). However, there are legitimate reasons to suggest that this ability of linguists may have come at the price of a loss of objectivity. Labov (1972a) argues that linguists have become removed from everyday experience. Greenbaum (1976a) believes that linguists are bound to make unreliable subjects, for at least three reasons. First, after long exposure to closely related sentences their judgements tend to become blurred; a famous quotation from Fraser (1971, p. 178, cited in Ney 1975, p. 146), exemplifies the point: "I think this issue is fairly clear. It will be resolved by speakers whose intuitions about the sentences in question are sharper than mine, which have been blunted by frequent worrying about these cases." Second, linguists are liable to be unconsciously prejudiced by their own theoretical positions, tending to judge in accordance with the predictions of their particular version of grammar.[18] (Botha (1973) also expresses this view, among many other critics. Additionally, Levelt (1974) suggests that hypercritical linguists might be biased *away from* the judgements predicted by the theory they are working on.) Carden and Dieterich (1981) hypothesize the subconscious process by

[18] Elan Dresher (personal communication) suggests that the "reputed argumentativeness" of linguists and the existence of multiple competing theories would guard against such bias. But this will only be true if those whose theories make a different prediction interact with the linguist in question; if the source of bias is an uncontroversial assumption within G-B, say, but disputed by proponents of Lexical-Functional Grammar, this cannot very well lead to discovery of the bias, because the two camps rarely interact.

which this could arise in a particular case. Third, they look for reasons behind their acceptance or rejection of a sentence, which takes away "spontaneity" and makes their judgement processes different from naïve subjects, who presumably have neither the inclination nor the knowledge necessary to perform this analysis. On the issue of whether this is actually less desirable, see our discussion in Chapter 4, §2.7 on the relative merits of spontaneous versus reasoned judgements. Nonetheless, we must agree with Greenbaum that this constitutes an additional difference between the two groups. Let us now see whether these hypotheses have been borne out empirically.

We begin with a brief summary of differences found by Ross in the study mentioned in §2, brief because its methodological short-comings make the results suspect at best. On average, his linguists were more unsure, i.e. had less confidence in their ratings, perhaps because thinking about language makes you realize how little you know about it and shatters your confidence in your own judgements—"Doing syntax rots the brain."[19] Nonlinguists rated themselves more conservative, were tougher graders (i.e. rated sentences less grammatical overall) and made fewer distinctions between levels of grammaticality (i.e. tended not to use the whole scale). We will find a counter-example to the relative stringency finding in another study.

The most widely cited work on linguist-nonlinguist differences is that of Spencer (1973). The paper is perhaps more important for the many issues it raises than for Spencer's experimental results. She starts from the position that

> it is possible that the behavior of producing linguistically relevant intuitions has developed into a specialized skill, no longer directly related to the language behaviour of the speech community (Bever [1970a]). The linguist views language in a highly specialized way, and perhaps is influenced by a perceptual set. The resulting description may not be an ideal representation of linguistic structure. It may be an artifactual system which reflects the accretion of conceptual organization by linguists. (p. 87)

Her experiment used two groups of subjects: the "naïve" subjects were students in introductory psychology, while the "nonnaïve" subjects were graduate students who had taken at least one course in generative grammar.[20] She states that Chomsky's (1961) definition

---

19 Ross attributes this adage to John Lawler without providing a reference.

20 Apparently the nonnaïve subjects did not possess a uniform amount of linguistic background, however, since some were graduate students in linguistics, while others were psychology or "speech" students. The latter groups may have "watered down" the linguistic biases of the first group.

of grammaticality and examples were used as the basis for the instructions in her experiment, but all she actually tells us about these instructions is the following: "Each [subject] was read the same instructions—he would be asked to make a decision on each statement as to whether it was complete and well-formed or not. There were a series of guidelines and examples as to what the [experimenter] meant . . . After the instructions had been read, the [subject] was asked to tell the [experimenter] what he had understood his instructions to be, and any confusions or omissions were corrected" (p. 91). Apparently Spencer (or her editors) did not consider it important to describe the details of these instructions, but they are crucial for interpreting the results: if they did not correspond to the concept of grammaticality that linguists use, then we have a confounding variable.[21] The stimulus sentences were drawn from six linguistic articles, and had all been labelled unequivocally good or bad by the original author. Unfortunately, none of the sentences are reported in the article; Newmeyer (1983) surmises, on the basis of the source articles, that many of them were pragmatically very odd and required an odd context to sound acceptable. Spencer's design was intended to draw out two possible results that would undermine linguists' use of their own intuitions: inter-subject variation by naïve subjects on allegedly clear cases, and naïve subject consensus that conflicted with a linguist's judgement. There was also a check for consistency: six randomly-chosen sentences were resubmitted for judgement at the end of the experimental session, and subjects who contradicted themselves on three or more of these had all their results discarded.

The first result was that an average of 81.4% of the 150 sentences were considered clear cases, as defined by the degree of consensus among subjects: at least 65% in each group gave the same rating (either good or bad, there were no other available answers). That is, the division between accepters and rejectors had to be at least 15% from an even split. But this does not strike us as a particularly strong consensus: 35% of the subjects could still have disagreed. If a 75% criterion had been set, the percentage of clear cases would have been lower; Spencer does not provide figures from which we can calculate it exactly. She acknowledges that her choice of cut-off is arbitrary. (For comparison, Snow and Meijer report 20% of their sentences as unclear cases among naïve native speakers; their definition of unclear is that a sentence received "approximately equal numbers" of acceptances and rejections.) As for whether naïve and nonnaïve subjects differed in their responses, it is impossible to be certain based on Spencer's reported figures, for two reasons. First, while she shows that the proportion of sentences accepted by the two groups

[21] Newmeyer (1983) makes this criticism as well.

differs by 6%, she reports no statistical test of significance for this difference. Second, this comparison would not reveal a situation where the groups differed on *which* sentences were accepted, but total *numbers* of acceptances happened to come out roughly the same. Spencer merely states that there were "no noticeable differences in the distribution of exemplars found unacceptable, unclear, and acceptable."[22]

As for comparing the subjects to the linguist authors, 73 of the 150 sentences showed disagreement, defined by the subjects' pooled rating being either unclear or opposite to that of the linguist; Table 1 (from Spencer 1973) gives a break-down of the results. Of the disagreements 81% were unanimous across the subject groups, and in the majority of the remaining cases it was the naïve subjects who disagreed with the linguists while the nonnaïve subjects agreed, but again this difference is not analyzed for significance. We must keep in mind, however, that this 50% disagreement rate is made up by comparing the pooled judgements of 65 subjects with those of individual linguists, a point that many subsequent articles have emphasized. Thus, while we can certainly conclude that the published judgements did not show a good correspondence with the population as a whole, we crucially cannot conclude that linguists *as a group* have systematically different judgements from nonlinguists: a comparison with any single randomly-chosen naïve subject could well have shown just as much disagreement. Nevertheless, Spencer tries to conclude that linguists should not trust their intuitions: "it is reasonable to state that the judgments of the linguists used are representative of many linguists as a group," since there had not been any published rebuttals in the 4–5 years since the original articles appeared, but there are many alternative explanations possible for that state of affairs and I remain unconvinced. As for the direction of the disagreements, the table shows that on 42 sentences nonlinguists were more accepting, while on 17 they were more stringent and on 14 they were mixed. This pattern, though not overwhelming, contradicts Ross's finding that linguists are more accepting on average.[23] Thus, the only firm recommendation we can draw from this study is that a reasonable sample size be used in determining the representativeness of judgements; we cannot conclude that this sample should not consist of linguists.

---

22 An "unclear" sentence is one on which the subjects did not show consensus by the measure defined above.

23 If we expect that linguists should be more aware of their actual speech tendencies than untrained speakers, then this result also contradicts the general recommendation of Hindle and Sag (1975) to trust "OK" judgements more than stars. If naïve subjects were unrealistically conservative, linguists ought to be more liberal.

Table 1

Comparison of Linguists' and Nonlinguists' Acceptability Judgements

Number of sentences	Judgement (+ = acceptable; – = unacceptable; ± = unclear case)			
	Linguist (as published)	Naïve group	Nonnaïve group	Total
Consensual Agreement				
51	+	+	+	
26	–	–	–	77
Consensual Disagreement				
17	+	– or ±	– or ±	
42	–	+ or ±	+ or ±	59
Judgments Mixed				
3	+	+	– or ±	3
4	+	– or ±	+	
7	–	+ or ±	–	11

Despite the less-than-convincing nature of her findings, Spencer goes on to point out that linguists who use only their own intuitions as data are really no different from trained introspectionists, whose intuitions ended up being totally removed from the layman's experiences (see Chapter 2, §4 for a discussion of introspectionism in psychology). In addition to the possibility that their theoretical perspective influences their judgements, she suggests that having worked with many sentences revolving around the same issue may also contribute to context biases in the judgements reported by linguists. Finally, she addresses the question of whether linguist-nonlinguist differences might not in fact be a good thing:

> It might be claimed that any difference between linguists and naïve speakers found in this experiment is due to the increased awareness and sophistication in language that linguists have developed through their study. Perhaps linguists are simply more sensitive to language and therefore are able to detect finer differentiations than naïve speakers in intuitions concerning natural language, rather than creating differentiations which do not exist within the natural language. If linguists are dealing with artifacts, however, nonnaïve speakers, who have studied modern linguistics, should perform in a manner similar to naïve speakers. Thus, to anticipate this criticism, nonnaïve speakers also participated in the experiment. (p. 90)

Of course there is a certain Catch-22 quality to this last point: one could always counter that however much linguistic training these nonnaïve subjects had had, it had not raised

them to the same level of linguistic sophistication as practicing linguists, and so the latter's judgements may still be valid. Conversely, if the nonnaïve subjects behave more like linguists than like naïve subjects, one could maintain that linguists' judgements were artifactual and that the nonnaïve subjects had too much linguistic training, such that they were exhibiting the same biases as linguists. Thus, it seems to me that subjects with some knowledge of linguistics can never be used to decide this issue definitively: what is needed is truly naïve subjects who nonetheless have been given a very good understanding of what is meant by grammaticality. (One might question whether this is possible even in principle, or whether the criteria are mutually exclusive.)

At least two other studies have attempted to compare linguist and nonlinguist judgements.[24] One of these, reported in a very brief article, is by Rose (1973). He also took his stimulus items from linguistic articles, asking subjects to classify them as acceptable or unacceptable (details of the method are not given). Half of the subjects were told to play the role of an editorial assistant working for a strict editor, while the other half had to play the role of a person attempting to help a foreign friend speak properly. Rose states that overall, subjects agreed with the linguist authors 89% of the time; we assume this is a percentage of the total individual judgements, rather than a pooled scheme like Spencer used. This number is not nearly as informative as Spencer's, since it could represent a variety of scenarios, e.g. each sentence showed strong agreement, or most showed uniform agreement and some showed uniform disagreement, etc. A chi-square analysis showed that linguist judgements and subjects' judgements were significantly related, but we have no indication as to which direction the disagreements went. There was no difference between the two roles played by subjects.

Snow and Meijer's second experiment repeated the procedures of the first, as reported in §2, but used eight linguists as subjects, allowing direct comparison with the results of their nonlinguist group. The linguists showed significantly greater within-subject consistency than the nonlinguists in the first experiment: 94.3% on the absolute judgements.[25] In part this may be attributable to a bias towards "–" responses, which exceeded

---

24 The only other empirical basis we have for comparing linguists and nonlinguists would have to come from separate studies that used the same procedure but with different kinds of subjects. For example, a study by Elliot, Legum, and Thompson (1969) used mostly linguists, whereas Greenbaum's (1973) replication, described in Chapter 4, §2.2, used all nonlinguists and got different results, but Greenbaum tried to eliminate other procedural problems with the design of Elliot et al., so the studies were no longer directly comparable. This is the only such instance I am aware of.

25 Their consistency between absolute ratings and rank-orderings was also significantly higher.

that of nonlinguists. (The authors do not report sentence-by-sentence comparisons, so we cannot say with certainty how often linguists were more stringent than nonlinguists; there is no basis for comparison with Ross or Spencer on this issue.) Linguists also showed greater between-subjects agreement, with Kendall coefficients of between .581 and .844. As for whether the linguists' judgements differed from the nonlinguists', the mean rankings of sentences by the two groups showed a high correlation (Spearman $\rho = .89$), as did the absolute ratings ($\rho = .84$). While this is a higher rate of agreement than Spencer found, we must consider that the present authors are using the mean ratings of a group of linguists, rather than a single linguist's judgements. Also, as they themselves point out, Spencer counted as disagreements any cases where nonlinguists showed disagreement among themselves; this was not taken account of in the present study. Thus, the two ratings are not directly comparable. The authors draw a number of methodological conclusions, including the interesting suggestion that while comparing absolute judgements with rank-orderings provides a useful check of judgemental consistency, the fact that a sentence is judged inconsistently may say more about the sentence than about the quality of the judges, for instance that it has some 'shifty' properties. With regard to the implications of linguists' higher consistency of judgement, they suggest two alternative interpretations: either linguists have learned to ignore minor irrelevant differences among sentences, e.g. their semantic plausibility, or they have learned to apply their theory to unclear cases. The extent to which each of these turns out to be right will obviously determine whether this improved consistency is a desirable property.

4.2 Literacy and Education

Birdsong (1989, pp. 31–44), Bialystok and Ryan (1985), and Masny and d'Anglejan (1985) provide extensive reviews of research examining the relationship between literacy, education and metalinguistic skills, including grammaticality judgements, and comment on the debate over which one(s) may be prerequisite(s) for the other(s). Bialystok (1986, cited in Birdsong 1989) suggests that schooling contributes to her dimension of linguistic control, implicated in the ability to objectify language for judging purposes, while literacy adds to one's analyzed knowledge. (See Chapter 5, §2.1 for more discussion of this model.) We present a few major studies here.

The largest and most fascinating project on this topic was conducted by Scribner and Cole (1981), who did several years of field work among the Vai people of Liberia. These people have invented their own syllabic writing system, which is taught to some

children in the home; formal schooling, for those who manage to get it, is conducted in English; some Vai also know Arabic. Scribner and Cole were interested in teasing apart the effects of schooling and literacy, and so the fact that there were Vai monoliterates who had had no formal schooling was crucial.[26] It was their hypothesis that writing contributes to the objectification of language, independent of any general cognitive advantages it might entail. (In fact, they found very little evidence that literacy in either Vai or Arabic produced advantages for problem-solving or other cognitive tasks.) More specifically, they believed that deliberate written composition in one's native language increases one's understanding of its formal properties, an idea that dates back to Vygotsky.

They used three kinds of metalinguistic tasks to test this theory. The first involved orally presenting paired sentences, one good and one bad, and asking subjects to choose the good one and explain why the other one was bad. Examples (5) and (6) below give rough English equivalents of the type of structures involved:

(5) a. He shot me at the gun.
b. He shot the gun at me.

(6) a. These children, what is its name?
b. These children, what are their names?

The second task called for subjects to explicitly identify some grammatical principle of Vai; this is illustrated in (7), where the relevant distinction is alienable versus inalienable possession.

(7) People say 'my (*ŋ*) father,' but 'my (*na*) book'; they say 'my (*ŋ*) sibling,' but 'my (*na*) wife.' Why do people sometimes say *ŋ* and sometimes say *na*?

(Apparently a wife is viewed as an acquired possession rather than a relative.) Subjects' explanations on these two tasks were scored on a scale of 0–7. Zero denoted irrelevant answers, such as "The old people say it like that," "Bad Vai," "Not a good Vai speaker," etc. A score of 1 was given to responses that claimed the sentence was semantically inappropriate, and higher scores denoted increasing degrees of grammatical relevance. While all groups were able to identify the bad sentence in the first task, their explanation abilities on the two tasks differed according to literacy and education: on one survey, the

[26] We should point out that theirs was a huge anthropological and psychological study, of which the metalinguistic tasks reported here constitute a tiny part.

average explanation scores were 3.9 for illiterates, 4.6 for Vai literates, and 5.6 for Vai-Arabic biliterates; a replication found scores of 2.3, 2.9, and 3.2, respectively. Multiple regression analysis showed that, of all the demographic data that were available about these subjects, Vai literacy was the only factor that predicted these differences.

The third task involved correcting errors of various types (shown in (8)) and explaining what was wrong with the sentence.

(8) a. My child is crying yesterday.
b. This house is fine very.
c. I don't want to bother you (plural) because you (singular) are working.
d. This is the chief's child first.
e. These men, where is he going?
f. They have planting the oranges.

This time explanations were scored 0–5. The authors provide Table 2 summarizing the number of errors correctly fixed and the total of the explanation scores on the six sentences. Here the regression analysis showed that schooling was the biggest contributor to explanation scores, and Vai literacy was also a factor. It is important to note that literates and illiterates had performed no differently on other tasks examining their ability to explain things, i.e. the effect is specific to the linguistic content of the problem. We can conclude from this work that literacy and schooling have little effect on the ability to identify ungrammaticality, and hence to make grammaticality judgements in the strict sense, but both factors appear to affect explicit grammatical knowledge, and hence will confound many other metalinguistic tasks.

Table 2
Comparison of Vai Error Correction and Explanation as a Function of Literacy

	Maximum possible score	Nonliterate	Arabic monoliterate	Vai monoliterate	Schooled literate
Number correct	6	5.1	4.5	5.0	5.6
Explanation score	30	6.9	8.1	9.9	15.7

Scholes and Willis (1987) studied 10 English-speaking adult illiterates and found that they seem to process sentences without making use of all the syntactic information they contain. For instance, they report anecdotally that a spoken sentence like *The win-*

dow in the room with the chair was broken is taken to mean that the chair got broken.[27] Birdsong (1989) cites other work by these authors suggesting that illiterates are insensitive to passive morphology, and that they judge grammaticality according to pragmatic validity and moral correctness or desirability. Scholes and Willis conclude that illiterates have "vastly different" grammars from literates, but Birdsong counters that their judgements may be based on different criteria, without the underlying grammars necessarily differing. Heeschen (1978) seems to have had similar experiences with the Eipo, an illiterate, neolithic horticultural people of West New Guinea. He states that they are "uneasy and unsuccessful" in trying to objectify language, and concludes that 90% of their grammaticality judgements of possible but rarely-occurring verbal affix combinations were simply wrong.[28] However, their judgements on word order were "absolutely correct." Heeschen suggests why this difference should be found: some affix combinations are rare and hence hard to see as correct out of context, whereas word order is a feature of every utterance that cannot be avoided. This hypothesis is supported by that fact that in *natural* situations (e.g. when native speakers corrected him in conversation), as opposed to structured judgement tasks, "their judgments as native speakers proved to be perfectly reliable" (p. 177). Thus, at least for this culture, it seems that illiteracy does not imply the inability to make accurate judgements, but just makes it hard to do so in an abstract context.

4.3 Other Experiential Factors

As in the previous section, we will conclude with a collection of remarks on other types of experience that might systematically affect judgements of grammaticality. The most obvious would be the amount of experience with the language in question. There have been numerous studies of metalinguistic skill in non-native learners of a second language, as part of the second language teaching literature, which is beyond the scope of this investigation. Clearly, one would expect non-native speakers to differ from their native counterparts in judgements as well as in language use, but the results of a third experiment in Snow and Meijer's study (cf. §§2 and 4.1) suggest that native intuitions may be acquired before native skill in language use.

---

[27] One might suspect the presence of some third, pathological factor affecting both ability to acquire literacy skills and ability to comprehend sentences, but Scholes and Willis's very brief description gives no indication of such a factor.

[28] He does not explain how he determined what the possible forms actually were.

This experiment involved the same procedure as before, this time with non-native speakers of Dutch as subjects. Their within-subject consistency was at least as good as native speakers, but predictably they showed more between-subject disagreements, since their degree of familiarity of Dutch was not matched. Nonetheless, their pooled judgements agreed somewhat better with the native speaker group than those of the linguists did. And surprisingly, the three virtually bilingual non-natives did not match the native group better than the remaining poorer Dutch speakers (as measured by correlations in rank-ordering). The authors interpret this to mean that one's skill in speaking a language can improve without one's syntactic intuitions becoming more native-like;[29] conversely, they suggest that classics scholars, for instance, show the opposite: they develop strong intuitions without being able to speak the language. Together with the large amount of variation in judgements among native speakers found in the first two experiments, their results lead Snow and Meijer to conclude that speaking and understanding involve a different language faculty from judging, since skill in one is not a good predictor of skill in the other. On the other hand, Coppieters (1987) claimed that his subjects appeared to have achieved native levels of production and comprehension, and yet showed significant differences from native judgements. But as discussed in Chapter 2, §5, the study had not actually shown that the two groups were identical in their *use* of the crucial forms, but only on unrelated general measures of fluency, mastery of various constructions, etc. Thus, we have no basis for concluding that non-native speakers display differences unique to their judgements: more likely, their grammars simply differ from natives on the points investigated, and this would show up in everyday use as well if these constructions occurred.

One would expect certain types of nonlinguistic experience to influence judgements as well, e.g. factual world knowledge, cultural and social experiences and beliefs. A fascinating example of how world knowledge is relevant to grammaticality is provided by Belletti (1988): according to her, the following two sentences involving subject postposing contrast in grammaticality in Italian:

(9) a. È stato rubato il portafoglio a Maria.
'has been stolen the wallet to Maria'

[29] Chaudron (1983) points out that there were only eight non-native subjects in this experiment all together, so due caution is advised in interpreting the results.

b. *È stata rubata la pianta a Maria.
('has been stolen the plant to Maria')

The crucial difference here is that we can assume people normally own only one wallet, but the same is not true for a plant. Presumably, someone from a different culture would not show this distinction. G. Lakoff (1971) has argued that the well-formedness of a sentence can *never* be assessed without reference to a set of presuppositions about the nature of the world, and cites numerous sentences where people differ in this regard, e.g. whether *My cat enjoys tormenting me* is grammatical depends on whether one believes cats to have minds; in cultures where events are believed to have this property, the equivalent of *My birth enjoys tormenting me* is perfectly normal.

5. Conclusion

The studies reviewed in this chapter show that a considerable proportion of individual differences can actually be attributed to specific linguistically-relevant features of the person, be they in-born or the result of experience. Nonetheless, we can be fairly certain that there remains much variation that we cannot 'factor out' in this way. In this regard grammaticality judgements are like most other forms of behaviour, including other metalinguistic tasks such as ambiguity detection (Kess & Hoppe 1983): a common genetic endowment provides for a certain degree of commonality, and certain gross parameters of variation, but beyond that differences abound. This state of affairs, however immutable, presents frustrating problems once we acknowledge that the study of grammar, while in principle a study of each individual's mental structures, must appeal to the judgements of many individuals. This does not mean the judgements are not valid; however, before we resign ourselves completely, we should consider that not all the variation that shows up within and across experiments is attributable to real differences between subjects: subtle differences in procedures or in the sentences themselves can add error to the actual variation. In the next chapter we turn our attention to such confounding sources.

Chapter 4

Stimulus and Procedural Factors in Grammaticality Judgements

Meander (a linguist):	*I have a theory that everybody's eyes are colourless.*
Simplon (a psychologist):	*But, Meander, everybody's eyes look brown, blue or green to me.*
Meander:	*That's because they are actually wearing contact lenses to color their eyes.*
Simplon:	*But, Meander, I know that I don't wear contact lenses, and when I look in the mirror my eyes look blue to me.*
Meander:	*Ah: but then, there's a lot we don't know about mirrors.*

(Bever 1974)

1. Introduction

By now it should not be a surprise to find that grammaticality judgements may vary depending upon the procedure by which they are obtained and properties of the stimulus items that are presented. In fact, the latter assertion may seem tautologous: obviously if judgements are to be of any value they must vary depending on the sentences being judged. Our concern here, therefore, will be on variation caused by factors that are *irrelevant* to the concept that we are trying to access through grammaticality judgements, namely grammaticality. Clearly there is room for disagreement here, since what should "count" towards grammaticality is a matter of theoretical assumption or fiat. Similarly, whether an experimental procedure "interferes" with grammaticality judgements depends on one's view of how best to obtain them. For the most part, the variables we examine in this chapter would be uncontroversially labelled as confounds by the majority of linguists. Even where there is disagreement, for instance as to whether context is a nuisance or an integral part of the grammaticality of a sentence, systematic study should lead to a better understanding of the phenomenon, and thus improve the linguist's chances of de-

signing effective elicitation procedures. The recommendations we make on the basis of the research reviewed here reflect one particular point of view on the nature of grammaticality.

The remainder of this chapter is essentially a top-down survey of the experimental literature. Section 2 covers features of the elicitation process as a whole, beginning with what subjects are asked to do, that is, how the procedure of judging grammaticality is explained to them (§2.1).[1] This issue will end up pervading the entire chapter, since differences in instructions are largely to blame for the staggering discrepancies between experiments, and much of the existing literature is undermined because vague instructions make many of the results virtually uninterpretable.[2] Then we examine the effects of the order in which sentences are presented for judgement (§2.2). The next three subsections largely follow the research program of one experimenter, Hiroshi Nagata, whose initial work looked at the effects of repeated exposure to the same sentences (§2.3). This also raises the issue of intra-subject consistency, which was a subsidiary concern of several other experimenters mentioned in this chapter. Later work by Nagata and others brought in mental state manipulations and their interaction with repetition (§2.4), and then sought support for his hypothesized explanations by correlation with subjects who were explicitly told to use certain judgement strategies (§2.5). The following subsection explores the least-studied procedural variable, which nonetheless could arguably have the greatest impact on judgements, namely the presentation modality—spoken versus written; closely tied up with this is the matter of register, since together these two factors define to a large degree the nature of the discourse situation, and hence how grammatically strict or permissive we are liable to be as listeners (§2.6). Finally, we take a very brief look at speed of judgement (§2.7).

Section 3 moves in for a more close-up view of the properties of stimulus sentences themselves. We begin with the role of the context in which the sentence is situated; since we are following Chomsky's narrow definition of acceptability, we restrict the

---

1 We do not concern ourselves here with the much larger question of the range of tasks one might use to elicit information about acceptability. This was discussed to some degree in Chapter 2, §2; further exploration is beyond our scope, since our focus is judgements.

2 One feature of the task instructions not covered explicitly in this chapter is the type of judgement required: good/bad versus numeric rating versus relative ranking. This issue was discussed in Chapter 2, §3. We are also omitting mention of certain standard confounding effects that psychologists typically seek to avoid but that do not seem to have any special impact in the domain of language judgements; some of these will be mentioned in the proposed methodology in Chapter 5, §3.

term "context" to the purely linguistic context, ignoring social factors that obviously influence acceptability in the broader sense (see van Dijk 1977 for a discussion of the latter). Nevertheless, we will see that the term is used in at least three different ways (§3.1). Our next concern is the extent to which the meaning of a sentence or apparent lack of it affects people's judgements of grammaticality, in cases where (we assume) it is an orthogonal issue (§3.2). We then ask the same questions about (perceived) frequency of occurrence of sentence types (§3.3). The final two subsections concern the level of individual words: what happens when one is replaced by another that is grammatically equivalent (§3.4), or with one that is grammatically identical (§3.5). We acknowledge here that some potential stimulus variables do not have separate headings devoted to them. Intonation is mentioned briefly in conjunction with modality in §2.6. Its written counterpart, punctuation, does not appear to have been studied for its effects on grammaticality judgements in general (but see Levelt 1974 for some discussion of its possible effects), although it is occasionally mentioned anecdotally in other types of psycholinguistic studies. Other features of printed text, such as type style, have not been studied in this regard; capitalization was mentioned by one of Hill's subjects (§2.1). Finally, Section 4 summarizes the major implications of the reviewed research.

2. Procedural Factors

2.1 Instructions

Hill (1961) performed some of the earliest investigations into the nature of grammaticality judgements; he used 10 subjects, of which 3 were linguists and several others were English professors, which should immediately lead us to suspect that his results will not generalize to the population at large. They were instructed to "reject any sentences which were ungrammatical, and to accept those which were grammatical," but there was apparently no definition or explanation of these terms given, nor any examples of their application. The results and anecdotal comments he reports show quite clearly that subjects had no clear notion of the concept; for instance, while all 10 subjects rejected *Those man left early*, 6 of them accepted *The child seems sleeping*. Even more troubling is the fact that two rejecters of the sentence *I never heard a green horse smoke a dozen oranges* changed their judgements to accept it once it was pointed out to them that the sentence was true. Other subjects explained their acceptance of a sentence by saying "it sounds like poetry" or rejected a sentence because it did not start with a capital letter. The conclusion to be drawn from all this should be obvious: even subjects who are supposedly

experts on language cannot be expected to know what linguists mean by "grammatical" (or "acceptable," for that matter),[3] so if you do not explain to them what you want, each subject takes his or her own interpretation and the results are meaningless. This criticism was made in the same journal volume by Chomsky (1961). We will see in §3.2 that Maclay and Sleator (1960) encountered the same problem with linguistically naïve subjects. As Chaudron (1983) puts it, "grammaticality, acceptability, and meaningfulness . . . are not socially uniform concepts." Similarly, in a very widely cited passage, Carden (1970a) states, "You must define 'grammatical' or 'acceptable,' words that naïve informants use in widely varying ways. It is of no value to know that 13 informants consider a sentence acceptable unless you know that they mean the same thing by 'acceptable.'" Newmeyer (1983, pp. 63–64) and Botha (1973) make similar points. Birdsong (1989) suggests that the problem is particularly acute when the forms in question occur in speech but are proscribed in writing.

Unfortunately, as we have seen in previous chapters and will continue to see in this chapter, many studies have fallen into exactly the same trap. In fact, if we were to ignore all studies where we believe the instructions to subjects were inadequate to convey the subtlety of a linguistic definition, the remainder could likely be counted on one hand. Thus, we will continue our practice of describing the experimenter's instructions in considerable detail, in order that the usefulness of the results can be assessed, but in order to make any progress, we will have to assume the major findings would hold up under more careful procedures. (Therefore, some of Hill's other results will be reported in subsequent sections.) This is not meant to condone the existing practice or deny the need for replication, but merely to accept the fact that somewhat confounded data are better than none at all. By way of ending on a positive note, we will also report occasionally instances of very well designed instructions, and proposals for how to test their effectiveness; for instance, Chaudron (1983) suggests asking subjects what they considered valid judgement criteria, and how they made use of these criteria in particular sentences. See Greenbaum & Quirk 1970 for examination of the instructions surrounding performance tasks.

---

3 Actually, it is not even clear that linguists agree among themselves as to what exactly is supposed to count towards grammaticality; as mentioned in Chapter 1, the concept changes as the theory evolves. At the very least, researchers must clarify this point in their own minds before trying to design a set of instructions.

2.2 Order of Presentation

Greenbaum (1973, 1976a) describes an experiment that looked at the effects of order of presentation of sentences on judgements. It required nonlinguist subjects to rate sentences containing participial *while* phrases attached in various places in a sentence, e.g.

(1) Sophia Loren was seen by the people while enjoying herself.

(2) The people saw Sophia Loren while enjoying themselves.

(3) Judy was seen by the people while enjoying themselves.

(4) The people saw Karen while enjoying herself.

Subjects had a choice of four responses to each sentence: "acceptable," "uncertain, but probably acceptable," "uncertain, but probably unacceptable," and "unacceptable." Two subjects were assigned to each possible ordering of the four sentences, and statistical analysis showed that the first sentence for each group was rated significantly lower than the others. This is presumably a type of primacy effect, although it is not clear how the direction of change (worsening) would have been predicted based on order effects in nonlinguistic domains. No significant effect was associated with any other position in the list. Clearly, then, sentence order should be controlled for, either by randomization or counter-balancing. The study was essentially a replication of one by Elliot, Legum, and Thompson (1969), who apparently used the same order for all subjects, thus severely confounding their results. Problems remain with Greenbaum's procedure as well, however. For one thing, he apparently gave his subjects no explanation of the term "acceptable." For another, he ignored the standard psychological practice of using warm-up trials to get subjects comfortable with the procedure. If he had done so, the effect of first position could have been removed rather than just counter-balanced, thus reducing the amount of variability in the scores. See Labov 1975, p. 21 for a review of these two studies and ensuing work. Greenbaum and Quirk (1970) also reported order effects in their test batteries, both for judgement and performance tasks. In one case they found a significant difference between the two orders (between-subjects groups) in the number of subjects giving "grammatical" ratings for 5 out of 51 sentences tested. Interestingly, relative rankings showed almost no changes as a result of varied orders.

Certain effects of presentation order that arise due to relationships among stimulus sentences will be treated as context effects in §3.1.

2.3 Repetition

Nagata has performed a number of experiments investigating the effect of repeated exposure to sentences on judgements of their grammaticality, and the interaction of repetition with other manipulations. According to him, no previous experiments have examined this variable systematically (I have not come across any either), but it has important implications, because linguists, the most common producers of judgement data, often consider the grammaticality of sentences many times over the course of investigating some theoretical issue. (Spencer (1973) also speculates on such effects of repeated exposure on linguists.) Thus, if we had reason to suspect that their judgements would not be stable, by the time they were drawing their conclusions the judgements might be quite different from first impressions. (This issue will also recur in §2.7.) Nagata suggests two possible a priori outcomes of a repetition treatment, to which we will add a third. He proposes that judgements might become more lenient (more grammatical) because subjects would construct additional linguistic or situational contexts for sentences, eventually finding cases where even fairly bad sentences would be reasonably acceptable. This would accord with the general psychological phenomenon of habituation, whereby repeated exposure to the same stimulus has diminishing effect, e.g. the same painful stimulus will evoke less and less reaction. On the other hand, Nagata postulates, we might expect judgements to become more *stringent* because people "differentiate" more syntactic or semantic properties of the sentence, that is, the more they look at it, the more things they may find wrong with it. Graeme Hirst (personal communication) has suggested a third possible outcome, namely that repetition might increase subjects' confidence in their judgements. In that case, we would expect a polarization of judgements, i.e. good sentences would get better and bad sentences worse.

In his first study, Nagata (1988) performed three experiments to examine the basic effect of repetition and its interaction with the presence of context. The procedure was essentially the same for many subsequent experiments as well. His stimulus materials were pairs of grammatical and ungrammatical sentences drawn from the Japanese linguistics literature, matched as closely as possible, plus pairs of filler sentences. Whether the target sentences were labelled good or bad was determined by whether or not they received any question marks or stars in the original source articles. Thus, the number of good and bad sentences would be roughly equal; the total number of sentences was 48. Subjects were asked to rate the extent to which the sentences were grammatical, i.e. "correctly expressed in Japanese," on a scale from 1 to 20. They were told that correct

sentences should be rated as 1, while 2–20 indicated increasing degrees of badness. Subjects were also told to make use of the full scale.[4] First, the sentences were presented one at a time in random order on a CRT and the subjects were asked to give their numeric judgements of each, to be used as the baseline measure. In the second part of the procedure, each sentence in turn was presented in a repetition phase, followed immediately by another judgement. In the repetition phase, the sentence was displayed nine times in a row for 3 seconds each, with 1 second between presentations. During these repetitions, the subject was told to think of the grammaticality of the sentence. Then, upon the tenth presentation the subject was asked once again to rate the sentence. (The order of sentences in this part of the procedure was again random.) The first, unsurprising, result was that the supposedly good sentences received significantly better ratings than the bad ones both before and after repetition, confirming that the a priori division was reasonable. As for the effects of repetition itself, the grammaticality of both kinds of sentences decreased significantly after repetition (i.e., the rating numbers were higher), as compared to before. Nagata concludes that subjects were engaged in differentiation rather than enrichment during the repetition phase. If this result is general, we must re-examine why the theory of "mere exposure" has been widely accepted in accounting for language change. It holds that as people hear a form more and more, they like it more, deem it more acceptable, etc.: "familiarity breeds content."[5] To the extent that this is true in language change, why is it not true in Nagata's repetition paradigm—is the time span involved too short, i.e. is repetition in quick succession different from repetition over a long period of time? Is the problem that all the repetitions come from the same source?

In the first follow-up experiment, the same sentences were used but the final judgements were made with the sentence preceded by a context string.[6] As compared to

---

4 The only justification given by Nagata for the unusually large rating scale and its asymmetric division (as opposed to making the best sentences 1, the worst 20, and the remainder on a continuum in between) is that the same scheme was used by Moore (1972). But Moore himself gives no justification for these choices. We can only speculate that they may have been inspired by the psychometric results mentioned in Chapter 2, §3. We suggest that Nagata's results might profitably be replicated using a smaller, symmetrical rating scale, but it does not seem to us that his scale would have biased the results. If anything, the large scale should increase variability in the results and make it harder to find significant effects.

5 Attributed in Bradac et al. 1980 to Walker (1973). Of course, in everyday situations, repeated exposure to a form is not accompanied by an instruction to ponder its grammaticality.

6 Nagata provides translations of his target and context strings only for the grammatical sentences, of which we give two examples; the targets are italicized:

(i) Look out of the window. *It is raining.*
(ii) What's the matter with you? *If you don't eat, you'll be hungry.*

post-repetition ratings in the first experiment, the with-context condition showed that ungrammatical sentences were judged significantly more grammatical; they showed no significant change from the pre-repetition ratings. The grammatical sentences did not differ significantly from either the post-repetition ratings in Experiment 1 or the pre-repetition ratings in Experiment 2. Nagata believes that this points to a change in encoding or organization of the bad sentences when embedded in context, somehow undoing the change induced by repetition. Apparently context had some mitigating effect for the good sentences as well, since they failed to show the decrease in grammaticality found in the first experiment. Nagata suggests that the effect of context was somewhat masked by a ceiling effect, i.e. the sentences were already rated about as high as they could get. A second follow-up, in which context preceded the target sentences before, during, and after repetition, confirmed the basic finding that context blocks the repetition effect, supposedly because it provides a stabilizing base for judgements.[7] The pre-repetition ratings were also compared with those of the first experiment, allowing a direct analysis of the effect of context alone. No significant differences were found, apparently contradicting numerous other studies that found that context raises grammaticality. See §3.1 for further discussion of this point.

In two subsequent studies (Nagata 1987a, 1987b), two alternative accounts for the basic repetition finding were ruled out. First, one must consider the possibility that the subjects' use of the rating scale had changed, independent of repetition, because the first set of ratings were made before all the sentences had been seen. Since subjects were told to use the full range of 20 values, and since they would only know which were the best and worst sentences after the first round of ratings, this is a distinct possibility. Thus, a new experiment was designed to seek out such a trend: sentences were all judged once, then all judged a second time (in a different random order). Since no changes were found between first and second ratings, a "change in the modulus of judgmental scale" account,

---

Apparently the nature of the bad sentences was such that reasonable contexts could still be provided for them; since I speak no Japanese, I cannot assess this.

7 Spencer (1973) cites several relevant background studies on repetition effects in word recognition, among them one by Taylor and Henning (1963) that reportedly shows a similar type of stabilizing effect: if subjects are told that they will only hear actual words, they do not report that some of the repetitions sound like nonsense syllables, whereas subjects who expect nonsense forms claim to hear them.

as Nagata calls it, is ruled out.[8] This result appears to contradict Carden's (1976) survey of a number of studies that examined the internal consistency of their data by seeking a second rating from subjects some time after the initial data collection: many of these found it highly inconsistent. However, Nagata was comparing *mean* ratings of all the good sentences pooled and all the bad sentences pooled, not individual sentences—a change could have been washed out by inter-sentence variability. A second potential confound comes from satiation: prolonged repetition of symbols (e.g. words) has been shown to lead to temporary loss of their meanings and concomitant illusory changes in their perception; if Nagata's subjects reached satiation for the stimulus materials, the results do not necessarily bear on "normal" judgements. Since satiation is a short-term phenomenon, this possibility was tested by looking for long-term maintenance of the changes induced by repetition. The subjects from the original experiment were re-tested on the same sentences 4 months later; their results were not significantly different from the original post-repetition judgements, and in most cases were still significantly higher than the original *pre*-repetition judgements, i.e. whatever had changed in their approach to these sentences was still true long after any satiation effect would have worn off. But had they encoded something *specific* to these 48 sentences, something that was maintained in their minds for 4 months without reinforcement, or was it that their judgement process *in general* had changed as a result of greater experience with the task? Nagata does not consider that latter possibility, yet it strikes us as somewhat more plausible, and could be easily tested. For instance, 4 months after the repetition treatment we could give the same subjects novel sentences, and compare their ratings to those of subjects who had never undergone repetition. If our interpretation is correct, we expect the former group to show significantly more stringent ratings.

A fourth study (Nagata 1989d) was designed to assess the extent to which the repetition effect applies generally to sentence types other than those used earlier. Its first experiment factored out the differential effects of repetition on sentences marked with a question mark as opposed to a star in the original sources. Nagata's hypothesis was that the truly bad sentences could not get any worse through repetition, but in fact both groups of sentences were rated worse in post-repetition judgements. The second experiment

8 Nagata does not discuss the possibility that subjects could have remembered their initial ratings and striven to be consistent by duplicating them the second time around. Since much less time intervened between first and second ratings as compared to conditions in the repetition experiment, the possibility is worth considering. However, given that there were 48 sentences and 20 possible scores for each, and the two presentations were in different orders, we doubt that accurate memory for one's ratings would be possible.

used new stimulus materials altogether, instances of the three types of violations identified by Chomsky (see Chapter 2, §3): incorrect lexical category versus subcategorization violation versus selectional restriction violation. Here he found that repetition had no significant effect on any of the three types of badness; the latter two in fact did not show significant differences between them. His explanation is that these violations were all more blatant than those used in the earlier studies, which involved subtle uses of particles, reflexives and honorifics. A more blatant violation may be easier for subjects to detect and explain, thus tending to anchor more on initial judgements and resisting change. Thus, one must conclude that, at least for ungrammatical sentences, the repetition effect has limited external validity.

The only other study I am aware of that has involved repeated judgements of the same stimuli was one by Carroll (1979). The issue for Carroll was the extent to which complex compound nouns such as *girl that irons her clothes doll* (referring to a doll that looks like a girl and that irons her clothes) were judged acceptable in a sentential context as a function of the syntactic structure of the elements making up the compound. He was cognizant of the potential for a change in use of the (5-point) rating scale on the basis of the range of stimuli he was presenting, especially since subjects might never have seen such complex compounds before, and so he asked his subjects to make a second pass through the data, judging them again. While the mean ratings of several sentences did increase from the first to the second judgement, Carroll does not analyze the differences for statistical significance, so we cannot compare his results to those of Nagata. However, the statistical tests that *were* performed showed that there were fewer significant differences *among* the 10 sentence types in the second set of judgements than in the first. Apparently, subjects see the range of sentences as more homogeneous the second time around. This study can also be held up as a rare example of one that took care to ensure that subjects had a strong understanding of the basis on which they were to make their judgements; Carroll gave example sentences with their ratings, discussed why the ratings had been chosen, encouraged questions about the rating system, etc. (see Carroll 1979, pp. 874–875).

2.4 Mental State

The next step in Nagata's project was to investigate the interaction of repetition with mental state, specifically the effect of objective versus subjective self-awareness. Before describing his study, we digress briefly to explore the nature and history of this

manipulation and its application to language. There is a standard operational technique from social psychology that is used to manipulate the 'introspective set' of subjects, inspired by the 'social facilitation' effect: when you observe yourself or others engaged in the same activity it makes you do it more intensely, e.g. people will ride a bike faster if they see other people riding bikes. Duval and Wicklund (1972) brought together a large number of findings in this area and unified them under the theoretical distinction of subjective versus objective self-awareness, states of consciousness directed at the external environment or at oneself, respectively. By their definitions, subjective self-awareness (SSA) is a state of consciousness in which attention is focused on events external to the individual's consciousness, personal history or body: you are the *subject* of consciousness directed outward, the source of perception and action but are not aware of yourself as experiencer. Objective self-awareness (OSA) is exactly the opposite state: your consciousness is focused on yourself, your own conscious state, personal history, body, etc.; you are the *object* of your own consciousness, a state that often leads to self-evaluation and negative affect, by inducing self-comparison with external standards. SSA is humans' primary or default state: the environment normally draws your attention; OSA requires a reminder of your status as object in the world—stimuli that cause a shift of attention to yourself, such as looking in a mirror, hearing your voice on tape, seeing a photograph of yourself, having a TV camera pointed at you, etc. Once you are in the OSA state, attention shifts to your relevant features, regardless of which sort of stimulus induced the state. One experiment that Duval and Wicklund used to demonstrate this manipulation went as follows. The experimenter describes to the subject a hypothetical scenario involving him or her, such as a traffic accident. For each situation, subjects are asked to rate how responsible they were for the outcome, i.e. how much of the causality was attributable to them. There are two conditions: the experimental room may have a mirror in it, positioned so the subjects will see themselves in it (the OSA condition), or it may have the non-reflective back of the mirror facing them (the SSA condition). The result was that OSA subjects attribute significantly more causality to themselves than the SSA subjects. In another experiment, subjects are given an intelligence test, then told they scored below average on it, left alone in a room with a clock, and instructed that if no one returns, they may leave after a certain number of minutes have passed. Again, self-awareness was manipulated by the presence or absence of a mirror. The OSA subjects tended to leave the room significantly sooner than the SSA subjects, supposedly because the mirror was leading to negative feelings: subjects were constantly reminded of their "below-average intelligence." Note that in this case, unlike the previous experiment, there was no reporting involved; the manipulation affected the subjects' actions, not just their statements.

Carroll, Bever, and Pollack (1981) conducted the first study that I am aware of that applied self-awareness manipulations to linguistic intuitions. Their premise was quite similar to ours: since intuitions are produced by performance mechanisms, they wanted to control and study these mechanisms, specifically by manipulating mental state. They had a quasi-theoretical goal as well, which was to show that different mental states could produce intuitions that corresponded to two competing theories of the relation between syntax and semantics, namely "Autonomous Syntax" and "Abstract Syntax"; therefore, it was not a case of choosing one theory over another, but rather both theories were correct, they simply accounted for different kinds of intuitions. Since this theoretical issue is not of concern to us here, and in fact is no longer much discussed, we will ignore the potential theoretical implications of the results and merely concern ourselves with the effects produced on judgements themselves. Carroll et al. suggested that linguists rendering intuitions need to be in something like the OSA state: unlike a regular speaker communicating, who is subjectively preoccupied and will tend to produce speech errors and ambiguities without noticing them (as Duval and Wicklund themselves suggest), linguists must cease being speaker/hearers to pause and reflect on the linguistic signal, to "objectify the sentence from all the specific potential functional contexts of its utterance." In a pilot study, they had subjects rate the truth of categorial statements like those listed below on a scale of 1 (least true) to 10 (most true); in one condition, subjects had to use an answer key that was stuck to a mirror to fill out the questionnaire, but were not otherwise instructed to look at it; the other condition had no mirror. The mean ratings for the two conditions are shown in Table 3 (from Carroll et al. 1981, p. 372). Overall, subjects in the OSA condition gave higher truth ratings. Furthermore, the greatest difference between the groups occurred on sentences like the one highlighted in the table: technically true but pragmatically unlikely ones. Carroll et al. propose the explanation that OSA subjects consider more potential communicative situations than SSA subjects, and this is most important for the marginal cases; the false sentences are pragmatically appropriate in very few situations, and the paradigmatic ones require no contextualization to establish their truth. The more general conclusion is that self-awareness manipulation does make itself felt in linguistic tasks.

Table 3

Mean Truth Ratings of Sentences as a Function of Self-Awareness

Sentence Type	SSA Condition	OSA Condition
A house is a building.	9.55	9.60
A garage is a building.	7.60	8.60
A lean-to is a building.	**5.40**	**7.35**
A tent is a building.	3.25	4.50
OVER-ALL	6.45	7.51

Thus, Carroll et al. proceeded to their central investigation. Since it does not concern grammaticality judgements, our summary will be rather brief, focusing on those aspects that will be relevant to Nagata's study. The task was to rate the pair-wise similarity of sentences from the following sort of paradigm:[9]

Active: The morning sun dried the sweet raisins.
Passive: The sweet raisins were dried by the morning sun.
Inchoate: The sweet raisins dried in the morning sun.
Were-Inchoate: The sweet raisins were dried in the morning sun.
Cause: The morning sun caused the sweet raisins to dry.
Because: The sweet raisins dried because of the morning sun.

The relevant feature of this group of sentences is that they are semantically very close but syntactically quite different: hence, the prediction that OSA subjects who are more aware of social interaction will be more sensitive to communicative similarity, since they consider a wider range of potential situations for the utterances, and thus differ from SSA subjects who will focus on the surface form of sentences. This prediction was borne out: OSA subjects gave higher similarity ratings overall; in addition, the multidimensional scaling plots come out quite different for the two groups. Carroll et al. take their results to show that people may use different *strategies for interpreting intuitions*, depending on the situation. Extrapolating from their statement, one could imagine having intuitions about *both* the structural and communicative properties of sentences, but how these are

9 Subjects were not told what to use as a basis for measuring similarity, so once again we have the potential for widely-varying interpretations.

weighted in coming to an overall similarity measure would depend on whether the situation prompted communicative versus sentential assessment.[10]

We now return to the realm of grammaticality judgements, and specifically to the experiments reported in Nagata (1989a), which investigated the effects of self-awareness and its interaction with the repetition manipulation previously described. Nagata started from the assumption that in his earlier repetition studies, subjects were in an SSA state (since there was nothing to trigger OSA), and hence used sentential strategies like Carroll et al.'s SSA subjects; perhaps repetition would have a different effect on OSA subjects: over multiple repetitions they might consider more potential situations or contexts for the sentences and thus rate them more grammatical (recall this idea from Nagata's first experiment). The first test of this hypothesis involved the same repetition paradigm as before, except that in the OSA condition there were mirrors on either side of the CRT where sentences appeared, and the subjects were told to look at themselves in the mirror while making judgements and while thinking of the grammaticality of sentences during repetition. The SSA subjects showed a worsening of ratings, as in previous studies, but OSA subjects showed no change of ratings after repetition. Thus, it seems that the mirror manipulation did negate the effects of repetition, although it failed to induce greater leniency in judgements. Nagata was convinced that such a leniency effect should be demonstrable, and suggested that it was undermined by possible ceiling effects, unclear instructions and over-exposure to the mirror (that is, it may have begun to induce the communicative strategy on initial judgements, leaving less room for measurable change after repetition). A follow-up experiment tried to solve these problems by using only intermediately-rated sentences (to avoid the ceiling, but in the process limiting the generality of the result), omitting mirrors from the initial judgement phase, and explicitly telling subjects to simultaneously look at themselves in the mirror and think about the sentences' grammaticality. We note that this is a much more explicit and forceful use of the mirror manipulation than the one that Duval and Wicklund or Carroll et al. had used; furthermore, no significant effect of self-awareness had been found for the before-repetition judgements in the first experiment. Apparently this particular procedure is not as susceptible to self-awareness manipulations as those others; this may be because judging grammaticality is more inherently a structural task, as compared to judging sentence similarity or truth. Despite all

[10] If all that is involved is a communicative orientation, as opposed to seeing *oneself* objectified, one might expect other procedures to have the same effect, e.g. showing someone a photograph of another person instead of the mirror. To my knowledge, such an experiment has not been done.

this emphasis, the OSA judgements only came out marginally more lenient after repetition. A second follow-up experiment was done to rule out a potential confounding variable: it is possible that the division of subjects' attention between mirror-watching and sentence-pondering could have created ratings different from those of the SSA subjects, independent of the fact that the competing activity was related to self-awareness. Thus, Nagata gave subjects a simple arithmetic problem to solve as a distractor during the repetition phase, instead of mirror-gazing, to see whether division of processing resources could account for the previous finding. In this condition, there was no change in judgements after repetition as compared to before, so division of attention *can* nullify the repetition effect. However, there was no trend towards *increasing* ratings, so to the extent that such an effect is reliable, it cannot be explained by processing factors alone: self-awareness must be considered.[11]

2.5 Judgement Strategy

Nagata concludes from the preceding three experiments that, in judging grammaticality, SSA subjects focus on syntactic and semantic structure, while OSA subjects look at pragmatic use; if so, then Carroll et al.'s suggestion that linguists need to be in the OSA state seems misguided, at least for the purposes of judging grammaticality. But note that so far we have only circumstantial evidence concerning the actual strategies used anyway. Nagata (1989c) wanted to explore this by explicitly telling subjects what sort of strategy to use in their judgements, rather than inducing it indirectly with mirrors and such. If the two explicit strategies show the same respective effects as the mirror versus no-mirror conditions, we have suggestive (though not conclusive) evidence that the interpretation of those effects is on the right track. This experiment again used sentences of intermediate grammaticality, where the leniency effect of OSA had shown up. One group of subjects was told to "analyze each sentential structure independently of sentential and/or situational contexts," and consider the parts of speech involved, during the repetition phase; the other group had to "[supply] sentential and/or situational contexts to each sentence such that each sentence could be used in such contexts."[12] The

---

11 This is not an airtight argument: perhaps the arithmetic task, which involved subtracting 2 from an integer, was not as demanding as mirror-gazing.

12 It is not evident from Nagata's description whether these are exact translations of the instructions, or whether subjects were given more explanation. Even knowing the purpose of the experiment, I do not find this wording particularly clear.

standard repetition paradigm was used, except that after the repetition phase subjects had to describe what they had been thinking, so the experimenter could be sure the desired strategy had been followed. Those who did not had their results discarded. While the differentiation condition produced significantly more ungrammatical ratings after repetition, the enrichment group showed only a nonsignificant tendency towards leniency. Again, Nagata tries to explain why the expected trend did not reach significance: apparently the enrichment strategy is hard to use, and even the subjects who seemed to describe the appropriate thoughts may not have used it as intended. But it is hard to see why subjects should use less of a strategy when explicitly told how to follow it than when it is induced indirectly by the mirror manipulation. This question casts some doubt on Nagata's interpretation of the OSA leniency effect. Perhaps the possibility that OSA affects reporting more than linguistic analysis is worth investigating further after all, if a more convincing demonstration of the change in judgement strategy cannot be made. As Bever's epigraph states, there certainly is a lot we don't know about mirrors.

Nonetheless, some more general conclusions can be unequivocally drawn from Nagata's studies. First, it is clear that the details of the process of intuitive judgement cannot be ignored when using intuitions for theoretical purposes; on that point, we agree with Carroll et al. as well as with Nagata.[13] More specifically, we can conclude that it is easy to make sentences get worse in people's judgements, but hard to make them get better. Given the stringency effect of repetition, we should expect linguists' judgements to be more stringent than non-linguists', at least on sentences that they have studied in detail. I am not aware of any studies having been done specifically on sentences that linguists have worked on extensively; more general studies have differed as to whether linguists are more or less lenient than normals (see Chapter 3, §4.1). I would still maintain that the influence of repetition is another valid reason why linguists' judgements should not be used as crucial theoretical evidence. With regard to where the effects of self-awareness come from, they seem to transcend language and thus fit the general description of a cognitive manipulation whose effects carry over into linguistic judgements. The repetition effect is more problematic in this regard, however. It runs contrary to basic habituation effects; in fact, I have not been able to find any parallel manipulations in other cognitive domains. If Nagata's suggestion is right, then the effects stem essentially from

[13] Nagata proceeds to argue that *any* variability or manipulability in judgements contradicts Chomsky's claim that native speakers know which sentences are grammatical. He is under the mistaken impression that Chomsky claims that we must manifest this knowledge in a consistent ability to judge grammaticality.

discerning more fine-grained properties of the stimulus through repeated consideration. If the effects are limited to the particular sentences used rather than overall ratings, then this is *not* a case of developing expertise, i.e. increasing the ability to discriminate. Rather, a parallel effect would have to involve complex stimuli whose properties are not all apparent on first exposure, e.g. a complex geometric figure containing multiple sub-figures that must be picked out. Finally, it is evident that psychological effects can interact in unpredictable ways, so that a complete understanding cannot be achieved merely by identifying each effect in isolation.

2.6 Modality of Presentation and Register

Vetter, Volovecky, and Howell (1979) were interested in the potential effects of modality of presentation and intonation, although their main interest was with meaningfulness—see §3.2. They used five conditions for sentence presentation: visual presentation only, auditory presentation only, or both presentations simultaneously, and the latter two modes could involve normal or monotone intonation. Interestingly, they found no overall effect of mode of presentation, although 16 of 60 particular conditions did show significant differences, which suggests to me that this variable is worth investigating more closely. But the basic result that intonation is not a factor echoes similar results in the domain of spoken surface-structure ambiguity resolution. Studies by Berkovits (1981, 1982) have shown that intonation plays a very limited role in disambiguating such sentences, being easily overridden by the inherent bias of the sentence or the surrounding context unless a subject's attention is explicitly drawn to prosodic cues. On the other hand, Hill (1961) describes some cases where reading a sentence with normal intonation, as opposed to presenting it in written form, increased the number of acceptances. For reasons discussed in §2.1, his results should be taken very cautiously, however.

The absence of any modality effect is at odds with the widely-held intuition that our judgement criteria are much stricter for written materials than for speech. (In line with this intuition, Bialystok and Ryan (1985) argue that oral presentation stresses meaning, whereas written presentation more naturally elicits attention to structure.) See Biber 1986 (cited in Birdsong 1989) for an analysis of some of the actual differences between spoken and written language. If modality turns out to be relevant to judgements, we should consider whether it can account for some of the differences between literates and illiterates discussed in Chapter 3, §4.2. However, the issue of register is tied up in this as well: a formal speech would have to meet higher standards of grammaticality than

a casual conversation. Greenbaum (1977b) suggests that written questionnaires present a fairly formal context for subjects, which may show up in sentence ratings as a preference for the more formal of two alternatives if they are compared side-by-side. Yet another confounding factor could be the degree of preparation: prepared text, whether spoken or written, can be freed of errors, whereas in spontaneous production speakers (or writers in certain circumstances) must be allowed some leeway in extricating themselves from grammatical culs-de-sac. One might approach the unravelling of these factors by using linked computer terminals that allow written communication with various speeds of transmission: instantaneous letter-by-letter, line-by-line, or complete messages (Graeme Hirst, personal communication). Whichever factor or factors determine tolerance, we are then left with explaining how the various levels of grammaticality criteria are encoded in the mind: different grammars, different parsing rules, a reduced threshold on the same parsing rules, etc. We will return to these issues in Chapter 5.

There are additional features of language that are related to register to some extent and that Ross (1979) speculates may have systematic effects on grammaticality judgements. These include clarity, awkwardness, slanginess and floweriness. While these have likely been examined in a sociolinguistic context, I am not aware of any research looking for them as confounds where grammaticality was the property subjects were targeting.

2.7 Speed of Judgement

Studies have differed as to the amount of time subjects are given to make their judgements: in most cases, written questionnaires are self-paced, although they may also be "speeded," i.e. subjects are told to work quickly; experiments using computer control (usually also measuring RT) may limit the amount of time a sentence is visible, and also limit the time available for judgement before the next sentence appears. This raises the issue of whether we want a subject's initial reactions to a sentence, or a carefully-reasoned decision resulting from some amount of deliberation. Presumably these two kinds of judgements would differ, although the matter has not been studied directly. (A study by Warner and Glass (1987), to be detailed in §3.1, found that context effects were attenuated by the delay in self-paced as opposed to on-line judgements. Also, we suggested in §2.3 that prolonged consideration of a sentence might induce effects similar to Nagata's repetition treatment.) Obviously if our goal is to examine the on-line processing of grammaticality, its effects on parsing, etc., then first reactions will be most useful; but if it

is the status of sentences that concerns us, it is not clear which should be preferred. One advantage to first impressions is that there is little time for the subject to consider (potentially irrelevant) extra-sentential factors. In cases where initial reactions are desired, we need a methodology for getting them: the costs of computer-controlled experimentation may be prohibitive as compared to questionnaires, so some authors have tried to use the latter. For instance, the instructions in Heringer's (1970) study (discussed in Chapter 3, §2) told subjects to refrain from changing their response after the initial judgement or rereading sentences that had already been judged; Greenbaum (1977b) told subjects not to turn back to previous pages in the questionnaire booklet, and to work as quickly as they could without being careless; Greenbaum (1973) told subjects every 5 seconds that they should turn the page in their questionnaire booklet to move to the next sentence.

3. Stimulus Factors

3.1 Context

We agree wholeheartedly with Bever (1970a) on the issue of context: "A science of the influence of context on acceptability judgements is as necessary in linguistic research as in every other area of psychology" (p. 347). First, however, we must set straight exactly what is meant by the term, which tends to be bandied about rather casually. While the common folklore states that "sentences usually sound better in context," we shall see that this really only applies to one of the possible kinds of context. In this subsection we will report on four types of context manipulations. First, we look at a few studies dealing with a context consisting of semantically or pragmatically related content. This is the most extensively studied of the four, and we cannot hope to cover the entire literature here. One large sub-domain we will systematically exclude is the area of discourse-dependent utterance forms such as ellipses, cross-sentential anaphora, etc. (see van Dijk 1977 for discussion); this strikes us as a reasonable omission, because there do not seem to be too many interesting issues that bear on elicitation methodology: obviously if a sentence is dependent on prior sentences for coherence, they must be included when the sentence is judged. The second type of context we consider consists of paradigmatically related sentences; very little work has been done in this area. The same is true for the third type, which consists of the theoretical context under which linguists consider data. The fourth type of context, which seems to have the most insidious impli-

cations for grammaticality judging, is made up of structurally related sentences that can set up extraordinary contrasts or "prepare" us for later sentences.

Bolinger (1968) prefaces his discussion of context effects by saying, "it is worth a moment to consider how a normal sentence can come to be thought abnormal" (p. 35; see also Bolinger 1971). By this he means that disembodying a sentence from its (semantico-pragmatic) context can make it appear unacceptable when in its original setting it was unexceptional. For the most part, the type of judgements he is concerned with are of semantic rather than syntactic well-formedness: for example, he assumes that *I'm the soup* is ungrammatical in isolation. Nonetheless, some of his observations have implications for our study as well. For instance, a sentence heard out of context will tend to trigger dominant senses of the words it contains; once a situation for the sentence as a whole is derived from these meanings, secondary senses are not likely to come to mind, even if they would make the sentence grammatical.[14] Situating sentences within a larger discourse (possibly by expanding a single sentence) also improves their acceptability by providing the motivation for marked constructions, such as the clefts in the following examples:

(5) ?It's a lawyer that he is.

(6) It wasn't a lawyer that he wanted to be but a doctor.

The low-bias reading of an ambiguous word can sound bad out of context, as in the following sentence when spoken:

(7) ?Never have too close friends.

We can reasonably expect that when subjects are asked to judge sentences in isolation, they may attempt to call up a suitable linguistic context. If we provide them with such a context instead of leaving them to their own devices, we will most likely find less variation in the resulting judgements. If we further assume that context cannot make a truly ungrammatical sentence seem acceptable (which is likely true for the vast majority of sentences), we are not biasing the outcome of the experiment by 'giving the sentence its best shot' in this way. Furthermore, by testing the same sentence in multiple contexts, we can examine the grammatical and discourse factors that distinguish various readings.

---

[14] The example Bolinger gives is not a particularly good one: he claims that *The girl was turned to* tends to be considered ungrammatical in isolation because the extended meaning of *turn to* does not come to mind, but I do not have any trouble with this sentence.

This was the kind of context that Nagata (1988) looked at as well (§2.3); recall that while it did cancel the effects of repetition, it made no significant difference to pre-repetition judgements. He points out that often experimenters specifically design their stimuli to be good only under a fairly obscure interpretation, which the context is then designed to bring out. This assessment applies to some of the contexts used in Heringer's (1970) study. For instance, he compared reactions to sentences like (8) with and without the bracketed context:

(8) John left until 6 pm. [John left earlier and is going to come back at 6.]

While none of his 20 subjects accepted (8) without the context, 15 of 39 accepted it with context.[15] It should not be surprising that context improves grammaticality under such conditions, but we cannot conclude from this that *any* semantically coherent context will improve ratings. This was certainly not true for Nagata's contexts, which were appropriate to the target sentence but did not bring out any abnormal readings. Under these conditions, context apparently has no effect, perhaps because some "default" context is assumed when none is directly supplied (Danks & Glucksberg 1971). Snow (1975) refers to this type of context as "paralinguistic context," and suggests that it should always be supplied by the experimenter, to avoid the variation that could results if subjects differ on the contexts they imagine.

Let us turn now to "paradigmatically related" contexts, by which we mean sentences that fill a parallel role in a paradigm. This can best be seen with an example. One finding of Hill's (1961) that probably would hold up under more controlled conditions was that a more structured design (as compared to individual sentence judgements) produced a reduction in inter-speaker variation: this involved presenting several sentence groups following the same paradigm, e.g.

(9) The plate is hot. The plate seems hot. The plate seems very hot.

(10) The child is sleeping. The child seems sleeping. The child seems very sleeping.

---

15 These numbers represent a pooling of subjects who indicated "acceptable" or "uncertain, but probably acceptable" on the four-choice questionnaire. In explaining this analysis, Heringer acknowledges that "some people apparently use a stricter interpretation of acceptability than others, while what is of interest here is not absolute acceptability but relative acceptability with respect to other sentences" (p. 291, fn. 5). There is also variability in the relative certainty of subjects, i.e. some will give many more "uncertain" responses than others. We must be extremely careful thinking about what information we are trying to extract from judgements, in choosing what to do with raw ratings.

In this example, *The child seems sleeping* would presumably be the target sentence of interest. It is surrounded in (10) by two sentences that are related to it in a way that is parallel to the relations among the sentences in (9). This allows subjects to 'see where the sentence came from,' by analogy to an unequivocally good sentence; apparently this procedure helps them to focus on the relevant features of the sentence. This type of parallel analysis is certainly common in linguistic argumentation, but no one else seems to have used it in studying the judgement process itself. In cases where it is feasible, it may prove to be a useful tool. (Recall a related finding by Scott and Mills (1973) from Chapter 2, §3: viewing all the rearrangements of a sentence together increased grammaticality ratings.)

Spencer (1973) mentions a type of context made up of "the set of rules for which [a] sentence is an exemplar"; Snow (1975) seems to mean something similar by the "context of discourse," which she defines as the linguistic issue on which a sentence bears. In both cases, such context is only relevant to linguists, and may actually be entirely implicit, without mention in the materials themselves. For instance, certain examples become closely associated with particular theoretical proposals or disputes by virtue of repeated discussion or published citations. Spencer seems to suggest that when a linguist's initial intuitions about a sentence fail to conform to the context (presumably this means they contradict the rules), the sentence is reorganized to bring the intuition in line. If so, this would be an explanation at the processing level for the hypothesis that linguists' judgements are sub-consciously biased by their theoretical positions. Unfortunately, her experimental results do not show this in any direct way, and it is in fact hard to imagine a conclusive demonstration of this effect, so it must remain as intriguing speculation for now.

Now we focus on the effects of the fourth kind of context, namely the neighbouring stimulus sentences displayed for judgement. It has been common folklore among linguists that marginal sentences can be made to seem more acceptable when preceded by much worse examples (e.g. Snow (1975), who calls this the "context of judgement"; Levelt (1974, vol. 3)).[16] Bever (1970a, 1974) may have been the first to make this explicit, in connection with the law of contrast from psychology:

[16] The use of this effect for purposes of theoretical argumentation is known in some circles as "Chomsky's trick."

> . . . one's "absolute" judgement of a stimulus can be exaggerated by the difference between the stimulus and its context. This influence by contrast clearly can occur in "intuitions" about grammaticality. For example, [(11b)] preceded by [(11a)] may be judged ungrammatical, but contrasted with [(11c)] it will probably be judged as grammatical.

(11) a. Who must telephone her?
b. Who need telephone her?
c. Who want telephone her? (Bever 1970a, pp. 346–347)

Bever describes an analogous effect in color perception: a pale green spot may appear blue when surrounded by a yellow field, but appears green if surrounded by red field. (Although the examples cited here are very closely related, contrast effects can be found with unrelated stimuli—see the discussion of Snow 1975 below.) To test the hypothesis for linguistic context, he proposes taking a bunch of sentences from linguistics articles and presenting them in two different orders to two groups for judgement: one group would see them in their original order as they appeared in the source publications, while the other group would see them in random order. Bever predicts that the former group would come much closer to the published judgements than the latter. Spencer (1973), as part of the study described in Chapter 3, §4.1, did exactly that: in one condition the order of sentences was completely randomized, while in the other they appeared in their originally published order (the order of the articles was randomized). Unfortunately, the results are reported in the same vague manner as her comparison of naïve and nonnaïve subjects: we can see that the mean number of sentences accepted by the two groups differed by almost 6%, but we know nothing of the significance of this difference or to what extent the distribution of good and bad ratings differed for the two groups.

Greenbaum (1976a) performed an experiment that made the same point. The crucial sentences exemplified various uses of the verb *dare*:

(12) We didn't dare answer him back.

(13) We dared not answer him back.

(14) We didn't dare to answer him back.

Two of the three sentences appeared together on one page of the experiment booklet, and subjects were implicitly encouraged to compare the two by having to rank which was better, in addition to rendering absolute judgements on the following scale: "perfectly natural and normal"; "wholly unnatural and abnormal"; "somewhere between"; "not sure." Sentence (14) showed a significant change in absolute rating depending on which

of the other two sentences it was paired with: it was much better alongside (13) than (12). Among the latter two sentences, (12) rated significantly better overall. That is, greater contrast produced polarization of the results: seeing the better alternative, subjects judged (14) even worse. This confirms the prediction made by Bever (1970a), although the results would be more convincing if they could be replicated without any explicit suggestion that subjects should compare the adjacent sentences. Greenbaum's conclusion is that closely related sentences should be presented for judgement as a group, with ordering counterbalanced across subjects, because he believes that in the absence of comparable sentences provided by the experimenter, subjects may try to think up their own related items, so that inter-subject differences in ratings could be related to differences in their ability to make such inventions. (This parallels Bolinger's point for semantic contexts.)

Snow (1975) conducted an experiment that demonstrated contrast effects with unrelated sentences. Her test consisted of alternating target and filler sentences; in one condition all the fillers were clearly grammatical, in the other they were clearly awful. Subjects judged acceptability on a yes-no basis. Although no statistical analysis of the raw data is reported, there was clearly a substantial shift in judgements between the two groups: 18 of the 20 target sentences were accepted by more subjects when surrounded by bad fillers, showing a mean increase of 11.7% in the number of subjects who accepted them; the most dramatic example showed a 32% increase. As Carden and Dieterich (1981) put it, "'ungrammatical' often should be interpreted as 'clearly worse than the "good" examples [a sentence] is being compared to'" (p. 589). They describe a data disagreement over "backwards coreference" constructions such as *I knew him when Harvey was a little boy*, where *him* and *Harvey* are taken as coreferential; those linguists who claim the sentence is bad pair it with a clearly good example, and vice versa. Carden and Dieterich argue that both good and bad related sentences should be presented for subjects' consideration.

The results of some experiments by Warner and Glass (1987) bear on the effects of context by both structural and semantic relatedness, that is, the first and third types in our taxonomy. Their main interest was to examine the processing of garden-path sentences,[17] but what surfaces as well is a striking case of judgements not reflecting the underlying grammar, because a majority of subjects judged sentences bad that are uncontro-

[17] The authors use the term "garden-path" to refer to all sentences with temporary ambiguities, regardless of whether people actually tend to fail on their first parse of them.

versially grammatical. Their design allowed the authors to measure the effects of two kinds of context sentences: those that were structurally similar to the target garden paths and those that were semantically related. Since garden path sentences can mislead the reader by virtue of temporary ambiguities, context sentences could either help or hinder their parsing, by priming either the correct or the misleading choice at the point of ambiguity. Where positive bias was induced, we find examples of the type Nagata mentioned: sentences that would probably be judged bad unless subjects were directed towards the necessary situation or structural analysis. Below are examples of the four possible relations between context and target; in each pair, the first (context) sentence is unambiguously parseable and grammatical, the second (target) sentence is a garden path:

Syntactically related, positive bias:

(15) a. When the girl sleeps the cat eats.
b. When the boys strike the dog kills.

Syntactically related, negative bias:

(16) a. If the girl pets the cat she sings.
b. When the boys strike the dog kills.

Semantically related, positive bias:

(17) a. The cat attacks because the boy harms the man.
b. While the boy kills the man the cat strikes.

Semantically related, negative bias:

(18) a. The boy attacks when the man is hurt by the cat.
b. While the boy kills the man the cat strikes.

Their first experiment elicited speeded grammaticality judgements and found there was a significant main effect of context: sentences preceded by positive-bias context received an average 87% rating, while those with negative-bias context received only 65% grammaticality. There was no significant difference between syntactic and semantic contexts. A subsequent self-paced judgement task showed no context effects in most cases; the authors suggest this change is attributable to the relative speed of judging: at their own pace, subjects would not be reaching final judgement decisions until much longer after reading the context sentences; thus it appears that context-induced priming is a fleeting phenomenon, which may account for some of the discrepancies among other findings. Interestingly, the class of garden paths that are hardest to process, namely those requiring

an intransitive reading of a transitive verb and with a delayed resolution of the ambiguity (e.g. *Before the boy kills the man the dog bites strikes*), were judged grammatical only as often as ungrammatical control sentences (e.g. *Who is strong killed that strike men*) in the absence of any biasing context: 25%. Apparently, people are either not very persistent or not very creative in looking for alternative parses, because this result held up in the self-paced experiment as well.

Milne (1982) presents both anecdotal and quantitative evidence corroborating this finding for other kinds of garden paths: for instance, when asked to judged whether *The prime number few* was a complete sentence or only a fragment, all 47 of his subjects thought it was a fragment. In a timed comprehension task, *The horse raced past the barn fell* took an average of 10.13 seconds to read, with many subjects still reporting they had not understood it after that period. (Bever (1970b) hypothesizes that such reduced relative garden-path sentences would be parsed much more readily if an example pair consisting of full and reduced versions of a sentence were presented first.) Thus, methodological caution is advised: if we suspect that the specific reading of a sentence that we want to test is hard for human parsers to arrive at, independent of whether it is grammatical or not, we should make every effort to ensure subjects think of the right reading; otherwise, rejections on the basis of ungrammaticality are confounded with those based on never having 'found' the sentence in question.

Our main conclusion from these studies of context is that it does not make sense to speak of *the* effect of context on judgements, because the type of context and its relation to the sentences in question must be considered. It can be used to make sentences look better or worse than they are in isolation; whether this is desirable will depend on the goals of particular investigations.

3.2 Meaning

The earliest study that I am aware of that looked specifically at the nature of linguistic intuitions as expressed in judgements about sentences was done by Maclay and Sleator (1960). They were specifically interested in the extent to which subjects could judge grammaticality independent of meaningful content and likelihood of occurrence, so they asked subjects whether each stimulus sentence was grammatical, meaningful and ordinary. (By the latter term they meant "occurring with high frequency," so that this portion of the study might belong in §3.3, which deals with frequency, but since it is not

clear whether their subjects interpreted it this way, we keep the main discussion together.) Unfortunately, they apparently did not give subjects any further explanation as to the intended meanings of these terms, and some of their results clearly show that at least some subjects did not take the desired interpretation. One good feature of this procedure, however, is that it allows subjects to voice opinions on these issues separately: if someone feels a sentence is meaningless or has no chance of occurring in natural speech, they will want to convey this opinion; if they are not asked specifically for the information, they will likely allow it to affect their responses on other matters, e.g. grammaticality (Elizabeth Cowper, personal communication). The experimenters had designed the sentences to represent various combinations of the three dimensions, e.g. grammatical but not meaningful and not ordinary. In addition, one group of grammatical sentences contained deliberate violations of "grammar school" rules, e.g. incorrect uses of *I/me*, that do occur in casual speech and ought (according to Maclay and Sleator) to be generated by a linguistic grammar. Sentences were presented orally with "normal" intonation.[18] For sentences that were intended to be grammatical but not meaningful or ordinary (e.g. *Colorless green ideas sleep furiously*), significantly more subjects said "yes" to the grammaticality question than to the other two questions. Maclay and Sleator take this as evidence that subjects were making the intended distinctions and judging grammaticality independent of the other two variables (but see below). Across all the sentence types and all three rating criteria, the relative ratings conformed to prior classifications: positive instances got a greater percentage of "yes" responses than non-instances. However, the absolute numbers were less convincing: the aforementioned grammatical–not meaningful–not ordinary sentences only received 50% grammatical rating, as did those that violated only the prescriptive rules;[19] the other absolute numbers were similarly disappointing, often indicating approximately neutral ratings on average for all three criteria. From their lack of clear-cut outcomes, the authors conclude that there is no empirical basis on which to classify sentences as grammatical versus ungrammatical, or even into multiple discrete levels of grammaticality, but we must be content with comparative rankings only. How they can discount the latter possibility without attempting to elicit multi-valued ratings is not clear to me. But perhaps the most telling part of their conclusion is the admission that 3 of their 21 subjects said that *Label break to calmed about and* was grammatical; since

---

18 It is not clear to me how the strings of word salad could be read with "normal" intonation; a standard contour would have to be placed arbitrarily over the words.

19 Bradac et al. (1980) also looked at errors of "school grammar," but found most subjects oblivious to them.

these were all native speakers of English, they clearly were not applying grammaticality in the intended way, and so the experimental results do not represent measurement of a unitary phenomenon.

Vetter, Volovecky, and Howell (1979) performed a follow-up to Maclay and Sleator's experiment, because they felt that the latter authors did not have statistical justification for the claim that grammaticality was being judged independently of the other two variables; their new experimental design allowed for the direct assessment of such claims. They used the same 36 stimulus sentences as Maclay and Sleator, but made a small attempt to improve the instructions; for instance, in one condition subjects were asked, "Is this word sequence grammatical? In other words, is it acceptable English?" In my opinion, this still leaves a lot of room for interpretation, and this time we have direct evidence concerning how the subjects tried to perform the various discriminations; but first, let us look at their results. As in the previous study, an ANOVA showed a significant effect of type of sentence: grammatical versus meaningful versus ordinary, but Vetter et al. correctly point out that such a finding is difficult to interpret because these do not represent values on a single dimension. Pair-wise chi-square tests of independence showed that in some of the conditions grammaticality and ordinariness were significantly related, whereas in other conditions meaningfulness and ordinariness were related. Thus, this study contradicts the earlier claim and suggests that these factors do influence each other. Other results largely replicated those of Maclay and Sleator: sentence groups that were supposed to differ only on grammaticality did show a significant difference on that parameter, and similarly for the other variables. Once again, however, the most definite conclusion we can draw from this study is that much more work is needed on conveying to all subjects the same notion of grammaticality, as evidenced by the following remarks from Vetter et al.'s high-school aged subjects regarding how they decided whether a sentence fit the criteria. For grammaticality, they considered punctuation, making sense, whether the sentence was "smooth," and "what I learned in elementary school about correct grammar"; for meaningfulness, they considered "pausing and verbalization," "if words could be rearranged to make sense," whether the "order of words seemed similar to reverse order in German," and whether it was "true or something that could happen"; for ordinariness, they considered that "word order inverted was ordinary, since it's natural in French," answered 'yes' if the sentence "didn't make sense but had normal subject and

verb order," and factored in "the way the words were typed."[20] From these reports it is clear that the ratings could not represent anything approaching a unitary phenomenon, a fact that may invalidate all the other conclusions of the experiment anyway.

One study took a different analytical approach to the problem of factoring apart meaningfulness and grammaticality:[21] Danks (1969) was interested in the comprehension process, and so used comprehensibility ratings to see what variables are relevant and how consistently it is judged, although he explicitly suggested to subjects that comprehensibility might involve grammaticalness, meaningfulness and familiarity, among other things. He manipulated grammaticalness, meaningfulness and word frequency (and other variables) in the stimuli but then elicited a single kind of judgement, comprehensibility, on a scale of 0–10, and used principal components analysis to examine the relations between them. His conclusion was that grammaticalness and meaningfulness were the only important factors in these ratings, the latter having a stronger influence. To the extent that comprehensibility tends to affect people's grammaticality ratings, this result is consistent with the notion that meaningfulness cannot be entirely separated from grammaticality.

3.3 Frequency

Another possible factor in judgements is the frequency of occurrence of the stimulus materials. This could be taken in at least two ways: to refer to the frequency of the lexical items in the sentences, or of the sentence structures themselves. I am not aware of the former having been studied, but Greenbaum (1976b, 1977b) looked at the latter.[22] His experiments involved judgements on closely-related sentence pairs such as active–passive and dative movement contrasts. In the first phase, subjects (who were linguistically naïve) had to judge the "overall frequency in the English Language" on a 5-tiered

---

20 There were also some responses that seemed to bear some resemblance to the desired interpretation of the terms.

21 Also, Moore (1972) (see Chapter 2, §3) reports a case where the existence of a metaphorical reading of a literally ungrammatical sentence may have contributed to it being rated significantly higher than other structurally-identical ones, which could be viewed as a meaning confound. The sentence in question was *College students read many professors*, which supposedly violates selectional restrictions on the verb *read*. But, as Moore correctly and humorously points out, "A college professor may be read in the sense that Plato is read; alternatively, professors may have such transparent neuroses that they are easily 'read' by their students" (p. 558).

22 Note that his measure of frequency was subjective, i.e. people's judgements of it, rather than objective, as might be obtained by corpus analysis.

scale from "very rare" to "very frequent." In the second phase, which occurred a week later, the same subjects were asked to rate *acceptability* of the same sentences, again on a 5-tiered scale, from "completely unacceptable" to "perfectly OK." Greenbaum compared mean numeric scores across 87 subjects and found that, for each sentence, the acceptability rating was within one point of the frequency rating. (In most cases, acceptability was rated higher than frequency.) On the surface, this suggests that the two ratings are highly correlated, but no statistical tests were carried out and there is a potential confound in the experimental procedure: we do not know if subjects were aware that they were supposed to be judging something completely different the second time; they may have taken the instructions as merely a variation in wording of the original procedure. (This possibility could easily be tested by a between-subjects design.) To the extent that Vetter et al.'s results have any validity, their finding that grammaticality was not independent of ordinariness points in the same direction. Greenbaum also examined the data on a subject-by-subject basis, finding that while identical ratings on the two scales were relatively rare, ratings within 1 point of each other occurred for 88% of the sentences among 65% of the subjects. By this measure there was also reasonable consistency in the relationships between the ratings for the members of a related pair of sentences, i.e. whichever of the two was rated more frequent (e.g. the active version) would be rated more acceptable by half of the subjects in 64% of the cases. If Greenbaum's interpretation is correct, we must be wary of grammaticality judgements on very obscure types of sentence constructions, which may reflect their infrequent nature despite their grammaticality. Following one of our earlier suggestions, a way to reduce this effect might be to allow subjects a separate frequency rating when judging acceptability, so that they can "express" this intuition and perhaps factor it out of the other judgement. This is presumably what Maclay and Sleator were trying to do with their "ordinariness" scale, although the meaning of the term was probably obscure to most subjects. We must also keep in mind, however, that Greenbaum has only shown a correlation between perceived frequency and acceptability, with no evidence about causality.

Some other interesting results were by-products of the fact that sentences were presented in related pairs such as active–passive. In general, across various other constructions actives are judged more acceptable than passives (no analysis of significance was done by Greenbaum, but on the basis of his sample size and reported means I expect it would hold up). Such a bias must be taken into account in other analyses. For example, in Chapter 1 we cited an instance where the fact that the active and passive versions of a sentence were equally marginal formed part of a theoretical argument (example (69)

from Belletti & Rizzi 1988). Of course, no experimental data had been used, but if they were, the general bias against passives would probably cause the result not to hold up, and yet this difference would have nothing to do with the particular theoretical issue involved. Other general biases were found as well, e.g. favouring present perfects over simple past tenses with both durative and iterative events, and subjunctives over indicatives and modals in subordinate clauses of demand or persuasion.

3.4 Lexical Content

Levelt et al. (1977) wanted to bring home the point that the particular lexical items chosen could affect grammaticality ratings even when, from a normative point of view, the range of words should all result in a grammatical utterance. The particular feature of lexical items they investigated was the imagery content of compound words in Dutch, by which they meant the extent to which they were concrete as opposed to abstract, and hence more easily imaginable. (This idea was inspired by phenomenological reports of subjects performing judgement tasks by trying to imagine a situation in which the phrase in question could be said.) RT was measured as grammaticality ratings[23] and paraphrases of novel Dutch compounds (not complete sentences) were elicited. The purpose of avoiding lexicalized compounds was to encourage computation of a rating, rather than the use of lexical 'look-up.' The basic prediction of Levelt et al., that the facilitation effect of imagery would show up in these judgements, was confirmed: more easily imaginable compounds were rated significantly more grammatical and judged and paraphrased faster than harder-to-imagine ones,[24] supporting the notion of search for interpretation as (part of) the decision procedure. (The results for speed of judgement are confirmed by results in Grant et al. 1977). The additional twist, however, was that imagery showed a much greater effect on RT for paraphrasing than for judging grammaticality. Levelt et al. considered two explanations: first, perhaps imagery is involved in the task of generating a paraphrase, as well as in comprehending the original sentence; they consider this unlikely, for reasons not given. Second, perhaps it is the grammaticality judgements that do not require full "imagistic search"; that is, once a preliminary check (involving some

---

23 Subjects were asked to rate the extent to which the stimulus could be a Dutch compound word, i.e. was "good Dutch." As the authors rightly recognize, "it seems wrong to ask a naïve subject to use an ill-defined linguistic technical term" (p. 94).

24 The imagery content of compounds was assessed in a pretest where other subjects reported how easily the expression lead to mental images of things or events.

imagistic component) succeeds, the compound does not need to be fully processed or understood before being judged as good. As the authors suggest, this could be tested by timing a task that requires full interpretation but does not involve paraphrase, such as verification (truth judgement). Whichever explanation is correct, we have another case of an "irrelevant" variable influencing judgements of grammaticality. McCawley (1985) views this as an instance of inaccurate *reporting* of judgements, i.e. the subjects think they are reporting on grammaticality, but really are reporting their (degree of) success in imagining a context.

Another of Greenbaum's (1977b) experiments also examined the effects of lexical substitution by comparing acceptability ratings between two instances of the same sentence structure that differed only on certain lexical items. (He does not give a large range of examples, but the intent seems to have been that these substitutions could reasonably be expected to have no grammatical implications.) Again, judgements were on a 5-point scale; for 27 of 50 sentences at least half of the subjects gave identical ratings to the two lexical variant sentences, and on 47 out of 50 70% of subjects were within one position on the scale of their initial judgement. Variants were presented in separate parts of the questionnaire so that memory would be extremely unlikely, and subjects were explicitly told not to try to remember their earlier ratings. A third study that found grammaticality differences triggered by lexical substitution was performed by Sobin (1987), mentioned in Chapter 1. The focus of the study was the comp-trace effect, whose grammaticality varied as a function of the particular complementizer (e.g. *Who did you ask if/whether Bill kissed?*) and the matrix verb. In these cases, however, it is possible that some of the items involved do differ on grammatically relevant properties, so we cannot draw many conclusions from it.

3.5 Morphology and Spelling

Langendoen and Bever (1973) claim that there are acceptability differences that depend on the transparency of morphology in cases where a pronoun later in the sentence refers to an implicit morphologically-related word, e.g. (19) versus (20):

(19) ?Flutists are strange: *it* doesn't sound shrill to them.

(20) *Flautists are strange: *it* doesn't sound shrill to them. (p. 407)

They claim (1) is more acceptable because *flute*, the implicit antecedent of *it*, is more obviously part of *flutists* than *flautists*; unfortunately, they do not cite actual data on this

point: they may simply be expressing their own intuitions. To my ear, both sentences are equally (quite) bad, but if such a result should be found systematically, it would constitute another factor to be taken into consideration. Some authors (e.g. Birdsong 1989) have cited this paper as showing that variant *spellings* cause changes in acceptability ratings, but Langendoen and Bever make no mention of spelling, which is confounded with pronunciation in this case. Still, it would not be surprising to find acceptability differences between alternate spellings of identically-pronounced variants, e.g. *night* versus *nite*, triggered by relative frequency in the dialect of the subject; I am not aware of this issue ever having been studied. The flip-side of this point is made by Hill (1961): before people can judge a sentence, they must first identify all the morphemes in it, but the presence of a normal intonation contour can lead one to interpret apparently familiar morphemes as novel ones. For instance, the sentence *I saw a fragile of*, read with declarative intonation, leads speakers to think that there is a novel noun *of*, rather than identifying it as the familiar preposition. This can lead them to judge bad sentences good unless we inform them that "all the words will be familiar," or some equivalent statement.

4. Conclusion

In this chapter, we have seen at least suggestive evidence for effects on grammaticality judgements of just about every stimulus and procedure variable one can think of. Serial order, repeated presentation, deliberate judgement strategies, modality, register, preparation and judgement speed are all features of the elicitation process that may contribute systematically to variation in judgements. So can stimulus features including the various types of contextual material, the meaningfulness of the sentence, the perceived frequency of the sentence structure, and various idiosyncratic properties of its lexical items. We have also seen some unpredictable interactions between variables, such as context with repetition, and mental state with repetition. But perhaps the biggest lesson is the importance of the instructions we give to subjects. In the face of all the disparity in subject interpretations of the intended tasks, there is a strong temptation to propose that our first order of business should be to replicate all these studies with much more carefully designed instructional procedures. There can no longer be any doubt of the importance of this experimental design feature with regard to the elusive definition of grammaticality/acceptability, and until this knowledge is acted upon we cannot say much about the other experimental factors with any certainty. Nevertheless, we will attempt to derive some methodological recommendations from the findings of this chapter; these will be presented in Chapter 5, §3.

Chapter 5

Theoretical and Methodological Implications

More and more subtle theory is now being constructed on less and less clear cases. In such a situation one would expect linguistics to turn to appropriate behavioral methods of data gathering and (statistical) analysis. Nothing of the sort occurs, however . . . What sort of process underlies the formation of a grammaticality judgment? The only way to approach this question is to ignore all a priori *linguistic restrictions and to regard it as a problem in human information processing.*

(Levelt et al. 1977)

1. Introduction

The purpose of this chapter is to bring together many of the issues raised and results reviewed in the preceding chapters to consider what we can learn from them. We will entertain two particular angles on that question, namely, what we can learn about what goes on in the mind to allow grammaticality judgements to be made, and what *should* go on in linguistic experimentation in order for those judgements to be useful.

In §2 we will take up the idea proposed in Chapter 1 that a useful way to make sense of a large collection of diverse experimental results is to try to fit them into a single coherent model of the mental structures that underlie the behaviour. That is, following the advice of the epigraph above, we treat this like any other problem in human information processing. In §2.1 we review what very little previous work of this sort has been done. In §2.2 we propose a preliminary model of our own that provides a useful way of picturing how the many mental components discussed so far might be brought together in the judgement process. Then we give some examples of ways in which this model, in conjunction with specific assumptions about how the underlying cognitive processes work, could account for some major findings (§2.3). In §3 we move on to the applied issues. In light of the demonstration in Chapter 1, §3, of the apparent necessity for complex and fine-grained judgements being used in current theoretical argumentation, and the dismal record of "insufficient reporting of results or data, poorly elaborated stimulus materials, or . . . lack of adequate controls" (Chaudron 1983, p. 367) evinced in experimental

work to date, it should be obvious that considerable care and effort must be put into the elicitation of grammaticality judgements if we are to stand a chance of getting consistent, meaningful, and accurate results. As is also pointed out in the epigraph to this chapter, this is not being done; therefore, we will make specific recommendations on how to improve almost every phase of the experimental process. We will examine first the stimulus materials (§3.1), then the elicitation procedures themselves (§3.2), and finally the statistical analysis and interpretation of the results (§3.3). We conclude in §4 by exploring the potential benefits of the theoretical hypotheses and methodological proposals put forth in the chapter.

2. Modelling Grammaticality Judgements

2.1 Previous Work

Almost no work has been done to date by way of modelling the psychological representations and processes involved in making grammaticality judgements, despite the proliferation of models of other language behaviours, most notably sentence processing. The first attempt in this area that I am aware of was the work of Bialystok and Ryan (1985), described in Chapter 2, §4. From our earlier description, it should be clear that this is a very high-level model, whose constructs are so abstract as to have almost no concrete content. This is not to say that they do not exhibit useful insights, but the details are left for others to work out. While the authors claim that the dimensions of analyzed knowledge and cognitive control underlie many of the more specific properties on which people differ, e.g. field dependence, nothing more specific is said. At the level of detail we wish to work, they do not have much to offer us.

One other line of work that could be considered a model of certain aspects of the grammaticality judgement process is that of Catt (1988; Catt & Hirst 1990), although this was not his main goal. Catt designed a computer program for computer-assisted language instruction that was designed to perform automatic error diagnosis and correction of ungrammaticalities produced by second-language learners. In effect, the system was a model of a foreign-language instructor: it would classify the source of errors in a sentence as phrase structure, transformations, morphology, verb subcategorization, or certain direct translations from the learner's native language. The heart of the system was a parser made up of constraints that could be selectively relaxed when an initial parse failed: once a parse was eventually found in this way, the constraint(s) that had been relaxed indicated the nature and location of the ungrammaticality. (We will return to the idea of constraint

relaxation in §2.3.) It is possible that people do something similar when they encounter ungrammaticality, and if so, the nature and degree of constraint relaxation might be reflected in their grammaticality ratings. Unfortunately for us, Catt was not concerned with extra-grammatical factors that could enter into this process, so we shall have to forge the rest of the way alone.

2.2 A Preliminary Model

In this subsection we will propose some high-level accounts of the phenomena discussed so far in this paper, culminating in suggestions for what the relevant cognitive representations might be and what the steps in the grammaticality judgement process might be. The motivations for modelling this process were given in Chapter 1 and will not be repeated here. At this point it is important to stress the preliminary and speculative nature of these proposals: much 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. In addition, what our model should look like will depend in large measure on many larger unanswered questions in language processing that also display a lack of well-articulated, well-motivated models, since a major issue of interest to us is the interface between metalinguistic components and those related to regular processing. In these cases we can only adopt some fairly well-accepted assumptions and proceed.

To begin, we ask whether there is in fact a static representation of grammatical knowledge independent of production and comprehension mechanisms? Lachter and Bever (1988), among many others, have argued for such a "psychogrammar," "an internalized representation of the language, that is not necessarily a model of such behaviors as speech perception or production, but a representation of the structure used in those and other language behaviors" (p. 221). Although I find their arguments thoroughly unconvincing, I will adopt this assumption, in part for clarity of exposition. (It may well be that our concept of the grammar as a separable black box of static knowledge will eventually have to change; perhaps it is not encoded separately from the production and comprehension mechanisms, but this might still be the most useful way to think about linguistic competence.) The next question is, what does the grammar look like? Here I will assume a principles-and-parameters model of Universal Grammar (UG), but only because it is the theory I am most familiar with. Now, what is the relationship between the principles and parameters, the grammar as a whole (including areas not covered by UG), and the parsing and processing mechanisms? Again mostly for expository ease, I shall assume that UG is

part of the grammar that is separate from processing mechanisms, which are based on it in some unspecified way, but function independently. (See Gerken & Bever 1986 for discussion of the possible relations between linguistic universals, language-specific sentence structures and perceptual mechanisms.)

The major question we are attempting to answer by way of a psychological model is summed up by Klein (1979): "We must ultimately answer the question as to how much of acceptability, or what kind of acceptability, falls within the scope of grammar, and how much is to be accounted for by other parts of the linguistic description or by disciplines outside linguistics" (p. 8). This brings us back to two major issues touched on in Chapter 2: how does linguistic judgement differ from language use, and are any of the manipulations to which judgement is susceptible particular to language? In Chapter 2, §4 we discussed the conceivable extremes in response to the first question: judgement involves all of the same components as conversation, or the two processes are entirely separate. Under either of these scenarios, the answer to the second question would be fairly uninteresting: if the mechanisms are identical, the only possible source of influence on judgements is the comprehension mechanism, so whatever we cannot attribute to that (linguistic) faculty cannot be accounted for; on that basis, we can probably rule out this model immediately. If the mechanisms are separate, there is no potential for "normal" language mechanisms to contribute to linguistic judgements, but since no other process relies on the judgement module, it can have arbitrary properties: whatever effects we find must originate there. However, if we make the reasonable assumption that reality lies somewhere between these extremes, then the question of how the various components contribute to the total process becomes more interesting. Figure 1 represents a first attempt at demonstrating this scenario. It must be stressed that this view is based on my impressions as accumulated over the course of this work; below I will explain what I have in mind with the various pieces of the picture, but I cannot justify the details in any rigorous way. (With reference to our discussion in Chapter 2, §6, of whether nonlinguistic properties should be attributed to separate modules of the mind or to the underlying substrate, we have taken the former interpretation here mainly for diagrammatic convenience.)

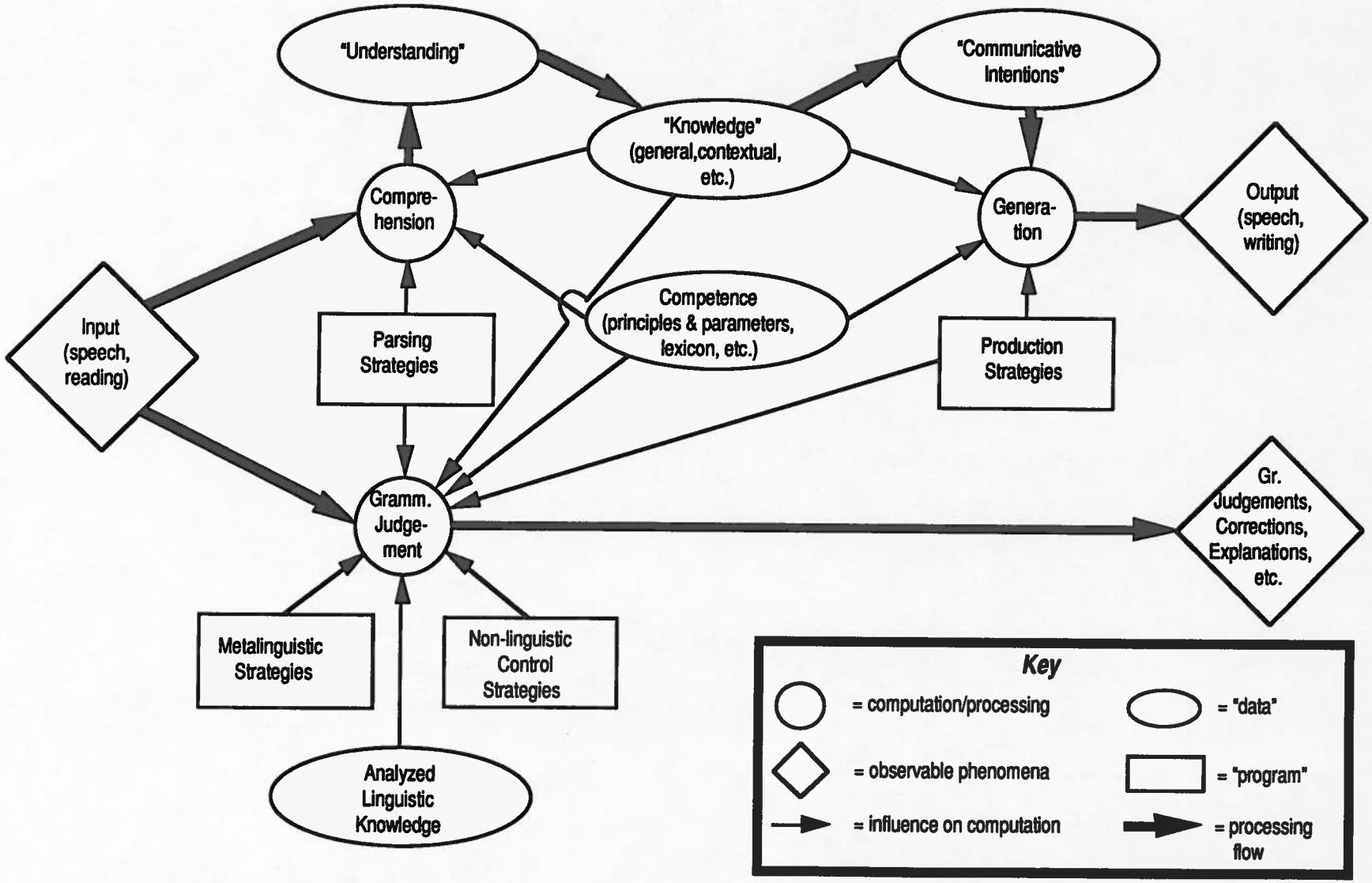

Figure 1. Proposed model of the mental components of linguistic and metalinguistic performance

The basic computational metaphor used in Figure 1 is that a program, procedure, or strategy uses static data or knowledge to process incoming information and yield information as output; the entire process is viewed as a computation. Thus, each computation symbol is connected by arrows to the program(s) that execute in order to perform that computation and the data source(s) that must be referenced; the thick arrows represent the processing flow. Let us first examine the upper portion of the diagram, implicated in the comprehension and production of language. Working from left to right, we see that the input to the system is a linguistic signal (in whatever modality), which undergoes a comprehension process to yield an "understanding" of the sentence (whatever that means); we take this understanding to be a *temporary* product of the computation, comprising information that may or may not subsequently be stored as part of one's long-term knowledge. Comprehension involves the use of parsing strategies, which we construe quite broadly here to include heuristics that do not involve assigning any hierarchical structure to the sentence string, such as interpreting a Noun-Verb-Noun sequence as Agent-Action-Patient; we certainly do not wish to claim that every sentence is assigned a complete constituent structure by the parser. Additionally, comprehension draws on information from competence (taken broadly here to include the lexicon); it also makes reference to general knowledge and specific memories to resolve ambiguities, etc. On the production side, an utterance begins with the intention to communicate something; that message is used in the process of generation, which employs production strategies (whatever it takes to produce a sentence word by word) that also make use of linguistic competence and world knowledge (e.g. to decide which referential expressions are appropriate) and yield language output (in whatever modality).

Now let us examine the lower portion of Figure 1, which illustrates the grammaticality judgement process. The input to this process is the same sort of input as to the comprehension process, namely sentences. The output will consist of judgements themselves, plus other related information that might be elicited in response to presentation of a sentence (for more examples, see Chapter 2, §2); since these need not be expressed through language (they could involve circling numbers on a questionnaire, for instance), no connection to the generation process is shown, although language generation will incidentally be required in some cases. The central question is, what processes or strategies are operative in generating these judgements, and what sources of knowledge do they draw upon? The answer provided in the diagram is in some sense a maximal one: we have included all the major components that *might* be involved; perhaps not all of them are, and certainly they need not all be implicated in judging a particular sentence. How-

ever, we will accomplish nothing in modelling if we merely supply a new component for each new experimental result we wish to account for; therefore, we must hypothesize the minimal set of components that could reasonably capture the range of facts we are concerned with. (I do not claim to have struck the perfect balance between these criteria.) Let us consider these components in turn. On the assumption that one of the first things one does when processing a sentence for judgement is to simply try to understand it, the usual parsing strategies (in our broad sense) will be involved, and therefore by assumption also the linguistic competence that they may draw on. If this step is exactly like normal comprehension, then general knowledge must also be involved; in reality, I suspect its role is somewhat smaller in the case of judgement, if concerns about plausibility, truth, etc. might be suspended, but world knowledge is also relevant to grammaticality per se, as we saw in the examples from Belletti 1988 in Chapter 3, §4.3.

The remaining influences are likely to be more controversial. The diagram suggests that production strategies might be involved, perhaps minimally in simple yes/no judgements, but more significantly in the scalar rating, locating, explaining and correcting of errors. Intuitively, it seems plausible that all of these activities involve comparison to a correct or predicted version of the sentence, and so it must be generated somehow; I am assuming that, while the parsing mechanism might embody some expectations about its input, it is not capable of generating a grammatical sentence from an intended meaning. In rating a marginal sentence, for example, one might first extract the intended meaning, then generate a grammatical sentence that is the expression of that meaning, then compare the two to decide how far off the original sentence was. But we cannot simply employ the regular generation process for this purpose, because we wish to follow the syntax of the original sentence insofar as it is correct; hence the use of production strategies operating on the input utterance.

Finally, we consider those components implicated in grammaticality judgements but not standard linguistic processing. The "analyzed linguistic knowledge" component is taken directly from Bialystok and Ryan, and is meant to include consciously-known rules of language such as the prescriptive rules one learns in school. These would be responsible for labelling a sentence as ungrammatical if it ends with a preposition, for instance. The control strategies are also inspired by Bialystok and Ryan's proposal; these strategies, which I take to be independent of language or any other particular cognitive function, are responsible for bringing the focus of attention to the form of an utterance, coordinating all the sources of information brought to bear in the process, perhaps even

coordinating the other sets of strategies. The latter might be required when, for instance, the parser fails on a sentence that one has been told is grammatical, e.g. a garden-path. While many parsing models have been proposed to account for initial parsing failure on such sentences, none to my knowledge have dealt with how one can eventually succeed in finding a valid structure for the sentence. One possibility is that control strategies intervene to prompt a search for obscure alternatives, such as verbs that could also be passive participles, as in *The horse raced past the barn fell*. Another possibility is that some sort of sentence-frame matching takes place, so that while attempting to figure out *The prime number few*, one would try to think of other sentences that end in *number few*.

The final set of strategies has been simply labelled as metalinguistic, following the terminology we have seen in much of the literature. This module is meant to include any algorithm or heuristic used expressly for the purpose of making grammaticality judgements (in the broad sense). This is where we would find, for example, the strategy of trying to imagine a plausible context for some questionable sentence, or enriching the given context to make it seem more natural, as proposed by Nagata and discussed in Chapter 4, §§2.3–2.5. It would also include procedures for interpreting the trace of execution of the parser (if this is possible) or the final state of the parser; for instance, failure with a certain set of conditions indicates that the parser could not find a suitable antecedent for some anaphor. Other possible strategies include thinking of a parallel construction to compare the given sentence to (e.g. *The child seems tired* for *The child seems sleeping*); considering the truth or plausibility of the situation described, by consulting world knowledge (to the extent that these criteria influence judgements); preliminary checking routines that would be run before parsing begins, as proposed by Levelt et al. (1977); and other strategies discussed in Chapter 2, §4. Recall that in Chapter 2, §6, it was hypothesized that there are no *language-specific* extra-grammatical factors involved in the process of grammaticality judgement. It might appear that the presence of a set of "metalinguistic strategies" contradicts this hypothesis, but in fact it does not, unless the strategies contained in that module require language-specific means to be implemented. For example, the imagining of plausible contexts could be carried out in the same way as any other sort of imagination, (presumably) independent of language; the fact that such a strategy can be *employed* for the purpose of judging grammaticality must be a fact about language behaviour, however. One possible component that we have not shown explicitly is a decision-making or judgement component for cognition in general, which might reflect, say, the tendency to use rating scales in a particular way. Obviously the mind must have a mechanism for decision-making, and it will be implicated in judgements, but it is hard to

conceive of this as a separable module of the mind, as opposed to an inherent feature of the low-level processes that underlie higher cognition. We also have not included any reference to general cognitive processes such as analogical reasoning, and the oft-mentioned but ill-defined "perceptual strategies" that are sometimes implicated in language processing; at this point I consider the evidence for their involvement to be marginal at best. A couple of other omissions are mostly for the sake of clarity. For instance, it is possible that some or all of the three components that are shown connected exclusively to the judgement computation could be used in everyday comprehension if some special need arose. Also, under certain circumstances, regular communicative output could be "filtered" through the judgement component as a means of quality control or self-monitoring; perhaps language teachers regularly do something like this, and everyone might do it to some extent (Graeme Hirst, personal communication).

If this picture is at all on the right track, then it is clear that we do not expect judgements to give all sentences the same status as linguistic performance does; the differences would be attributable to any or all of the components discussed in the previous two paragraphs. (The results of different metalinguistic tasks will differ as well, by virtue of how the contents of the modules are used, e.g. which metalinguistic strategies are employed under what conditions.) In fact, it might appear that grammaticality judgements are the *worst* way to get at linguistic competence, as compared to production and comprehension, because they involve the interaction of many more factors. This conclusion has been reached by others before: "Contrary to what has generally been claimed, the relations between explicit intuitions and underlying competence are *less* direct than those between phenomena of primary language use (speaking, listening) and competence" (Levelt, Sinclair & Jarvella 1978, pp. 5–6). However, this by no means constitutes grounds for abandoning them as a source of data. Several arguments for their use were presented in Chapter 1, §1; we can now provide two more. First, while more factors are involved, they might be less mysterious than those that are connected to language use: for instance, what exactly is the "understanding" of a sentence, and how would we ever get at it and draw conclusions about grammaticality from it?[1] The same goes for communicative intentions. Second, grammaticality judgements provide an *alternative path* to the

[1] Peter Reich (personal communication) suggests that this line of reasoning might be a case of "looking for the keys under the lamppost," i.e. that issues of what understanding is will *have* to be faced and comprehension data will *have* to be used to get at the grammar. Of course, we cannot know for certain how important a particular source of data is to a problem until we actually solve it, but it seems to me that the short-term prospects for getting useful information from judgements are considerably brighter than from comprehension.

grammar; to the extent that they are subject to *different* influences from language use, we have a basis on which to search for the common core that underlies both kinds of behaviour. This in turn might help us to factor out various nongrammatical contributions, so that each path by itself becomes more informative as well. Studying nonlinguistic judgement tasks in the framework of our model, as proposed in Chapter 2, §6, will aid in this process by helping to clarify the role of its nonlinguistic elements.

One thing that is not reflected explicitly in the model presented in Figure 1 is the sequence of steps in the process of judging grammaticality; given the number of factors involved, it seems unlikely that it would all occur in parallel. My hunch is that it can be usefully thought of in three main phases: parsing, decision, and diagnosis. The first is fairly uncontroversial, when taken to include the kinds of heuristics mentioned earlier. The second involves determining what the rating will be: yes/no or a number on a scale or whatever. This might involve an analysis of the remnants of the parsing process, for instance. The third, diagnosis, which might not always occur, involves *consciously* determining reasons for the particular ratings chosen, which could be used in explaining the error or correcting it; as we have discussed before, these may or may not bear any relation to the actual causes of the decision. This sketchy account will suffice for the rest of our discussion here; obviously many details remain to be worked out.

2.3 Applications of the Model

In order to see how my model might actually account for some interesting facts about grammaticality judgements, a bit more needs to be said about how the pieces of this picture are implemented. I will assume the widely-used principle of spreading activation, together with the parallel-processing concept of competition,[2] that is, multiple processing paths might be active in parallel, and the first one to succeed will determine the outcome of the computation. This approach has already been used by McRoy and Hirst in so-called race-based parsing (McRoy & Hirst 1990). In their model, each parsing rule takes a certain amount of time to execute, determined partially by its complexity but also influenced by the lexical content of the sentence, previously-parsed sentences, etc. Multiple parse paths are tried in parallel, and whichever succeeds first is used to interpret the sentence; if none succeeds before a time-out deadline, the parse is considered to have failed. Thus, different readings of a structurally-ambiguous sentence may be found on

2 This is just one possibility that strikes me as promising, but certainly not the only one.

different occasions because the time weights associated with the relevant rules can change. Many of the order-of-presentation effects can be accounted for; for instance, a sentence that could not be parsed at one time (e.g. a garden-path) can be parsed at some other time if the required rules have been strengthened or sped up; this might be accomplished by employing them in parsing structurally similar non-garden path sentences. More generally, context effects due to structural similarity or dissimilarity (see Chapter 4, §3.1) can be accounted for in this way.

This scheme can be extended very naturally to deal with graded judgements of (un)grammaticality; we shall incorporate the idea of constraint-relaxation from Catt's (1988) work, discussed in §2.1. While it is fairly clear that relaxing constraints will allow ill-formed input to be processed and the well-formed parts recovered, not all constraints can be of equal status if we are to get graded judgements based on grammatical properties (as opposed to imagery content of lexical items or other nongrammatical features). That is, it is not sufficient, as it was for Catt, to know *which* constraint must be relaxed in order to "get through" the sentence; we must know how much of a "concession" it was to relax that constraint. While one can imagine an ad hoc weighting function for this purpose, it is perhaps more naturally served using the race-based framework: the speed with which a parse path can be followed if its constraint is violated is inversely proportional to the importance of the constraint on that path. That is, a path with a very fundamental constraint (e.g. the lexical category of an input word) can be traversed only extremely slowly if the constraint is not satisfied, whereas a path with a lesser constraint (e.g. a semantic feature restriction on an input word) can be followed faster than the path with the greater constraint when *its* condition is not met, other things being equal.[3] Thus, degrees of grammaticality could be equated with parsing "speed," where it is understood that speed is meant in a somewhat metaphorical sense that need not correspond to actual processing speed.[4]

---

3 It seems that this approach could by applied quite directly to account for prototypicality effects of the kind discussed in Chapter 2, §3: if an exemplar of a concept (e.g. bird) lacks a fundamental property (e.g. ability to fly), it will take longer to verify than one that lacks a less-important property of the prototype.

4 This scheme bears a certain resemblance to the concept of fuzzy grammar (see Mohan 1977 for a concise review). The idea was that each derivation rule that applied to generate a sentence would rate its output on the basis of features of its input, eventually yielding a well-formedness rating between 0 and 1 for the sentence as a whole.

This measure could yield both absolute and graded judgements: the absolute distinction is based on whether or not any constraints are violated, the graded measure is based on speed, so that a grammatical but hard-to-parse sentence (e.g. a garden-path) would be rated the same as a truly ungrammatical sentence, which is exactly what Warner and Glass (1987) found (see Chapter 4, §3.1). The relative leniency for ungrammaticality in spoken, as opposed to written, language could result from an across-the-board reduction in the time cost of traversing paths (including ones whose constraints are violated), in order to keep up with the real-time demands of continuous speech. Situationally-related context could speed up parsing by decreasing the time required to access meanings and other properties of relevant words; in the same way, a processing advantage is predicted for frequently- versus infrequently-occurring words. The relative strengths of parse paths should also be reflected in corrections, i.e. how you change a bad sentence to make it good: presumably, whatever was implicated in the constraint that got relaxed will be changed, since it is the only known locus of error. If you were to make a correction somewhere else, you could not be assured that it would be sufficient. Of course, we have seen that graded judgements can arise on a wide variety of nonlinguistic tasks as well; under our view, this occurs because race-based implementation is a feature of cognitive processing in general. Now I would like to conclude this section by suggesting that the same competition principle can be applied at a macro level as well.

Consider the situation where an ungrammatical but comprehensible sentence is encountered, e.g. *I just bought a CD for me.* Different knowledge components contribute to different views of this sentence: according to linguistic competence, the sentence is ill-formed because it contains a bound pronoun; our knowledge of the world, and specifically the knowledge that Boris just walked in the door with a CD in hand (still in its factory plastic wrapper), allows us to interpret the sentence without any difficulty. Which component will be allowed to determine our reaction to the sentence—to ignore the error and take up the conversation, to make a mental note that Boris does not speak perfect English, to tell him he has made a mistake and see if he can discover it, or to blurt out the correct version of the sentence? This decision can be seen as the result of a competition, where the "speed" of each processing module is determined by the demands of the situation. Thus, teaching an ESL class primes grammatical competence and correction strategies, everyday conversation favours parsing, the current situation activates relevant knowledge of the world, seeing oneself in a mirror might strengthen the communicative over the structurally-based strategies, etc. Under this interpretation, the "control strategies" themselves might not exist as strategies per se, but as the by-products of spreading

activation and race-based competition. Of course, for people we know, this choice may have been made for good when we first got to know them.

3. Methodological Proposals

3.1 Materials

There are basic precautions that could easily be taken in preparing materials for the elicitation of grammaticality judgements in order to avoid certain obvious kinds of bias.[5] One potential confound is the order in which sentences are presented to subjects: it has been shown experimentally (e.g. by Greenbaum (1973)) that sentences will be rated differently depending on their order of presentation. A simple way to factor this effect out of results is to counterbalance orders across different subjects, thus controlling for nervousness at the beginning of the session, fatigue at the end, practice effects, the influence of surrounding test items, and any other serial position effects. (Of course, this requires that one consult more than one subject.) A second kind of bias is introduced if there are substantially more grammatical sentences in the test materials than ungrammatical sentences or vice versa: subjects will tend to get into a "yea-saying" or "nay-saying" mode or come to expect deviance; thus, the numbers should be kept roughly equal. (Since we presumably do not know the outcome of all the judgements in advance, we cannot guarantee the exact proportions.) A third factor in our list of potential confounds in stimulus materials is the semantic content of the lexical items in the sentence.[6] As mentioned in Chapter 4, §3.4, it is simply not true that people will rate all structurally identical sentences equally grammatical; for example, Levelt et al. (1977) found that different ratings could be induced by varying the imagery content of a sentence, i.e. the degree to which it represented an imaginable or concrete situation. With a good understanding of such a factor, one can reduce its effect by avoiding sentences at the extremes of imagistic content,[7] and using several different exemplar sentences with the structure in

---

5 Several of the suggestions in §3 have been synthesized from Birdsong 1989, Ray & Ravizza 1988, and Snow 1975.

6 Obviously some semantic features of words are directly relevant to grammaticality; here, as in Chapter 4, §3.2, we are concerned with properties not generally considered grammatically relevant.

7 On the other hand, Birdsong (1989) proposes that only high-imagery content words should be used, so that all subjects can see the sentences as potentially referential and meaningful. Whether this is a good idea or not should probably be determined on the basis of experiments comparing the two methodologies; to my knowledge Birdsong's proposal has not been used yet.

question across subjects. That is, the lexical content of the sentences should be varied to guard against the influence of imagery, and any other potential biases of lexical items, such as word length, frequency, and semantic peculiarities. In light of the findings by Hill (1961), it might also be best to inform subjects that only common words will be used, or that if they are not sure about the status of a word, they should ask the experimenter; this would circumvent the possibility of subjects interpreting *of* as an unfamiliar noun, for instance.

More controversial than any of these issues in preparing materials is the surrounding contextual material, of all the various types: we have all had the experience of thinking at first that a sentence is totally ungrammatical, only to have someone suggest a real-world situation where it is quite plausible. As discussed in Chapter 4, §3.1, there are numerous ways that context can influence grammaticality, from bringing out rare word meanings to priming certain parsing procedures.[8] There is certainly no universally correct answer for what sort of context, if any, is suitable for particular elicitation purposes, but it is a variable that cannot be ignored: ratings of sentences in context cannot be compared with those made in isolation, for example. The consensus among the authors surveyed in the present work seems to be that a supporting pragmatically-related context should always be provided, unless that would somehow defeat the purpose of the experiment. Since only structural well-formedness is at issue, not pragmatic appropriateness, then if there exists a situation where the sentence would be appropriate, why should we not lead the subject to that situation? Furthermore, we will reduce between-subject variability by not leaving subjects to their own devices in imagining situations where the sentence might occur, which many researchers claim would otherwise be a major part of the judgement process.

Depending on the purpose of the experiment, one might wish to avoid choosing sentences whose rating is likely to be confounded by parsing difficulty: for instance, the garden paths studied by Warner and Glass (1987) showed extreme parsing difficulty; since these researchers were interested in the parsing process, rather than grammaticality per se, these were sensible choices, but if one wishes to know whether a sentence is accepted by the grammar, it does not make sense to confuse that with low parseability. Of course, the distinction might not always be obvious ahead of time, but one could call on a

---

8 For some striking demonstrations of how apparent word salad can be made plausible by context, see Hill 1961, especially fn. 4.

pilot study with a post-test questionnaire, where the intended interpretation of a sentence that was judged ungrammatical is stated, and the subject is again asked whether, under this reading, the sentence is still bad.

If one wishes to detect very small differences between sentences, then it is crucial that they be matched as closely as possible on as many features as possible, including semantic plausibility (Carden & Dieterich 1981); that is, they should be minimal pairs at the sentence level. When the relative grammaticality of two or more related forms is at issue, it is best to allow subjects to see them side-by-side and draw their attention to the comparison; the order among the related sentences, and the order among the *sets* of such sentences, should still of course be counterbalanced. Judgement tests are likely to give misleading results if the sentences used contain features that are unrepresentative of the whole range of sentences to which the results should generalize. Thus, Levelt (1974, vol. 3) pleads for the avoidance of additional unnaturalness that has nothing to do with the crucial issue at stake. He cites numerous cases where this seemingly obvious admonition has been violated, e.g. by what he views as unnecessary loading of short term memory (*That Tom's told everyone that he's staying proves that it's true that he's thinking that it would be a good idea for him to show that he likes it here*) or by the distracting semantic load resulting from an unusual situation (*I dreamed that I was a proton and fell in love with a shapely green-and-orange striped electron*) (Levelt 1972). Again, to guard against unintentional distractions of this type, multiple sentence frames should be built around the same crucial construction wherever possible.

3.2 Procedure

Having minimized potential confounds in the stimulus materials, the next logical step is to remove confounds from the process of gathering judgements. The first issue is the selection of subjects, perhaps the worst offense with regard to experimental method in linguistics to date. Here we would implore that these must be people with no linguistic training. 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. This is almost never done by theoretical linguists. (Bolinger (1968) and Greenbaum (1976a) also make this point.) They first consult their own intuitions (one cannot find a more biased subject than the investigator), then their colleagues in the next office (almost as biased), and if they are really ambitious, perhaps a couple of their students (not exactly objective either). While striking differences between linguists and nonlinguists have not been convincingly

demonstrated empirically due to poor experimental designs (see Chapter 3, §4.1), we have enough reasons to *expect* them to be different that linguists simply ought to be excluded. Also, the small samples of linguists that are usually available are bound to lead to unreliabilities (Bradac et al. 1980). Nonetheless, linguists continue to insist that the ease of obtaining data is a reason for preferring oneself as a subject, ignoring the inferior quality of the data so obtained. (Newmeyer (1983, p. 50) claims this justification is "uncontroversial"!) If linguists wish to live up to scientific standards of data validity, it is time for them to abandon the convenient fiction that data is never further away than their own minds.

Subjects must be sufficient in number in order for the assumptions of the required statistical tests to be met and to avoid distorting the results with atypical speakers. If there is any reason to suspect regional variation on the issue at hand, an effort should be made to find speakers of various dialects (this would usually be a good idea in any case). Snow (1975) suggests that subjects be pre-tested and screened for their ability to judge reliably, but I wonder whether such a procedure would systematically exclude a relevant class of judgements; a similar objection could be raised against the exclusive use of "expert" language users, e.g. prominent authors. Judgement tests should be carried out in a controlled setting, to decrease the chances that subjects will be "inebriated, inattentive, mendacious or whimsical" (Grandy 1981); the pub where everyone goes at the end of a conference is probably not an ideal locale. Individual differences on potentially relevant factors such as age, sex, education, etc. should at least be noted on a personal questionnaire so that variability attributable to them can be examined in the analysis; if multiple conditions are being used (e.g. with versus without context), random assignment of subjects or counterbalancing on these factors is important.

The next problem linguists have is with the instructions to informants: what exactly should one ask them to do? No two studies seem to agree. Certainly we have shown that one cannot hope for the terms "grammatical" or "acceptable" to have their intended meanings for naïve subjects; Chaudron (1983) and many others point out the potentially nonunitary measure that would result. Experimenters must put considerable effort into designing an explanation for them on how they want them to make their judgements, at least until such time as the field can agree on a standard set of instructions.[9]

---

9 Peter Reich (personal communication) suggests that standardized instructions are unlikely ever to be adopted, given that there are apparently very few such cases in psychology, which is generally much more concerned with procedural matters than linguistics. I would argue that, unlike in psychology, large numbers of linguists are interested in asking exactly the same questions about their stimuli

This will require linguists to make explicit exactly what counts toward grammaticality, which perhaps can only be done with reference to particular types of theoretical issues being investigated. For this reason, we cannot propose a generally applicable set of instructions here, but we can suggest how to make them effective. Most experiments seemed to have erred on the side of describing the task too briefly and vaguely. Instructions should be specific, mention possible reasons why a sentence *should* be considered bad, and also mention potential reasons that should *not* come into play. Give examples of sentences that you consider unequivocally good and unequivocally bad (but that do not contain the construction you wish to test) and explain why the good one is good, despite some irrelevant properties (e.g. meaninglessness) and why the bad one is bad, despite other irrelevant properties (e.g. interpretability). The examples should cover a wide enough range to avoid problems like the one Birdsong (1989) encountered: he reports that his subjects claimed a stimulus item was not a sentence because it was a question! Run some practice trials where the informants think aloud during the judgement process, so that you can point out if they are using inappropriate criteria. It is important to keep the statement of instructions itself down to a reasonable length; otherwise, it "becomes an essay on linguistics that only a sophisticated informant can understand, and only an unusually patient one will read" (Carden 1970a, p. 296).

If the field had a standard set of instructions, then at the very least it would ensure that everyone was testing the same thing, even though considerable refinement would be required to make it the thing we are interested in. Results could be meaningfully compared across experiments, which is currently not possible. This must be possible, however, if there is any chance of making linguistics a truly objective endeavour: if the only people we can gather data from are other linguists, all hope is lost. (See Newmeyer 1983, p. 61, for the view that this state of affairs is unlikely to change anytime soon.) An alternative suggestion, made somewhat tongue-in-cheek by Hirst (1981, p. 101), is the establishment of a central sentence-testing service to which linguists would send their crucial sentences (and some money) and get back in the mail a standardized set of experimentally-elicited ratings. This would eliminate the time and effort that would be required to set up appropriate testing facilities in each department, ensure consistent procedures, and reduce the overhead expense by dividing it among a larger user population.

(sentences); at the very least, we can hope to make widely know what sorts of directives do and do not work.

Having dealt with how the concept of grammaticality is to be conveyed, we must now consider what to ask subjects to do with it. The biggest issue here is whether to use absolute ratings, and if so on what scale, or relative rankings, and if so on how many sentences at a time; if rankings are used, should we ask for a grammaticality threshold to be drawn, as some studies have done? The issues surrounding this choice were discussed in detail in Chapter 2, §3, and will not be repeated here; the goals of the experiment will obvious come into this decision. All other things being equal, most researchers advocate comparative judgements on the basis of their higher reliability. If a rating scale is used, I would argue it should be a balanced one; that is, unlike Nagata's scales (see Chapter 4, §2.3) where 1= grammatical, and 2–*n* are degrees of badness, it should range evenly from good to bad, with middling being in the middle. If verbal descriptions of the various positions on the scale are given, some care is called for. Greenbaum and Quirk (1970) advise against calling a middle rating "not sure," for instance, because this carries the potentially negative connotation that the subject is unable to make a decision, rather than labelling the sentence as intermediate in grammaticality; in fact, both answers should be available, so that cases where the subject truly *is* unsure can be treated separately. They also suggest against calling the middle category "marginal or dubious," as they report that Quirk and Svartvik (1966) did, because this terminology sounds too technical. In fact, I believe the use of more than one rating criterion should be seriously considered. If one gives subjects a chance to rate grammaticality, stylistic felicity, likelihood of occurrence in conversation, and their own (un)certainty separately, this should reduce the chances that the latter factors will play a role in subjects' ratings of the first; people seem to want to express their feelings about these other matters, so it is best to give them the opportunity explicitly. The effectiveness of the rating scale(s) also depends on warm-up trials that encompass a representative range of sentences. At this stage (unlike the detailed examples advocated earlier) they probably *should* include sentences of the type that will occur in the experimental trials, otherwise there is a risk that novel stimuli will show a primacy effect. Using relevant sentences in practice trials is not a problem as long as experimenters do not bias the subjects with their own opinions of these sentences.

Carden (1970a) points out a problem with eliciting grammaticality judgements only on a questionnaire, rather than in an interview: "You often must focus on a particular reading or construction. It is of no value to know that informant X considers a sentence ungrammatical if you do not know that his reason for rejecting it is unrelated to the construction you are studying" (p. 296). He thus argues for greater use of interviews, an issue to which we return below. We agree that asking for some sort of explanation is cru-

cial to knowing that a sentence was rejected for the "right" reasons in many cases. However, there are ways of getting at which feature(s) of a sentence caused a subject to reject it, even with a written questionnaire. One can ask subjects to indicate the location of any errors they perceived, to explain why they are incorrect, and/or to correct them. If this is done, however, it is crucial to balance these tasks with corresponding ones to be performed if the sentence is good, otherwise subjects might be biased towards good ratings just to avoid the extra work. (Snow's (1975) and Hakes's (1980) studies had this problem.) As mentioned in Chapter 2, §2, the most obvious candidate task is paraphrase: ask the subject to re-write the sentences in a different way while preserving the meaning. This can additionally tell us how the sentence was interpreted, which could be useful information. (Of course, whether paraphrasing is as demanding as explaining errors is hard to assess, but it is a step in the right direction.) A similar kind of bias to that just discussed was exhibited by Rose's (1973) study, described in Chapter 3, §4.1. His materials contained equal numbers of good and bad sentences (according to the sources they were drawn from), but subjects were divided by being asked to mark either the good or the bad sentences, while leaving the other kind unmarked. The groups differed significantly in the number of sentences accepted, with each group leaving more than half the sentences unmarked. This seems to be another instance of bias towards minimal action; to avoid it, subjects must be given the same amount of work to do no matter how they rate a sentence.

The issue of by what means judgements are to be elicited is another important question of methodology. The most detailed examination of this problem is found in Carden 1976. His main concern was to find ways of increasing the *reliability* of elicited judgements, that is, the extent to which later ones by the same informants or separate ones from other informants of the same speech community will be consistent. Carden concurs with our own position that the major difficulties lie in explaining the task to naïve subjects, particularly in non-interactive forms of data elicitation, such as (forced-choice) questionnaires. At the opposite end of the interactivity scale is the open-ended interview, which is rarely used systematically by linguists, but is claimed to yield considerably cleaner data than questionnaires. Carden cites two examples of interview studies that were later replicated with questionnaires; in both cases, the interview results showed clear patterns, whereas virtually no systematic results could be found in the questionnaire results. A plausible explanation for this is that interviews provide more opportunity for the experimenter's bias to influence the subjects (Newmeyer 1983), but Carden argues that there is evidence to suggest interviews also allow real improvement in data quality, be-

cause the task can be explained in more detail, subjects' questions can be answered and misunderstandings set straight, etc. In follow-up interviews to the questionnaires, he found that much of the noise in the data was due to irrelevant readings of stimulus sentences, and in cases where statistical analysis was available, it did show significant patterns of the same sort found in interviews, although they were much less obvious from causal inspection of the unanalyzed data. Still, Carden acknowledges that until potential bias effects are studied in more detail, interviews remain suspect. In fact, he paints a gloomy overall picture that may not have improved much over the intervening years:

> The linguist's own intuitions are plainly untrustworthy. Direct observation of performance, while potentially important as a means of validating other methodologies is impractical as a primary technique. Performance tasks seem to be even less reliable than evaluation [judgement] tasks, and are difficult to adapt to the more interesting syntactic problems. Forced-choice questionnaires are also difficult to construct, and have at best marginal reliability and very noisy data. Open-ended interviews seem to produce clear results, but are very time-consuming and may have bias problems. (p. 103)

We would suggest following the standard practice in social science of using interviews in the preliminary phases of an investigation only: once potential ambiguities, misunderstandings, etc., have been discovered, the materials can be adjusted to deal with them and controlled experiments run, so that statistical analysis can legitimately be applied to the results. Another standard technique that could be useful in preliminary investigations is the focus group (Graeme Hirst, personal communication): informants could discuss test sentences among themselves, employ them in different contexts, point out problematic features, etc., while the experimenter observed surreptitiously; in this procedure, grammaticality judgement would be a group, as opposed to an individual, activity.

Psychology has identified several kinds of experimenter effects, induced by the behaviour of the experimenter, which can bias results (see also Labov 1975). In the linguistic case, there is great potential for the investigator to influence a subject's judgements, even if the experiment is not an interview per se: by demonstrating the procedure using sample sentences that are related to the test materials; by the idiolect of their own speech, which may be different from the subject's; by subtleties of their interaction with the subject, e.g. how they respond when the subject gives a judgement they do not expect; etc. Heringer (1970) raised many of these issues over 20 years ago in the following widely-cited passage, which seems to have been ignored by most theoreticians: "In the casual interaction between linguist and informant, there are many opportunities for self-fulfilling prophesies to take effect, both ones conditioned by theoretical position and also

ones conditioned by the linguist's own idiolect. This could occur even without the conscious knowledge of the linguist, especially if stress and intonation are not controlled" (p. 294; see also Bradac et al. 1980). These dangers are fairly easily removed, if one is aware of them, by not using the investigator as an experimenter and by scrutinizing the instruction and elicitation phases for potential influence. Carden (1970a) warns us that the linguist could also bias results by inconsistent coding of informants' responses,[10] so this should be done by disinterested parties as well. Typically, each set of responses would be coded independently by two judges, and consistency between them should be measured and reported.

Sentence judging is also particularly susceptible to what are known as maturation effects (also sometimes called order effects). These include the results of being asked for too many judgements at one sitting: boredom, frustration, and fatigue, which lead to inaccurate responses because the subject stops caring about the outcome. Satiation, whereby symbols lose their meaning after repeated exposure, strips subjects of their intuitions altogether. Short sessions and varied stimulus materials are the obvious remedies. Closely related to those just mentioned are testing effects. These include practice or training, whereby the subject gains skill in the judgement task over time, meaning that early results are not comparable with later ones; these can be controlled for across subjects by counterbalancing orders as discussed in §3.1, but they will still distort within-subject comparisons. There is also potential for the subject to become aware of what particular issues the experimenter is interested in, which can cause the crucial sentences to be treated specially. For instance, subjects might identify parasitic gap constructions as the items of interest and decide that they ought to rate every one of these identically, regardless of their actual intuitions. (In Chapter 1, §3.4, we suggested that linguists regularly do this.) This should be avoided by using enough filler or distractor sentences, i.e. ones that are unrelated to the crucial construction. These will also serve as anchors, to remind subjects of the range of potential goodness/badness; otherwise, after looking at marginal sentences for a long time, subjects might start spreading their ratings out further on the scale. A more bizarre variable that can apparently affect subjects' perception of the purpose of the study is experimental setting (Greenbaum & Quirk 1970): performance tests were conducted on two groups of college students, one in a lecture hall with white-coated strangers as experimenters, the other in the investigators' own English department with

[10] He proceeds with the non-argument that since his informants have often disproved his theories, he must be fairly successful in avoiding conveying bias in interviews.

familiar profs as experimenters. The test itself was tape-recorded, so there could be no bias in the stimulus materials themselves, and yet the authors found significant differences between the two groups in the number of relevant non-compliances (RNCs). Subsequent interviews showed they had put the two groups of subjects into different mental sets, thinking the test had a linguistic versus a psychological purpose: "We have seen that opinions of the test's purpose can importantly affect RNC scores and that the opinions themselves can be easily affected to a significant degree by small changes in test (and pre-test) conditions" (p. 58). A final procedural consideration is the mode of presentation of the sentences. It is common knowledge that spoken and written language have vastly differing norms (which may be attributable to the dimensions of interactivity and/or permanence of the communicative medium), so we should expect that judgements of sentences in the two modalities will reflect these differences (as discussed in Chapter 4, §2.6); the two cannot be directly compared. If oral presentation is used, sentences should be read by a disinterested person, not the linguist, and tape recorded to insure uniformity of intonation, and to edit out any speech errors; trained announcers serve this purpose well. If instructions are orally presented, these too should be recorded to insure uniformity.

3.3 Analysis and Interpretation of Results

Levelt (1974) has complained that linguistics lacks a *theory of interpretation*, that is, a standard specification of how data are considered to bear on the theory. We conclude this section with some specific suggestions about the interpretation of grammaticality judgement data. The first seems almost too obvious to mention, and yet linguists consistently ignore it: without performing statistical tests of significance, we cannot know whether trends in our data are likely due to chance or to actual facts about grammars (or some other part of the mind), unless we truly have "sledgehammer" results. The more levels of grammaticality we try to distinguish, the less unanimity we will find, and the more we will need to rely on statistics. Another problem of statistical ignorance, originally pointed out by Clark (1973), is the "language-as-fixed-effect fallacy." Clark's point is that even when we find statistically significant results on a grammaticality test, we cannot necessarily generalize from the actual sentences used in the study to all sentences of the same form. The statistical analysis must treat the materials as a random rather than a fixed factor, which results in more stringent criteria for significance, but many studies have failed to do this. Additionally, the implications of the way the stimuli were gathered must be considered; it might not be crucial to do true random sampling of sentences (it is

not entirely clear what that would mean), but, as suggested in §3.1, experimenters must consider the extent to which their materials are representative of the "population" of sentences to which they would like their results to generalize.[11] (Clark also points out other common statistical problems with psycholinguistic experimentation.)

Simple statistical comparison of judgement ratings is not the only type of analysis that can be used to learn about grammaticality. Bradac et al. (1980) were motivated by the belief that looking at a single measure, viz. grammaticality/acceptability, might conceal the "rich, multi-dimensional nature of language judgments"; thus, their technique was geared to multiple factor analysis. Their stimuli were broken down similarly to those of Maclay and Sleator (1960), by the various types of errors in sentences, including "school grammar" errors, typical foreign learner errors, and sentences that are supposedly grammatical although unacceptable. They asked 13 questions about each sentence that were answered on 7-point scales, including "Is this grammatical?", "Is this English?", "Is this clear?", "Is the speaker educated?", etc. As usual, none of these terms were explained to the subjects, so it should not surprise us to see the authors conclude that "persons may be quite sensitive to the precise way in which such questions are asked," especially since half their subjects were linguists, the other half not. They also recorded various other attributes of the subjects. The experimental procedure was very carefully controlled: for instance, ratings on the various scales were elicited in different orders on different trials; the positions of the extremes of the scales were reversed for half the trials, etc. The problem with this and most other multivariate studies is that it is very hard to draw any firm conclusions from them; what one is left with is a bunch of correlations between variables. For instance, the 13 rating scales were factored into four major dimensions, one of which was interpreted as grammaticality/acceptability.[12] Among the personal attributes, linguistic training or lack of it was a systematic source of variation in judgements, and so were the number of sisters the subject had and the subject's birth order, whereas sex and handedness accounted for very little variation. Still, this type of analysis can yield useful insights that might otherwise be overlooked.

---

11 An example of a study where the experimenters found that two stimulus sets did not seem to represent the same population is the second study reported by Bradac et al. (1980).

12 The scales that comprised this dimension were "is grammatical/is ungrammatical," "is acceptable/is unacceptable" and "is correct/is incorrect."

One principle of interpretation that many researchers in this area have stressed (e.g. Chaudron (1983)) is that *any* conclusions on the basis of a single kind of experimental test is dubious; wherever possible we should appeal to "cross-methodology validation" (Carden & Dieterich 1981). Even in cases where one kind of task (e.g. judgements) yields reliable results, their validity as indicators of linguistic competence is suspect because of the numerous potential intervening factors, as discussed in §2, but if the same results show up reliably across additional types of tasks, such as unwarned judgements, performance tasks like those used by Greenbaum and Quirk (1970), short-term memory measures, unintrusive reaction measures such as ERPs, sentence-completion tasks, or naturalistic observation of speech and writing, then the odds are much higher that the evidence does represent a convergence on fundamental underlying knowledge. In his review of early work, Carden (1976) concluded that such cross-task comparisons had yielded extremely low agreement to that point, suggesting that they were not measuring the same thing. On the other hand, Chaudron's (1983) review cites the following studies as showing that judgements *are* correlated with with other performance measures: Coleman 1965, Danks 1969, Moore 1972.

The final and perhaps most troubling problem we will comment on in interpreting grammaticality judgements is what to make of inconsistencies, be they changes in one subject's judgements over time or disagreements between subjects.[13] Generative linguists have often suggested that the between-speaker differences that are found represent minor disagreements on fringe data, but that the major substance of the grammar is the same for everyone, being a function of, say, UG plus parameter settings.[14] Others have disagreed, e.g. Grandy (1981), Levelt (1972), Carden and Dieterich (1981): "data disagreements, regrettably but perhaps not surprisingly, tend to center on theoretically cru-

---

13 In Chapter 2, §3, we raised the issue of what should count as an inconsistency. Different experimenters have used different criteria for deciding when two judgements are consistent: some require them to be identical, others allow a one-point variation on a scale of three or four values, etc. Birdsong (1989) points out that a yea-saying bias, as found by Mohan (1977, reported in Chapter 2, §3), can artificially inflate consistency scores.

14 Newmeyer (1983) explicitly argues that the vast majority of alleged data disagreements in generative grammar are actually disagreements about the role of the theory, not judgements. My own experience has been that such a position is totally untenable, as exemplified by the cases discussed in Chapter 1, §3.2.

cial examples" (p. 584).[15] Under a principles-and-parameters approach, we might expect the "periphery," that part of the grammar *not* specified by UG but somehow learned, to vary with people's learning abilities and experience, but we would presumably not expect variation on matters directly in the scope of innate universals. While it is hard to find data bearing directly on this point, my suspicion is that it is untrue.[16] If we take Binding and Subjacency conditions as paradigmatic examples of the domain of UG, my own experience is that these are two of the areas of greatest variation. (Again, see the discussion of *that*-trace effects in Chapter 1, §3.2.) Here one really cannot argue that the disagreements involve unimportant or fringe sentences: as rare as they may be in everyday speech, if they are governed by innate principles then this degree of variation is unexpected. The standard appeal to "performance factors" is also unconvincing here, at least in the usual narrow sense, which typically boils down to memory limitations or analogy. While it is reasonable to suggest that people's ability to process multiply centre-embedded sentences could be a function of their short-term memory capacity independent of their grammar, the same does not ring true for Binding and Subjacency.

Are we forced to conclude, then, that UG exhibits individual differences, that we are not all born with identical principles and parameters? This is certainly a possibility, and would not be a particularly surprising result—in general, people do exhibit individual differences on many, perhaps all, innately-specified behaviours, while sharing the gross features. Peter Reich (personal communication) has suggested that one could test this hypothesis on bilinguals by examining whether their judgements differ in the same ways in both languages. If not, one might suspect (learned) differences in properties of lexical items instead. A third possible explanation is that the differences lie not in the grammatical competence but in some other phase or aspect of the judgement process, i.e. they are attributable to metalinguistic performance in our broad sense. Most theoretical linguists have effectively given up on accounting for judgement differences and resorted to describing a single "dialect" or "idiolect," typically their own, when faced with data dis-

---

15 Newmeyer (1983) correctly points out that the particular example that Carden and Dieterich use to exemplify the situation (the debate between Chomsky on the one hand and Katz and Postal on the other over the interaction between passivization and quantifier scope) is probably not a true data disagreement.

16 This is not to deny that there is an (arguably very small) "core" of simple sentences that all speakers of a language will agree are grammatical. But this set is not identical to the core in Chomsky's technical sense; far from it.

agreements.[17] But linguists must also take responsibility for the range of variation that is actually found.

An explanation in terms of extra-grammatical factors seems even more likely in the case of changes in one person's judgements from one elicitation to the next. (Either that or, as Snow (1975) suggests, poor experimental design could be to blame.) While grammars certainly do change in some regards over the course of a lifetime, most linguists would probably not want to say this happens on a day-to-day basis in adulthood.[18] Another alternative that has been bandied about occasionally is that grammars are probabilistically defined, so that some sentence will be judged good 90% of the time, bad the other 10%, based on random variation in neural signal strength or some such factor. (In fact, the race-based approach predicts such a pattern of events.) The problem with this analysis is that it denies that there are any systematic causes behind the variation we find. If instead we start with the assumption that it has a cause within the system of judgement performance, then as we understand more about that process we might eventually be in a position to say precisely what governs variation over time and predict it as a function of other cognitive and situational variables. Only after we have exhausted the search for such an explanation should we resort to random probabilities.

Labov (1975) has proposed a widely-cited set of working principles for dealing with variation in grammaticality judgements and interpreting their relationship to the grammar:

> I. The Consensus Principle: if there is no reason to think otherwise, assume that the judgments of any native speaker are characteristic of all speakers of the language.
>
> II. The Experimenter Principle: if there is any disagreement on introspective judgments, the judgments of those who are familiar with the theoretical issues may not be counted as evidence.
>
> III. The Clear Case Principle: disputed judgments should be shown to include at least one consistent pattern in the speech community or be abandoned. If differing judgments are said to represent different dialects, enough in-

---

17 Ross has paraphrased the standard research directive as "Write a grammar of what you find in your heart" (MIT class lectures, 1966–7, cited in Carden & Dieterich 1981).

18 An exception to this is Peter Reich, who views each instance of language comprehension as a potential instance of learning, since under his view the grammar simply *is* the performance mechanism, which is constantly adjusting itself in response to input signals, in a manner similar to connectionist learning algorithms.

> vestigation of each dialect should be carried out to show that each judgment is a clear case in that dialect. (p. 31)
>
> IV. The Principle of Validity: when the use of language is shown to be more consistent than introspective judgments, a valid description of the language will agree with that use rather than introspections. (p. 40)

For the most part, these suggestions strike me as quite reasonable, although a caution is in order. Principle I is intended to allow the field to continue without having to resort to experimental verification of sentences whose (un)grammatical status no one has ever questioned. Principle II is intended to guard against experimenter effects; I would suggest strengthening it so that the investigating linguist's own intuitions are *never* counted as evidence, even if his or her data have not been disputed. Principle IV jibes well with our comments in Chapter 2 concerning the potential for differences between use and intuitions; unfortunately, there is still no way of knowing when primary data from linguistic use need to be sought out (since such relevant data are often not immediately available), and no obvious procedure for determining whether they are more consistent than judgement data. But the major problem comes with principle III: its wording (and that of principle I) indicates that Labov is interested in accounting for the grammar of *groups* of speakers (as inferred also by Newmeyer (1983)), but we have argued that it is entirely possible for *individuals* to have unique grammars, so that discarding judgements that are not shared by other speakers could involve throwing away real data. Certainly, the more speakers we can find who share a set of intuitions, the more confident we can be in the legitimacy of those intuitions, but at a certain point we will have to hope that our methods have removed as many confounds as possible and treat the resulting data as significant, even if it applies only to a single speaker. While there may be little interest in studying individual idiolects for their own sake, the *range* of variation that is found is crucial to the construction of theories.

4. Conclusion

In this chapter we have presented two major proposals that assemble the information gleaned from the first four chapters of this work. The first was an initial attempt at a model of mental components of metalinguistic activity, with a focus on grammaticality judgements. There is clearly much more work that could be done in this vein: first of all, we might now go on to suggest specific experiments that could clarify aspects of the model or show where changes are required. Second, we could derive new empirical predictions regarding how certain effects should manifest themselves in tasks that implicate

certain components of the mental structure, and run these tasks experimentally to see whether the predictions are borne out. Third, we might consider whether there is something to be gained from implementing a computer simulation of the model; since it is highly parallel, and relies on very many micro-computations, simulation could lead to better understanding and refinement, as it has in connectionism and race-based parsing. The second major proposal in this chapter took the form of methodological guidelines for eliciting grammaticality judgements. We did not go so far as to propose a particular experimental design, since this must vary with the specific purpose of the experiment, but if even some of our suggestions are followed, significant strides towards a solid scientific foundation for linguistic research will have been made. The question is whether theoretical linguists are likely to heed such advice. I would like to suggest that part of the reason for linguists' lackadaisical attitude in this regard is not so much that they believe the data are clear-cut, but that there is little motivation for putting effort into a systematic approach because, unlike in most of the social sciences, there is no standard publication format requiring authors to describe how their data were gathered. (Grandy (1981) makes a similar point, and suggests other possible reasons why the deplorable state of lack of rigour continues.) Also, since they typically have no training in experimental design, they do not appreciate how useful and important it is. On this question, we can do little more than keep our fingers crossed; it does seem, based on my assessment of the literature, that more and more linguists are coming around.

Chapter 6

Looking Back and Looking Ahead

Recent trends in linguistic research have placed increasing dependence on relatively subtle intuitions . . . whose psychological status is extremely unclear. Since there are many sources for intuitional judgements other than grammaticality, and since grammaticality judgements themselves can be influenced by context, subtle intuitions are not to be trusted until we understand the nature of their interaction with factors that are irrelevant to grammaticality. If we depend too much on such intuitions without exploring their nature, linguistic research will perpetuate the defects of introspective mentalism as well as its virtues.

(Bever 1970b)

1. Introduction

The epigraph from Bever above concurs very well with our own findings about grammaticality judgements in the preceding five chapters. By way of a response, it seems fair to say that the field has begun taking steps to "explore their nature," and I hope that the present work has made its own contribution to that exploration. In this final chapter we will concentrate mostly on what lies ahead in this endeavour.

We will not attempt to summarize the discussion to this point in any detail, but will very briefly review the structure of the argumentation and illustrate that a major achievement has been to provide substantive support for the views of grammaticality judgements that have been expressed succinctly and eloquently by previous researchers in this area. It may be hoped that their observations will carry more weight with the underpinnings of the extensive experimental and theoretical literature that the present work has assembled.

In Chapter 1 we reviewed some of the history of how the concept of grammaticality has evolved since the 1950s, various opinions on its empirical status, and how it is used today among theoretical linguists. On this basis we argued that theory is no longer being based on clear cases, and that therefore detailed study of the judgement process was required to establish how to deal with unclear cases. Botha (1973) extends the argument

even further: "consider the status of the so-called clear cases of linguistic intuitions. Today, it can be seen that native speakers may make, concerning a particular linguistic property of some sentence, judgements which are at once, clear, decisive, and consistent without there necessarily being genuine linguistic intuitions at the basis of these judgements" (p. 205). We also pointed out that concern with these problems has been sorely lacking to date in linguistic work: as Pullum somewhat cynically puts it, "The median number of speakers on whom the entire corpus of examples in an English syntax paper is checked before publication, including its author, is zero" (Pullum 1987, p. 453).

Chapter 2 was devoted to pursuing the suggestion of Chapter 1 that grammaticality judgements be studied as an instance of (meta)linguistic performance: as part of a larger family of such tasks, as an instance of graded behaviour, as an instance of introspective behaviour, and as just one more source of evidence about grammars. By this point, it was already apparent that "in many ways, intuition is less regular and more difficult to interpret than speech" (Labov 1972a, p. 199). We then made the specific proposal that the sources of the perturbations in grammaticality judgements exist independent of the language faculties of the brain. In retrospect, that position has probably turned out to be too strong: some of them might be attributable specifically to the parser, for instance. Nonetheless, the more phenomena we can reduce to language-independent sources the better, by Occam's razor, so we maintain that one should always seeks evidence for this position first.

Chapters 3 and 4 were devoted to detailed examinations of the range of causes for variability in judgements: Chapter 3 was concerned with the degree of variation between subjects and its attribution either to inherent attributes or to life experiences. Chapter 4 examined external factors, which were broken down by being (mostly extra-grammatical) features of the very sentences being judged, or features of the procedure used to elicit judgements. It is clear that in neither case do we have a full understanding of the way these factors work, so that Birdsong's plea still stands: "thorough study of the psychological and epistemological intricacies of metalinguistic performance is necessary if we are to achieve an understanding of the linguistic knowledge it is often thought to reflect" (Birdsong 1989, p. 49).

Finally, Chapter 5 was an attempt to integrate these findings in terms of a high-level view of the inferred mental structures and a proposed methodology for more rigorous data collection among linguists. The componential view of the judgement process as involving many more pieces than language use lent credence to another of Birdsong's

statements: "the hypocrisy of rejecting linguistic performance data as too noisy to study, while embracing metalinguistic performance data as proper input to theory, should be apparent to any thoughtful linguist" (p. 72); if it was not before, it should be now!

The format of the remainder of this chapter is very straightforward: in §2 we consider the sorts of research that should naturally follow most directly from the present paper, including both experimental and theoretical undertakings. I hope to have the opportunity to pursue them in the future. We conclude in §3 with some speculation about the future in the field of linguistics, specifically the role that grammaticality judgements are likely to play down the road, and the chances that attitudes toward their collection and application will change.

2. Directions for Future Research

As acknowledged in Chapter 1, §5, what we have presented here is far from a complete picture of the state of the art in studying grammaticality judgements. There is great potential for elucidating many of the issues we have confronted by considering kinds of data that have been excluded here. For example, experiments involving amnesics could clarify the memory mechanisms underlying structural priming effects, repetition effects, context, etc. Research on the development of metalinguistic skills in children should tell us more about their interdependence with primary language skills. Work with second-language learners should help to establish the relationship between intuitions and use as skill in the language increases, while avoiding some of the methodological problems involved in eliciting judgements from children; experiments with aphasics could serve a similar purpose. Finally, more about the process of linguistic judgement in general could be learned by more detailed work on the nature of lexical, phonological, semantic, and pragmatic judgements, in comparison with syntactic ones. The larger open question of the existence of linguistic competence and its role in language processing remains a major unresolved issue in the psychological investigation of language processing.

As for specific directions that would follow more directly from the present paper, many potentially informative experiments have been proposed in response to specific problems with published work; these will not be repeated here. One major area where we have barely scratched the surface is substantiating the hypothesis proposed in Chapter 2, §6, namely that we can find a nonlinguistic analog for every one of the perturbations that grammaticality judgements are subject to. Ideally, this would include finding indepen-

dent motivation (outside linguistic judgements) for the components of the model proposed in Chapter 5, §2, e.g. the general control strategies. The first major hurdle in this line of work is to find suitable domains of cognition in which to look for parallels to the manipulability of grammaticality judgements. By way of an example, in addition to possibilities discussed in Chapter 2, one intriguing area I have come across involves judgements in legal cases. Kaplan (1977) reports a number of phenomena that look promisingly parallel to what we have seen. For instance, it has been shown experimentally that jurors are influenced by factors totally irrelevant to the legal merits of a case in ways that depend on the nature of the crime: an attractive defendant will be judged less likely to be guilty of a burglary but more likely to be guilty in a confidence swindle. Personal traits such as race, sex and marital status have been shown to affect outcomes of cases even when jurors are explicitly instructed not to pay attention to them. Even when jurors are told that a certain variable is or is not statistically a predictor of guilt, they do not use this information in deciding how to treat the information in question. These findings have obvious parallels in the effects of irrelevant variables on grammaticality. A second major area that cries out for follow-up is the methodology of judgement elicitation itself. The logical next step would be to design and run a case-study experiment incorporating the proposals made in Chapter 5, §3, developing a specific set of instructions along the way. Such a study will undoubtedly point out problems with the proposals, suggest refinements, etc., and will allow the resulting data to be assessed for reliability, which could be compared with results from more casual data collection: to the extent that the data are more reliable, one of our goals will have been met.

3. The Future in Linguistics

A glance at the length of the reference list of this paper shows that many language researchers have concerned themselves with the problems that have been addressed here. Many of the experimental findings were published a number of years ago, but it seems that lately this activity is on the increase again, along with continued calls for greater use of formal experimentation for collection judgement data (e.g. Hirst 1981, pp. 100–101). Is all this work having any real effect on the way theoretical linguistics is carried out on a day-to-day basis? While the instances are still few and far between, I believe that issues in grammaticality judgement collection and interpretation are receiving greater attention. I would like to close by citing three examples of what I consider to be leading-edge work in this regard, studies that made appropriate use of judgement data within the framework of theoretical argumentation. The first, Grimshaw & Rosen 1990, has already been dis-

cussed in Chapter 1, §2: Grimshaw and Rosen eventually conclude that while children do not show perfect mastery of Binding Theory, they perform above chance, and treat violations of it differently from non-violations. The authors argue that inherent properties of the relevant constructions, as well as the experiments by which they are evaluated, conspire to worsen children's performance, especially as reflected in their apparent lesser mastery of Principle B versus Principle A.

The goal of our second exemplary paper, Carden & Dieterich 1981, was to establish the structural conditions on pronoun coreference. Carden and Dieterich were dealing with cases where a pronoun precedes the noun with which it is coreferent, e.g. (1) and (2), where co-subscripting indicates coreference:

(1) I knew him$_i$ when Harvey$_i$ was a little boy.

(2) We'll just have to fire him$_i$, whether McIntosh$_i$ likes it or not.

A handful of instances of these constructions had been found in texts, but proportionately very few compared to uncontroversial backwards coreference cases like (3):

(3) The boy who loves her$_i$ claims that Mary$_i$ is a genius.

The situation illustrates once again the problem with corpus data: "How do we interpret this data? Do we cheer because there *were* six examples, and conclude that Reinhart was right? Or do we boo because there were *only* six examples, as against hundreds of the uncontroversially good type? . . . We may have a good but (accidentally) rare construction; or we may have a bad construction occurring a few times because of errors" (Carden & Dieterich 1981, p. 591). They sorted out the status of sentences like (1) and (2) using an experiment that showed that these questionable forms were accepted no more often than an uncontroversially bad form. (In each case, only 1 of their 30 subjects accepted them.) The materials were constructed so that a preceding context sentence allowed a plausible reading where the crucial coreference relationship did *not* hold, as well as a reading where it *did* hold, so that subjects would not be forced by plausibility into accepting an ungrammatical structure. Further, they also tested the uncontroversially good and uncontroversially bad sentences preceded by the same context sentence, so that the results would be fully comparable.

Our third exemplary study also involved backwards coreference; it was conducted by Gerken and Bever (1986), who were apparently not aware of Carden and Dieterich's

work in this area. On the basis of inter-speaker differences in the interpretation of the same sorts of sentences, Gerken and Bever propose that linguistic universals, in particular Binding Theory, are not necessarily applied to complete sentence structures as given by linguistic competence, but rather are applied to the speaker's *perceived* structure as generated during sentence processing. They point out that for many sentences it is not necessary to compute the complete syntactic structure in order to extract the meaning, and suggest that this computation might therefore be delayed until after the initial parse, or never carried out at all. The specific contrast they were concerned with was this: Binding Theory dictates that there should be a strong contrast between VP-attached and S-attached subordinate clauses with regard to potential backwards coreference, so that (4), where the complement clause is under the VP, should be much worse than (5), where the adverbial clause is attached directly to the S node.

(4) The dog told him$_i$ that the horse$_i$ would fall.

(5) The dog hit him$_i$ while the horse$_i$ ate lunch.

However, Gerken and Bever's acceptability experiment failed to find any such difference. They argue that there are no surface cues for the difference between S-node versus VP-node attachment, so it is possible that the distinction is not made in on-line parsing structures. In fact, there is a tendency in English to segment sentences after a noun-verb-noun sequence, and those subjects who performed strong "perceptual closure" at this juncture (as revealed by another experiment) did not make the attachment distinction for pronouns, whereas those who made less use of the closure strategy did have the predicted contrast between (4) and (5). Those who exhibit strong closure do not have a VP node accessible for attachment when they get to the subordinate clause, because it has been "closed off," and therefore treat all such clauses as S-attached, thus allowing coreference in both sentence types. Thus, these data do *not* require us to posit individual differences in the formulation of Binding Theory. Besides the possibility that complete trees are never computed, an alternative interpretation suggested by Gerken and Bever is that we do compute full constituent structures but cannot access them for certain tasks, being left instead with the perceptual structure alone. This raises the possibility that linguists have developed ways to get at these fuller structures, which untrained informants cannot do. I must say I find this rather an unlikely outcome, although the idea is intriguing.

Obviously if the trend of linguists basing their theories on experimental data, as exhibited in the three studies above, is to continue and grow, linguists will have to be

trained in areas that they traditionally have been required to know nothing about: statistics and experimental design in general, and the psychology of grammaticality judgements in particular. This would seem to be a natural outgrowth of Chomsky's own suggestion that linguistics be viewed as a branch of cognitive psychology; somehow, the focus on cognitive issues has not yet been accompanied by adoption of the scientific standards of that discipline. But even if only a small proportion of linguists were to actually carry out their own experimental data collection, all could benefit by knowing more about problems of experimental bias, individual differences, introspection, etc. Will any of these recommendations be adopted? Carden and Dieterich (1981), who make proposals similar to our own, cite a typical response to their work, suggesting that many linguists will oppose such methodological changes: they cite Green (1978) as saying that if proposals like theirs were adopted, "research would come to a standstill." Certainly this would be true if *every* sentence had to be subjected to extensive experimental verification (Labov 1975), but that is unnecessary: if we adopt Labov's Clear Case Principle (see Chapter 5, §3.3), this will only be required when we have reason to believe that there is disagreement. Green continues with a second objection: "I doubt if any experimental results, no matter how clean, would affect the status of crucial disputed examples. Linguists will still trust their own intuitions of grammaticality." All I can say in response is, I hope not.

Somewhat more optimistically, Labov suggests that "introspective linguists" are most likely to resort to experimentation on data that are crucial both ways, i.e. which can either clinch their argument or destroy it. This would be a reasonable first step. Certainly, intuitive judgements by native speakers (but, one hopes, fewer and fewer linguists) will not be replaced by other kinds of language behaviour as the major source of data, at least on syntactic questions, in the foreseeable future, if ever. My view is that linguistics has much to gain and nothing to lose by taking data collection, and particularly judgement collection, much more seriously, both with regard to the insights that will be gained and theoretical issues clarified, and with regard to the standing of the field as a scientific endeavour in the larger academic setting. The realization does seem to be growing that the psychology of grammaticality judgements can no longer be ignored.

References

Andrews, Avery D. (1990). Case structures and control in Modern Icelandic. In Joan Maling & Annie Zaenen (Eds.), *Modern Icelandic syntax. Syntax and Semantics* **24**, San Diego, CA: Academic Press, 187–234.

Aoun, Joseph, Norbert Hornstein, David Lightfoot & Amy Weinberg (1987). Two types of locality. *Linguistic Inquiry* **18**(4), 537–577.

Armstrong, Sharon Lee, Lila R. Gleitman & Henry Gleitman (1983). What some concepts might not be. *Cognition* **13**(3), 263–308.

Asquith, Peter D. & Ronald N. Giere (Eds.) (1981). *PSA 1980: Proceedings of the 1980 biennial meeting of the Philosophy of Science Association. Volume 2: Symposia.* East Lansing, MI: Philosophy of Science Association.

Baker, Mark C. (1988). *Incorporation: A theory of grammatical function changing.* Chicago: University of Chicago Press.

Baltin, Mark (1977). Quantifier-negative interaction. In Ralph W. Fasold & Roger W. Shuy (Eds.), *Studies in language variation: Semantics, syntax, phonology, pragmatics, social situations, ethnographic approaches*, Washington, D.C.: Georgetown University Press, 30–36.

Barsalou, Lawrence W. (1987). The instability of graded structure: Implications for the nature of concepts. In Neisser 1987, 101–140.

Belletti, Adriana (1988). The case of unaccusatives. *Linguistic Inquiry* **19**(1), 1–34.

Belletti, Adriana & Luigi Rizzi (1988). Psych-verbs and θ-theory. *Natural Language and Linguistic Theory* **6**, 291–352.

Bergum, Bruce O. & Judith E. Bergum (1979a). Creativity, perceptual stability, and self-perception. *Bulletin of the Psychonomic Society* **14**(1), 61–63.

——— (1979b). Self-perceived creativity and ambiguous figure reversal rates. *Bulletin of the Psychonomic Society* **14**(5), 373–374.

Berkovits, Rochele (1981). Are spoken surface structure ambiguities perceptually unambiguous? *Journal of Psycholinguistic Research* **10**(1), 41–56.

——— (1982). On disambiguating surface-structure ambiguity. *Linguistics* **20**, 713–726.

Bever, Thomas G. (1970a). The cognitive basis for linguistic structures. In John R. Hayes (Ed.), *Cognition and the development of language,* New York: Wiley, 279–362.

——— (1970b). The influence of speech performance on linguistic structure. In G. B. Flores d'Arcais & W. J. M. Levelt (Eds.), *Advances in psycholinguistics*, Amsterdam: North-Holland, 4–30. [Reprinted in Bever, Katz & Langendoen 1976, 65–88.]

——— (1974). The ascent of the specious; or, There's a lot we don't know about mirrors. In David Cohen (Ed.), *Explaining linguistic phenomena*, New York: John Wiley & Sons, 173–200.

Bever, Thomas G., Caroline Carrithers & David J. Townsend (1987). A tale of two brains; or, The sinistral quasimodularity of language. In *Program of the ninth annual conference of the Cognitive Science Society*, Hillsdale, NJ: Erlbaum, 764–773.

Bever, Thomas, Jerrold Katz & D. Terence Langendoen (Eds.) (1976). *An integrated theory of linguistic ability*. New York: Crowell.

Bever, Thomas G. & D. Terence Langendoen (1971). A dynamic model of the evolution of language. *Linguistic Inquiry* **2**, 433–463.

Bialystok, Ellen (1986). Factors in the growth of linguistic awareness. *Child Development* **57**, 498–510.

Bialystok, Ellen & Ellen Bouchard Ryan (1985). A metacognitive framework for the development of first and second language skills. In Donna-Lynn Forrest-Pressley, G. E. MacKinnon & T. Gary Waller (Eds.), *Metacognition, cognition, and human performance. Volume 1: Theoretical perspectives*, Orlando: Academic Press, 207–252.

Biber, D. (1986). Spoken and written textual dimensions in English. *Language* **62**, 384–414.

Birdsong, David (1989). *Metalinguistic performance and interlinguistic competence*. New York: Springer-Verlag.

Bolinger, Dwight (1968). Judgments of grammaticality. *Lingua* **21**, 34–40.

——— (1971). Semantic overloading: A restudy of the verb *remind*. *Language* **47**(3), 522–547.

Botha, Rudolph P. (1973). *The justification of linguistic hypotheses: A study of non-demonstrative inference in transformational grammar*. (With the collaboration of Walter K. Winckler.) The Hague: Mouton.

Bradac, James J., Larry W. Martin, Norman D. Elliott & Charles H. Tardy (1980). On the neglected side of linguistic science: Multivariate studies of sentence judgment. *Linguistics* **18**(11/12), 967–995.

Browning, M. A. (1987). Null operators and their antecedents. In Joyce McDonough & Bernadette Plunkett (Eds.), *Proceedings of NELS 17, 1986, Volume 1*, Amherst, MA: Graduate Linguistic Student Association, Department of Linguistics, University of Massachusetts at Amherst, 59–78.

Burzio, L. (1981). *Intransitive verbs and Italian auxiliaries*. Unpublished Ph.D. Dissertation, MIT.

Carden, Guy (1970a). Discussion of Heringer 1970. In Mary Ann Campbell, James Lindhohm, Alice Davison, William Fisher, Louanna Furbee, Julie Lovins, Edward Maxwell, John Reighard & Stephen Straight (Eds.), *Papers from the sixth regional meeting, Chicago Linguistic Society*, Chicago: Chicago Linguistic Society, 296.

——— (1970b). A note on conflicting idiolects. *Linguistic Inquiry* **1**(3), 281–290.

——— (1976). Syntactic and semantic data: Replication results. *Language in Society* **5**(1), 99–104.

Carden, Guy & Thomas Dieterich (1981). Introspection, observation, and experiment: An example where experiment pays off. In Asquith & Giere 1981, 583–597.

Carroll, John M. (1979). Complex compounds: Phrasal embedding in lexical structures. *Linguistics* **17**, 863–877.

Carroll, John M., Thomas G. Bever & Chava R. Pollack (1981). The non-uniqueness of linguistic intuitions. *Language* **57**(2), 368–383.

Catt, Mark (1988). *Intelligent diagnosis of ungrammaticality in computer-assisted language instruction*. M.Sc. Thesis, Department of Computer Science, University of Toronto. [Published as Technical Report CSRI-218, Computer Systems Research Institute, University of Toronto.]

Catt, Mark & Graeme Hirst (1990). An intelligent CALI system for grammatical error diagnosis. *Computer Assisted Language Learning* **3**, 3–26.

Chaudron, C. (1983). Research on metalinguistic judgments: A review of theory, methods and results. *Language Learning* **33**(3), 343–377.

Chomsky, Noam [1955] (1985). *The logical structure of linguistic theory*. Chicago: University of Chicago Press. [Also New York: Plenum, 1975.]

——— (1957). *Syntactic structures*. The Hague: Mouton.

——— (1961). Some methodological remarks on generative grammar. *Word* **17**, 219–239.

——— (1965). *Aspects of the theory of syntax*. Cambridge: MIT Press.

——— (1973). Introduction. In Chomsky [1955] 1985, 1–53.

——— (1981). *Lectures on government and binding*. Dordrecht: Foris.

——— (1986). *Knowledge of language: Its nature, origin, and use*. New York: Praeger.

Clark, Herbert H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior* **12**, 335–359.

Coleman, E. B. (1965). Responses to a scale of grammaticalness. *Journal of Verbal Learning and Verbal Behavior* **4**(6), 521–527.

Coppieters, René (1987). Competence differences between native and near-native speakers. *Language* **63**(3), 544–573.

Cowart, Wayne (1989). Notes on the biology of syntactic processing. *Journal of Psycholinguistic Research* **18**(1), 89–103.

Crain, Stephen & Janet Dean Fodor (1987). Sentence matching and overgeneration. *Cognition* **26**(2), 123–169.

Danks, Joseph H. (1969). Grammaticalness and meaningfulness in the comprehension of sentences. *Journal of Verbal Learning and Verbal Behavior* **8**(6), 687–696.

Danks, Joseph H. & Sam Glucksberg (1971). Psychological scaling of adjective orders. *Journal of Verbal Learning and Verbal Behavior* **10**(1), 63–67.

Dellarosa, Denise (1988). A history of thinking. In Robert J. Sternberg & Edward E. Smith (Eds.), *The psychology of human thought*, Cambridge: Cambridge University Press, 1–18.

Dijk, Teun A. van (1977). Acceptability in context. In Greenbaum 1977a, 39–61.

Downey, Ronald G. & David T. Hakes (1968). Some psychological effects of violating linguistic rules. *Journal of Verbal Learning and Verbal Behavior* **7**(1), 158–161.

Duval, Shelley & Robert A. Wicklund (1972). *A theory of objective self awareness*. New York: Academic Press.

Elliot, Dale, Stanley Legum & Sandra Annear Thompson (1969). Syntactic variation as linguistic data. In Robert I. Binnick, Alice Davison, Georgia M. Green & Jerry L. Morgan (Eds.), *Papers from the fifth regional meeting of the Chicago Linguistic Society*, Chicago: Department of Linguistics, University of Chicago, 52–59.

Fillmore, Charles J. (1979). On fluency. In Fillmore et al. 1979a, 85–102.

Fillmore, Charles J., Daniel Kempler & William S-Y. Wang (Eds.) (1979a). *Individual differences in language ability and language behavior*. New York: Academic Press.

——— (1979b). Introduction. In Fillmore et al. 1979a, 1–10.

Forster, K. I. & B. J. Stevenson (1987). Sentence matching and well-formedness. *Cognition* **26**(2), 171–186.

Fowler, Roger (1970). Against idealization: Some speculations on the theory of linguistic performance. *Linguistics* **63**, 19–50.

Fraser, Bruce (1971). An analysis of *even* in English. In Charles J. Fillmore & D. Terence Langendoen (Eds.), *Studies in linguistic semantics*, New York: Holt, Rinehart & Winston, 151–180.

Gerken, LouAnn & Thomas G. Bever (1986). Linguistic intuitions are the result of interactions between perceptual processes and linguistic universals. *Cognitive Science* **10**, 457–476.

Gleitman, Henry & Lila R. Gleitman (1979). Language use and language judgment. In Fillmore et al. 1979a, 103–126.

Gleitman, Lila R., Henry Gleitman & Elizabeth F. Shipley (1972). The emergence of the child as a grammarian. *Cognition* **1**(2/3), 137–164.

Grandy, Richard E. (1981). Some thoughts on data and theory in linguistics. In Asquith & Giere 1981, 605–609.

Grant, David A., Jeffrey A. Kadlac, Marian Schwartz, Michael J. Zajano, Joseph B. Hellige, Louise C. Perry & Kenneth B. Solberg (1977). The role of noun imagery in the speed of processing the grammaticality of adjective-noun phrases. *Memory and Cognition* **5**(4), 491–498.

Green, Georgia (1978). Remarks on a proposal presented by Thomas Dieterich and Guy Carden at the NWAVE-VII Colloquium on the Validation of Introspective Judgements. Paper presented at the NWAVE-VII Conference, Georgetown University, 4 November 1978.

Greenbaum, Sidney (1973). Informant elicitation of data on syntactic variation. *Lingua* **31**, 201–212.

——— (1976a). Contextual influence on acceptability judgments. *Linguistics* **187**, 5–11. [Also published in *International Journal of Psycholinguistics* **6**, 1977, 5–11.]

——— (1976b). Syntactic frequency and acceptability. *Lingua* **40**(2/3), 99–113.

——— (Ed.) (1977a). *Acceptability in language*. The Hague: Mouton.

——— (1977b). Judgments of syntactic acceptability and frequency. *Studia Linguistica* **31**(2), 83–105.

Greenbaum, Sidney & Randolph Quirk (1970). *Elicitation experiments in English: Linguistic studies in use and attitude*. Coral Gables, FL: University of Miami Press.

Greenberg, Joseph H. & James J. Jenkins (1964). Studies in the psychological correlates of the sound system of American English. *Word* **20**(2), 157–177.

Grimshaw, Jane & Sara Thomas Rosen (1990). Knowledge and obedience: The developmental status of the binding theory. *Linguistic Inquiry* **21**(2), 187–222.

Grusec, Joan E., Robert S. Lockhart & Gary C. Walters (Eds.) (1990). *Foundations of psychology*. Toronto: Copp Clark Pitman.

Hakes, David T. (1980). *The development of metalinguistic abilities in children*. New York: Springer-Verlag.

Hardyck, Curtis, Hilary Naylor & Rebecca M. Smith (1979). How shall a thingummy be called? In Fillmore et al. 1979a 261–276.

Heeschen, Volker (1978). The metalinguistic vocabulary of a speech community in the highlands of Irian Jaya (West New Guinea). In Sinclair, Jarvella & Levelt 1978, 155–187.

Heringer, James T. (1970). Research on quantifier-negative idiolects. In Mary Ann Campbell, James Lindhohm, Alice Davison, William Fisher, Louanna Furbee, Julie Lovins, Edward Maxwell, John Reighard & Stephen Straight (Eds.), *Papers from the sixth regional meeting, Chicago Linguistic Society*, Chicago: Chicago Linguistic Society, 287–295.

Hill, Archibald A. (1961). Grammaticality. *Word* **17**(1), 1–10.

Hindle, Donald & Ivan Sag (1975). Some more on *anymore*. In Ralph W. Fasold & Roger W. Shuy (Eds.), *Analyzing variation in language: Papers from the second colloquium on new ways of analyzing variation*, Washington, D.C.: Georgetown University Press, 89–110.

Hirsh-Pasek, Kathy, Lila R. Gleitman & Henry Gleitman (1978). What did the brain say to the mind? A study of the detection and report of ambiguity by young children. In Sinclair, Jarvella & Levelt 1978, 97–132.

Hirst, Graeme (1981). *Anaphora in natural language understanding: A survey*. Berlin: Springer-Verlag.

Kaplan, Martin F. (1977). Judgment by juries. In Martin F. Kaplan & Steven Schwartz (Eds.), *Human judgment and decision processes in applied settings*, New York: Academic Press, 31–55.

Kess, Joseph F. & Ronald A. Hoppe (1983). Individual differences and metalinguistic abilities. *Canadian Journal of Linguistics* **28**(1), 47–53.

Klein, Eberhard (1979). The role of syntactic and semantic factors in explaining degrees of acceptability of non-finite verbal complement structures in English. *Linguistische Berichte* **61**, 1–20.

Kutas, Marta & Steven A. Hillyard (1983). Event-related brain potentials to grammatical errors and semantic anomalies. *Memory and Cognition* **11**(15), 539–550.

Labov, William (1972a). *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

——— (1972b). Some principles of linguistic methodology. *Language in Society* **1**(1), 97–120.

——— (1975). *What is a linguistic fact?* Lisse: Peter de Ridder. [Originally published as Empirical foundations of linguistic theory. In Robert Austerlitz (Ed.), *The scope of American linguistics*, Lisse: Peter de Ridder, 1975, 77–133.]

Lachter, Joel & Thomas G. Bever (1988). The relation between linguistic structure and associative theories of language learning—A constructive critique of some connectionist learning models. In Steven Pinker & Jacques Mehler (Eds.), *Connections and symbols*, Cambridge: MIT Press, 195–247. [Reprinted from *Cognition* **28**.]

Lakoff, George (1971). Presupposition and relative well-formedness. In Danny D. Steinberg & Leon A. Jakobovits (Eds.), *Semantics: An interdisciplinary reader in philosophy, linguistics and psychology*, Cambridge: Cambridge University Press, 329–340.

Lakoff, George (1987). Cognitive models and prototype theory. In Neisser 1987, 63–100.

Lakoff, Robin (1977). You say what you are: Acceptability and gender-related language. In Greenbaum 1977a, 73–86.

Langendoen, D. Terence & Thomas G. Bever (1973). Can a not unhappy person be called a not sad one? In Stephen R. Anderson & Paul Kiparsky (Eds.), *A festschrift for Morris Halle*, New York: Holt, Rinehart & Winston, 392–409. [Reprinted in Bever, Katz & Langendoen 1976, 239–260.]

Lasnik, Howard (1981). Learnability, restrictiveness, and the evaluation metric. In C. L. Baker & J. J. McCarthy (Eds.), *The logical problem of language acquisition*, Cambridge: MIT Press, 1–29.

Lasnik, Howard & Mamoru Saito (1984). On the nature of proper government. *Linguistic Inquiry* **15**(2), 235–289.

Lefever, M. M. & L. C. Ehri (1976). The relationship between field independence and sentence disambiguation ability. *Journal of Psycholinguistic Research* **5**(2), 99–107.

Levelt, W. J. M. (1972). Some psychological aspects of linguistic data. *Linguistische Berichte* **17**, 18–30.

——— (1974). *Formal grammars in linguistics and psycholinguistics*. 3 vols. The Hague: Mouton.

Levelt, W. J. M., J. A. W. M. van Gent, A. F. J. Haans & A. J. A. Meijers (1977). Grammaticality, paraphrase, and imagery. In Greenbaum 1977a, 87–101.

Levelt, W. J. M., A. Sinclair & R. J. Jarvella (1978). Causes and functions of linguistic awareness in language acquisition: Some introductory remarks. In Sinclair, Jarvella & Levelt 1978, 1–14.

Maclay, Howard & Mary D. Sleator (1960). Responses to language: Judgments of grammaticalness. *International Journal of American Linguistics* **26**(4), 275–282.

Marks, Lawrence E. (1967). Judgments of grammaticalness of some English sentences and semi-sentences. *American Journal of Psychology* **80**(2), 196–204.

——— (1968). Scaling of grammaticalness of self-embedded English sentences. *Journal of Verbal Learning and Verbal Behavior* **7**(5), 965–967.

Masny, Diana & Alison d'Anglejan (1985). Language, cognition, and second language grammaticality judgments. *Journal of Psycholinguistic Research* **14**(2), 175–197.

McCawley, James D. (1985). Review of Newmeyer 1983. *Language* **61**(3), 668–679.

McRoy, Susan Weber & Graeme Hirst (1990). Race-based parsing and syntactic disambiguation. *Cognitive Science* **14**, 313–353.

Miller, George A. (1962). Some psychological studies of grammar. *American Psychologist* **17**, 748–762.

Milne, Robert William (1982). Predicting garden path sentences. *Cognitive Science* **6**, 349–373.

Mohan, B. A. (1977). Acceptability testing and fuzzy grammar. In Greenbaum 1977a, 133–148.

Moore, Timothy E. (1972). Speeded recognition of ungrammaticality. *Journal of Verbal Learning and Verbal Behavior* **11**, 550–560.

——— (1975). Linguistic intuitions of twelve year-olds. *Language and Speech* **18**(3), 213–218.

Moore, Timothy E. & Irving Biederman (1979). Speeded recognition of ungrammaticality: Double violations. *Cognition* **7**(3), 285–299.

Nagata, Hiroshi (1987a). Change in the modulus of judgmental scale: An inadequate explanation for the repetition effect in judgments of grammaticality. *Perceptual and Motor Skills* **65**(3), 907–910.

——— (1987b). Long-term effect of repetition on judgments of grammaticality. *Perceptual and Motor Skills* **65**(1), 295–299.

——— (1988). The relativity of linguistic intuition: The effect of repetition on grammaticality judgments. *Journal of Psycholinguistic Research* **17**(1), 1–17.

——— (1989a). Effect of repetition on grammaticality judgments under objective and subjective self-awareness conditions. *Journal of Psycholinguistic Research* **18**(3), 255–269.

——— (1989b). Judgments of sentence grammaticality and field-dependence of subjects. *Perceptual and Motor Skills* **69**(3), 739–747.

——— (1989c). Judgments of sentence grammaticality with differentiation and enrichment strategies. *Perceptual and Motor Skills* **68**(2), 463–469.

——— (1989d). Repetition effect in judgments of grammaticality of sentences: Examination with ungrammatical sentences. *Perceptual and Motor Skills* **68**(1), 275–282.

——— (1990). On-line judgments of grammaticality of sentences. *Perceptual and Motor Skills* **70**(3), 987–994.

Neisser, Ulric (Ed.) (1987). *Concepts and conceptual development: Ecological and intellectual factors in categorization.* Cambridge: Cambridge University Press.

Newmeyer, Frederick J. (1983). *Grammatical theory, its limits and its possibilities.* Chicago: University of Chicago Press.

Ney, James W. (1975). The decade of private knowledge: Linguistics from the early 60's to the early 70's. *Historiographia Linguistica* **2**(2), 143–156.

Nisbett, R. E. & T. Decamp Wilson (1977). Telling more than we know: Verbal reports on mental processes. *Psychological Review* **84**, 231–259.

Oller, John W., Jr., B. Dennis Sales & Ronald V. Harrington (1970). Toward consistent definitions of some psycholinguistic terms. *Linguistics* **57**, 48–59.

Pollock, Jean-Yves (1989). Verb movement, universal grammar, and the structure of IP. *Linguistic Inquiry* **20**(3), 365–424.

Pullum, Geoffrey K. (1987). Seven deadly sins in journal publishing. *Natural Language and Linguistic Theory* **5**(3), 453–459. [Reprinted in Geoffrey K. Pullum, *The great Eskimo vocabulary hoax, and other irreverent essays on the study of language*, Chicago: University of Chicago Press, 1991, 84–91.]

Quirk, Randolph & Jan Svartvik (1966). *Investigating linguistic acceptability*. The Hague: Mouton.

Ray, William J. & Richard Ravizza (1988). *Methods toward a science of behavior and experience*. Belmont, CA: Wadsworth.

Reich, Peter A. (1969). The finiteness of natural language. *Language* **45**(4), 831–843.

Riemsdijk, Henk C. van & Edwin Williams (1986). *Introduction to the theory of grammar*. Cambridge: MIT Press.

Ringen, Jon (1975). Linguistic facts: A study of the empirical scientific status of transformational generative grammars. In D. Cohen & J. R. Wirth (Eds.), *Testing linguistic hypotheses*, Washington, D.C.: Hemisphere Publishing, 1–41.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General* **104**, 192–233.

Rose, Robert G. (1973). Linguist and nonlinguist agreement concerning surface structures. *The Journal of General Psychology* **89**(2), 325–326.

Ross, John Robert (1979). Where's English? In Fillmore et al. 1979a, 127–163.

Sachs, J. S. (1967). Recognition memory for syntactic and semantic aspects of connected discourse. *Perception and Psychophysics* **2**, 437–442.

Scholes, Robert J. & Brenda J. Willis (1987). Literacy and language. *Journal of Literary Semantics* **16**(1), 3–11.

Scott, Robert Ian (1969). A permutational test of grammaticality. *Lingua* **24**(1), 11–18.

Scott, Robert Ian & John A. Mills (1973). Validating the permutational test of grammaticality. *Language and Speech* **16**(2), 110–122.

Scribner, Sylvia & Michael Cole (1981). *The psychology of literacy*. Cambridge: Harvard University Press.

Sinclair, A., R. J. Jarvella & W. J. M. Levelt (Eds.) (1978). *The child's conception of language*. Berlin: Springer-Verlag.

Snow, Catherine E. (1975). Linguists as behavioral scientists: Towards a methodology for testing linguistic intuitions. In A. Kraak (Ed.), *Linguistics in the Netherlands 1972–1973*, Assen: Van Gorcum, 271–275.

Snow, Catherine & Guus Meijer (1977). On the secondary nature of syntactic intuitions. In Greenbaum 1977a, 163–177.

Sobin, Nicholas (1987). The variable status of comp-trace phenomena. *Natural Language and Linguistic Theory* **5**(1), 33–60.

Spencer, N. J. (1973). Differences between linguists and nonlinguists in intuitions of grammaticality-acceptability. *Journal of Psycholinguistic Research* **2**(2), 83–98.

Taylor, M. M. & G. B. Henning (1963). Verbal transformations and an effect of instructional bias on perception. *Canadian Journal of Psychology* **17**, 210–223.

Thráinsson, Höskuldur (1979). *On complementation in Icelandic*. New York: Garland.

Tottie, Gunnel (1977). Variation, acceptability and the advanced foreign learner: Towards a sociolinguistics without a social context. In Greenbaum 1977a, 203–213.

van Dijk, Teun A. See Dijk, Teun A. van.

Van Petten, Cyma & Marta Kutas (1991). Influences of semantic and syntactic context on open- and closed-class words. *Memory and Cognition* **19**(1), 95–112.

van Riemsdijk, Henk C. See Riemsdijk, Henk C. van.

Vetter, Harold J., Jerry Volovecky & Richard W. Howell (1979). Judgments of grammaticalness: A partial replication and extension. *Journal of Psycholinguistic Research* **8**(6), 567–583.

Walker, E. L. (1973). Psychological complexity and preference: A hedgehog theory of behavior. In D. E. Berlyne & K. B. Madsen (Eds.), *Pleasure, reward, preference: Their nature, determinants, and role in behavior*, New York: Academic Press.

Warner, John & Arnold L. Glass (1987). Context and distance-to-disambiguation effects in ambiguity resolution: Evidence from grammaticality judgments of garden path sentences. *Journal of Memory and Language* **26**, 714–738.

Weiner, Bernard, Willard Runquist, Peggy A. Runquist, Bertram H. Raven, William J. Meyer, Arnold Leiman, Charles L. Kutscher, Benjamin Kleinmuntz & Ralph Norman Haber (1977). *Discovering psychology*. Chicago: Science Research Associates.