# Gradient Data and Gradient Grammars*

Thomas Wasow
Stanford University

## 1 Introduction

Throughout the history of generative grammar there has been a tension between the categorical nature of the theories proposed and the gradient[1] character of the data used as evidence for those theories. This paper addresses the question of whether gradient grammars would provide better models for the phenomena syntacticians study.

The remainder of section 1 spells out in more detail the premise of the paper, namely, that linguistic data are gradient but generative grammars generally are not. Section 2 discusses several proposals over the past half-century for incorporating gradience into generative grammars. Section 3 addresses how the issue of gradience fits into recent debates over the relationship between grammar and usage, and the relevance of the competence/performance distinction. Section 4 considers what kinds of evidence might settle the question of whether grammars should be gradient, and considers some possible cases. Some rather inconclusive conclusions constitute section 5.

### 1.1 Gradient Data

The dominant type of data in the generative tradition has been judgments of well-formedness or of meaning. The widespread use of symbols like *?, ??, ?*, *?,* etc, in addition to the standard asterisk, constitutes a tacit recognition that acceptability of linguistic forms is not consistently an all-or-nothing matter.

The gradient nature of acceptability judgments has occasionally played a role in theory construction, when constraints are differentiated on the basis of how bad it sounds when they are violated. For example, the distinction between strong and weak crossover phenomena (first suggested in my 1972 MIT dissertation) is based largely on the intuition that examples like (1a) are a lot worse than examples like (1b):

(1) a. *Who$_i$ does he$_i$ think proved Fermat's last theorem?
b. ??Who$_i$ does someone he$_i$ knows think proved Fermat's last theorem?

---

* I am grateful to Ivan Sag for stimulating discussions relevant to this paper. Jerry Sadock also provided me with helpful feedback after my oral presentation of the paper. Work on the paper was supported in part by NSF Award No. IS-0624345.

[1] The term "gradient" appears in dictionaries I have consulted only as a noun, but is often used as an adjective by linguists. The noun "gradience" does not appear in dictionaries; Wikipedia attributes its coinage to Dwight Bolinger. I use "gradient" (as seems to be standard among linguists) as a rough synonym for "graded" and as an antonym for "categorical"; I use "gradience" as a nominalization of this use of "gradient", denoting the property of being gradient.

Similarly, the distinction between weak and strong islands is based on intuitions of relative acceptability. Such intuitions constitute a form of gradient judgment data.

Recently, several works (e.g., Bard, et al, 1996; Cowart, 1997, Schütze, 1996) have advocated treating the collection of judgments as a form of psychological experiment, subject to the accepted methodological standards of such studies, including use of enough participants and items to do meaningful statistical analyses. Typically, such studies offer participants more than a binary choice between acceptable and unacceptable, either through the use of a fixed scale of responses or through magnitude estimation. And participants invariably make use of a substantial range of response options; that is, the responses rarely cluster around the extremes of acceptability and unacceptability. For example, I was involved in one study (reported on in Rickford, et al, 1995) in which subjects were asked to rate 20 sentences on a 4-point acceptability scale. Of the 80 cells in the 20-by-4 matrix of results this produced, none ended up with a zero; that is, for each sentence and each acceptability rating, there was some participant who gave that sentence that rating.

Other forms of experimental data used in linguistics are even more clearly gradient. Paradigms like self-paced reading, visual world, and ERP typically produce results based on relatively subtle but significant statistical differences. They often require large numbers of subjects and items, and achieving statistical significance of results may require discarding outliers or other forms of massaging the data. In this respect, experimental data in linguistics resemble experimental data in psychology, where nobody would expect categorical results.

Corpus data are almost always gradient, as well. Indeed, corpora have been used to show that some putatively categorical constraints from the linguistics literature are in fact widely violated. For example, Bresnan and Nikitina (2007) found numerous counterexamples to the supposed unacceptability of personal pronouns as second objects in the English double object construction; (2) is one such example they found in a corpus of web documents.

(2) Please follow these simple rules and teach your children them.

I hasten to add that categorical constraints in syntax do exist. An example from English is that heavy NP shift is incompatible with the double object construction. That is, while heavy direct objects can follow other NPs that are used predicatively, as in (3a), no such non-canonical ordering is possible when the other NP is an object of the verb.

(3) a. They consider [a traitor] [anyone who opposes their war policies].
b. *I don't envy [the adulation] [rock stars their fans worship]

The unacceptability of (3b) is all the more remarkable because the canonical ordering, given in (4), induces a garden path, making the sentence extremely hard to parse.

(4) ??I don't envy [rock stars their fans worship] [the adulation].

Through years of collecting examples of heavy NP shift and searching corpora for them, I have never encountered an example of it in the double object construction.

My claim about gradience of syntactic data, then, is an existential one, namely, that some (indeed many) grammatical constraints are not categorical. Consequently, there are a great many sentences that are neither fully acceptable, nor totally unacceptable. But there are categorical constraints, as well.

### 1.2 Categorical Grammars

This is a rather weak claim, and it is unlikely that many linguists would disagree with it. However, the practice of generative grammarians has usually been to pretend that it is false. A particularly clear statement of the working assumption that has dominated the field is the following:

> The fundamental aim in the linguistic analysis of a language L is to separate the *grammatical* sequences which are sentences of L from the *ungrammatical* sequences which are not sentences of L and to study the structure of the grammatical sequences. The grammar of L will thus be a device that generates all of the grammatical sequences of L and none of the ungrammatical ones. (Chomsky, 1957; 13)

In practice, linguists assume that the sequences of words from the vocabulary of a language can be separated into those that are sentences and those that are not. We use native speakers' judgments of acceptability to decide which sequences should be regarded as sentences and which should not be. We construct grammars (or, more commonly, tiny fragments of grammars) based on these binary judgments. For strings that native speakers do not robustly classify as sentences or non-sentences, Chomsky (1957;14) suggested that we "let the grammar itself decide". In practice, however, examples of intermediate acceptability often play a crucial role in theoretical argumentation, with the authors pretending that they are fully grammatical or ungrammatical, as suits the argument.

In short, generative grammar has traditionally formulated categorical grammars on the basis of gradient data.

## 2 Some History

The gradient character of linguistic data has been acknowledged since the advent of generative grammar. In Chomsky's seminal work, *The Logical Structure of Linguistic Theory*, in a chapter entitled "Grammaticalness", he writes, "a partition of utterances into just two classes, grammatical and nongrammatical, will not be sufficient to permit the construction of adequate grammars" (Chomsky 1955/75; 129). In that chapter and in a 1961 article, Chomsky proposed that a language could be characterized by a hierarchy of grammars, with increasingly fine-grained category structure. For example, at a fairly coarse level, the sentence *Golf plays John* could be generated, because rules making reference only to basic parts of

speech (noun, verb, etc) would license the N-V-N sequence. At a finer level, however, it would not be generated, because the verb *plays* would require an animate subject. Thus, it would be less grammatical than *John plays golf*, which satisfies both the constraints on basic sequencing of lexical categories, and the selectional restrictions of the verb.

Katz (1964), in another early piece on degrees of acceptability, argues that Chomsky's proposal is inadequate to capture the full range of intermediate cases. He claims that, for example, omission of function words in examples like *Man bit dog* or *The ball hit by the man* render them less than fully acceptable sentences, but they are comprehensible and not totally unacceptable. Yet Chomsky's proposal for degrees of grammaticalness would assign them a low degree. The same could be said for minor deviations from normal word order, such as *I will be tomorrow in the office.*

As generative grammar became established as the dominant paradigm in linguistics, discussion of intermediate levels of acceptability waned. Chomsky's idea of a hierarchy of grammars was never taken up again. Ross's work on "squishes" attracted some interest, but not much of a following. In several works (e.g. Ross, 1972, 1973), he argued that many grammatical properties were "squishy" – that is, gradient. For example, he argued that different types of nominal constituents exhibit different degrees of "nouniness" in their behavior, offering an ordered list of diagnostics for nounhood, and claiming that any type of constituent satisfying any diagnostic on the list would also satisfy all the diagnostics ordered below it.

A common feature of Chomsky's proposal and Ross's is that both accommodate gradience in grammar without introducing any explicitly quantitative component into the theory. By relying on ordinal ranking, both proposals allow grammatical properties to have multiple levels, but these are not assigned numerical values. The same is true of Optimality Theory (Prince and Smolensky, 2004), though the connectionist foundations of OT are thoroughly quantitative (see Smolensky and Legendre, 2006).

Indeed, standard OT takes pains to avoid making non-categorical predictions, though certain versions of it have been adapted to account for gradient data. In particular, stochastic OT (Boersma and Hayes, 2001) assigns probability distributions to constraint rankings, with the result that specific forms are assigned probabilities of occurrence. A different way of deriving quantitative predictions of relative frequencies in OT is Anttila's (2006) proposal that constraint rankings are only partial, and that all total rankings compatible with the partial fixed rankings are considered in licensing forms. He claims that the frequencies of occurrence of varying forms reflect the frequencies with which they are selected by the different total constraint rankings that are considered. This interesting idea generates gradient predictions (of frequency of occurrence) entirely from the combinatorics of the formal system. Yet another way of accommodating gradience in OT is Coetzee's (2004) proposal that the ranking of "losers" in OT tableaux can account for language variation.

Long before the invention of OT, sociolinguists argued for accommodating gradience in grammar. Labov (1969) argued that at least some rules of grammar are variable; that is, he proposed "to associate with each variable rule a specific quantity φ which denotes the probability of the rule applying" (p. 95). He went on to say:

> we are in no way dealing with statistical statements or approximations to some ideal or true grammar. We are dealing with a set of quantitative *relations* which are the form of the grammar itself. (p. 125)

This straightforward idea for extending the formalism of grammatical rules to accommodate one type of gradience was never embraced by generativists, despite its central role in sociolinguistic research. It is worth considering why.

The formal foundations of generative grammar came from mathematical logic, not from any quantitative branch of mathematics. Chomsky's early work in formal language theory (e.g. Chomsky, 1956) explores the mathematical properties of categorical rule systems. Although other scholars of that era (most notably Zipf, 1949) were interested in the statistical distributions of forms in natural languages, Chomsky was very critical of such approaches. He wrote:

> If we rank the sequences of a given length in order of statistical approximation to English, we will find both grammatical and ungrammatical sequences scattered throughout the list; there appears to be no particular relation between order of approximation and grammaticalness. … [P]robabilistic models give no particular insight into some of the basic problems of syntactic structure. (Chomsky, 1957; 17)

Chomsky continued to dismiss quantitative modeling in linguistics for many years (although less frequently, as time went on). His central argument against probabilistic linguistics, repeated in slightly different form in various places, is the following:

> [S]ince most of the 'normal sentences' of daily life are uttered for the first time in the experience of the speaker-hearer (or in the history of the language, the distinction hardly being important at this level of frequency of occurrence), they will have had probability zero before this utterance was produced and the fact they are produced and understood will be incomprehensible in terms of a 'probabilistic theory of grammar'. (Chomsky, 1966; 35)

This argument specifically attacks the feasibility of identifying the probability of a sentence with the frequency of occurrence of that specific string of words, but he generalizes this argument to other means of assigning probabilities to sentences:

> Nor can we take the probability of a sentence to be an estimate based on the probability of its parts, in some way; such a decision would ... give a probability scale that has no useful relation to the intuitive scale

of grammaticalness on which all linguistic description is based (Chomsky, 1966; 36)

Chomsky's claims that quantitative methods could provide nothing of use to linguistics were difficult to challenge in the 1950s. Natural languages have vocabularies of tens of thousands of words, so even studying frequencies of three-word strings involves considering trillions of possibilities. The combinatoric explosion gets even worse as the strings get longer. Chomsky's arguments had a *prima facie* plausibility, and it was convenient to be able to focus on constructing grammars based on intuitive judgments of well-formedness, without examing usage data. This focus produced several decades of very productive research in syntactic and semantic theory (see Newmeyer, 1986, for a summary).

Eventually, however, computer technology changed this. By the mid-1980s, processors had become fast enough and memory had become inexpensive enough that serious statistical modeling of natural language became feasible. Individuals and companies involved in building practical natural language technologies largely abandoned categorical theories in favor of statistics-driven systems. Natural language processing systems based on theoretically-motivated grammars had suffered from two pervasive problems: very limited coverage, and a tendency to break when attempts were made to extend that coverage. Systems based on statistical patterns of usage might give wrong results quite frequently, but they always give some results, and improvements in the quality of those results can be achieved in small increments.

Some theoretical linguists soon followed suit. The availability of on-line corpora and computational tools for searching them made it possible for corpus linguists to do frequency counts in minutes that it would have taken earlier generations of linguists years to complete. In the 15 years or so that I have been engaged in this sort of work, the technological advances have been dramatic. The advent of linguistically annotated corpora (especially syntactically parsed corpora like the Treebank of the Linguistic Data Consortium) has greatly facilitated this line of research (as, of course, has Moore's Law). Searches that now take milliseconds to carry out on a laptop would have taken overnight on a supercomputer a decade ago, and might have filled a career prior to the advent of computers.

As noted above, corpus studies almost never yield categorical results, so analyzing corpus data requires statistical tools. This naturally gives rise to probabilistic models of language structure. Hence, as theoretical linguists have come increasingly to employ corpora in their research, they have begun to ask whether a grammar should represent not just what is possible in a language, but also something about relative frequencies of the possible forms (see, e.g. Wasow, 2002, Bresnan, 2006).

For the most part, Chomsky's old arguments against the use of frequency data in linguistic theorizing have not been revisited. One exception is Pereira's (2002) demonstration that a fairly straightforward computation of probabilities of occurrence based on corpus bigrams predicts that *Colorless green ideas sleep*

*furiously* is about 200,000 times more probable than *Furiously sleep ideas green colorless*, contrary to Chomsky's (1957) assertion that no probabilistic theory could distinguish between them.

## 3 The Contemporary Grammar/Usage Debate

In his LSA Presidential Address, Newmeyer (2003) criticized the use of corpus data and quantitative models in syntactic theorizing. Briefly, Newmeyer's central arguments on this topic are the following:

- Comparisons of corpus frequencies of alternate forms can't control for subtle meaning differences between the forms
- Grammars characterize the linguistic competence of individual speakers, but corpora average over multiple speakers
- Non-categorical, but non-random, usage by an individual can be explained in terms of non-linguistic knowledge and processing efficiency

Before turning to these arguments, it is worth noting that Newmeyer assumes that the burden of proof is on advocates of what he calls "stochastic grammars" – by which he seems to mean models of linguistic competence that involve quantitative components. His claim is a universal one: that all grammatical mechanisms are categorical. Hence, one putative counterexample – one *prima facie* example of gradience in grammar – puts the burden of proof on him. I will offer potential examples below.

If Newmeyer is right and grammars are never gradient, how can one reconcile the discrepancy between the gradience of linguistic data and the categorical nature of the dominant theories of grammar? Three possible answers (which are not mutually exclusive) correspond to the three bullet points above. First, it is possible that at least some cases of apparent gradience in the choice of grammatical form are actually cases in which different meanings are being expressed. Second, some cases of gradience may be the result averaging over the usage of multiple individuals, who have different internalized grammars. Third, there are gradient factors affecting language use that are not part of linguistic competence.

Newmeyer cites Lavandera's (1978) argument against positing variable rules in syntax, on the grounds that what are regarded as syntactic alternatives, such as active vs. passive, typically don't mean quite the same thing. The position that there is no such thing as true paraphrase stated most clearly by Bolinger (1968; 127), who wrote, "a difference in syntactic form always spells a difference in meaning". Paraphrase denial is implausible: there are certain alternations that do not appear to involve a meaning difference, e.g. the optionality of *that* in examples like (5).

(5) a. I'm planning on visiting Barcelona on the vacation (that) I'm taking.
    b. I was surprised (that) he got in an accident..

It turns out, however, that some linguists have argued for meaning differences based on the presence or absence of *that* (see Yaguchi, 2001 and Dor, 2005).

Kinsey, Jaeger, and Wasow (2007) embedded such sentences in brief contexts and had readers rate them for emotionality and temporal distance, two semantic properties Yaguchi claims differ in the versions with and without *that*. The presence of *that* made no difference in readers' responses. In contrast, readers who rated versions of the sentences containing emotional or temporal adverbs found clear meaning differences between those with and without adverbs. While this alternation might convey some other meaning difference, the burden of proof is clearly on the paraphrase deniers.

Moreover, even if it could be shown that no true paraphrases exist, the meaning differences between alternating forms are often extremely subtle. In many instances of use, what drives the choice between two forms is not the difference in their meanings (either one of which would be adequate to express the speaker's intentions), but other factors, such as relative length of constituents or the newness of the information they express. Hence, it is legitimate to look at all of the factors, including semantic ones, that influence the choices among linguistic forms. This involves statistical modeling.

Newmeyer does not object to this enterprise. But he claims that those of us engaged in it are studying performance, not competence. Categorical competence grammars can give rise to gradient performance data either through averaging across the outputs of individual grammars or through interaction with gradient performance mechanisms.

Newmeyer (2003; 696) writes, "There is no way that one can draw conclusions about the grammar of an individual from usage facts about communities"; after citing a "nonsequitur" from the literature, he continues, "The point is that we do not have 'group minds'." But variation in the usage of individuals certainly exists. Few, if any, speakers always use *that* or always omit it in examples like (5). Moreover, there are well-established statistical techniques for testing the effects of individual differences, both in experimental work (Clark, 1973) and in corpus studies (Bresnan, et al 2005). And the data remain largely gradient when the effects of individual differences are factored out.

Newmeyer's strongest argument against gradience in grammar is that all gradience in linguistic data comes from performance facts. As he puts it (p. 696): "the evidence for probabilities being associated with grammatical elements seems pretty weak. The numbers are overwhelmingly epiphenomenal." This kind of argument is not new. Kiparsky (1972; 223) gave very similar reasons for rejecting variable rules in phonology:

> Labov ... claimed that these frequencies are part of linguistic competence ... The alternative hypothesis would be that they are the result of general functional conditions impinging on speech performance.

### 3.1 Competence and Performance

What is the basis for attributing some feature of language to competence or to performance? Historically, there have been at least two distinct notions of competence that have been invoked in the linguistics literature.

The first, which I will call the "pure language" conception, stems from the observation that judgments of acceptability can be affected by factors that are not strictly linguistic, such as memory limitations or plausibility. For example, the fact that a sentence a million words long would not be comprehensible is irrelevant to the question of its grammaticality. Likewise, the oddness of a sentence like *Fermat's last theorem rallied the stock market* has to do with common-sense knowledge of the world, not with any constraints on language. A more subtle and hence more interesting form of this conception of competence is the idea that the architecture of the human parser makes certain sentences hard for people to process, even though they are grammatically well formed. Paradigm examples of this are center self-embedding examples like (6a) and "garden-path" sentences like (6b).

(6) a. The cheese the mouse the cat the dog chased killed ate was moldy.
b. The boat floated down the river sank.

The pure language notion is that competence is what is left when such functionally motivated factors are excluded.

The other notion of competence is what I will call the "shared knowledge" conception. It is based on the observation that there is a common body of knowledge that we draw on for distinct uses of language, including understanding what others say, formulating our own utterances, making judgments regarding example sentences, reading, writing, translating, paraphrasing, making puns, etc. Under this conception of competence, it consists of the shared knowledge that is required for various uses of language – particularly the two primary ones, producing and comprehending ordinary conversation. It is this conception that is suggested by Chomsky's frequent summary of the competence/performance distinction as the difference between knowledge of language and use of language (see, e.g. Chomsky, 1986).

The two notions of competence are not equivalent. Competent use of language requires many kinds of knowledge that would be excluded under the pure language conception. In particular, pragmatic knowledge of the world plays a role in many language tasks, but is not strictly linguistic in the way required by the pure language conception of competence. For example, since a sentence of the form *X walks Y* entails the corresponding sentence *X walks*, but not conversely, it follows that *X walks* truly characterizes more real-world situations than *X walks Y* (for any choice of Y). Consequently, speakers know that *walks* is used more frequently intransitively than transitively[2]. It would be straightforward to show that this sort of knowledge plays a role in production, comprehension, metalinguistic judgments, etc. So, under the shared knowledge conception, the

[2] This is a modified version of an argument given by Newmeyer (2003; 696), responding to a claim made in Wasow (2002).

quantitative fact that *walks* is more frequently intransitive than transitive is part of competence; but under the pure language conception, it is not.

Jaeger (2007) argues that the same (gradient) verb biases that influence comprehension also influence production. In particular, Garnsey, et al (1997) showed that a verb's preference for a direct object or a finite clausal complement influences how fast they are processed, and Jaeger (2006) showed that the same preference influence the rate at which the finite complements begin with *that*. Assuming the shared knowledge conception of competence, this provides a clear argument for including gradient information – verb subcategorization biases – in competence.

But this argument would not convince adherents of the pure language conception of competence. They could argue that linguistic competence is only part of the knowledge that is shared in diverse linguistic behaviors. Knowledge of the world – e.g. that walking a dog involves walking – is arguably not linguistic knowledge, and hence not part of competence. Certain processes that are involved in most uses of language, such as lexical retrieval and parsing, are excluded from competence under the pure language conception because of their functional motivation. We cannot do anything with our knowledge of the words and structures that a language contains unless we can access that knowledge during production, comprehension, and other linguistic tasks. The processes through which such access is accomplished are not part of the pure language conception of competence. From this perspective, involvement in multiple linguistic tasks is a necessary but not sufficient condition for being considered part of competence.

According to the pure language conception, then, knowledge that is used in linguistic behavior belongs to linguistic competence only if it cannot be explained functionally. That is, anything that facilitates efficient communication by exploiting knowledge of the world, resource limitations of the language faculty, the architecture of the parser, or the like, is not part of competence. This raises the question of whether pure language competence exists at all: it is conceivable that everything about natural language structure has a functional explanation. Even if there are some aspects of language that serve no communicative function, one might argue that they are the least interesting ones, since they appear to be accidents.

But such arguments seem rather pointless. Isn't this just a terminological dispute? Everyone agrees that gradient mechanisms are essential components of the language faculty. Does it really matter what one calls competence and what one calls performance? I claim that it does, though not for any deep theoretical reason.

Over the past fifty years, the central focus of mainstream theoretical linguistics has been the study of linguistic competence. The study of performance has been widely regarded as of secondary importance and necessarily dependent on a well-motivated theory of competence. Hence, what is deemed to be an aspect of competence influences what linguists study. In particular, it matters

whether the knowledge that results in the gradience of linguistic data is considered part of competence. Calling something competence affects how much it is studied, what kinds of hypotheses are formulated, what kinds of methods are used to test those hypotheses, and the interpretation of the studies. Consequently, it is worth considering what might constitute evidence of gradience in grammar, even for an adherent of the pure language conception of competence.

## 4 Possible Cases of Gradience in Grammar

In this section I tentatively offer three cases of gradience in language that might argue for including some quantitative information in the grammar of English. I am under no illusion that they will convince believers in pure language competence that grammars must contain numbers. But I hope that they at least serve to place the burden of proof clearly on those who deny gradience in grammar.

### 4.1 Collocations

Fixed multi-word expressions abound in natural languages, including both idioms and compositional collocations. It is not hard to see how their existence might facilitate both production and comprehension: storing chunks that might be used frequently can save time both in formulating and understanding utterances. But quantitative information about the frequency of collocations has no such evident usefulness.

Consider a fully compositional collocation like *unmitigated disaster*. The adjective *unmitigated* is a relatively rare one, but when it occurs, it is frequently followed by *disaster*. Speakers of English know this; that is, they know that the probability of hearing *disaster* next when they hear *unmitigated* is quite high. But they also know that *unmitigated* can occur without *disaster* (as it does about three quarters of the time, according to a Google search of the web). This is gradient linguistic knowledge that does not have any obvious functional explanation.

This is by no means a unique case. There are other words whose collocational occurrences constitute a very high proportion of their total uses, e.g., *hermetically (sealed), diametrically (opposed* or *opposite), (take) umbrage,* etc. The knowledge that these words tend to occur in collocations is conscious knowledge, unlike much linguistic knowledge. The exact strength of this tendency in any given case is of course not conscious but could in principle be determined through psycholinguistic experimentation.

More generally, we know a great deal about the cooccurrence preferences of the words in our vocabulary. This knowledge is not just what **can** occur with what, but how probable it is that words will occur together. Sometimes, there is a functional explanation for these cooccurrence constraints (such as the preference for *walk* to be used intransitively), but sometimes they are just something the language user learns from experience with the language. If such cases are to be

treated as qualitatively different from categorical cooccurrence restrictions, a principled reason for doing so needs to be given.

### 4.2 Experiencer Verbs

Transitive verbs that denote a psychological state of an experiencer with respect to some thing or event sometimes express the experiencer argument as subject (e.g. *enjoy*) and sometimes as direct object (e.g. *amuse*). Sometimes two verbs denoting much the same psychological state differ in terms of which grammatical function expresses the experiencer, e.g. *fear* vs. *frighten*. As is well known, this is true across languages.

One such pair in English is *like* and *please*, illustrated in (7).

(7) a. The reviewer liked the movie.
b. The movie pleased the reviewer.

Native speakers of English know that *like* in this sense is far more frequent than *please*. In the three corpora of the Penn Treebank, *like* with an experiencer subject occurs 1037 times, whereas *please* with an experiencer object occurs only 31 times. To the best of my knowledge, this is a rather arbitrary asymmetry of this one verb pair in English. For example, no such asymmetry exists between *fear* (39 occurrences in the Penn Treebank) and *frighten* or *scare* (67 occurrences combined). The German counterpart to *please* (*gefallen*) does not have the same somewhat formal or old-fashioned ring to it as examples like (7b), and it is much more frequently used.

The fact that *like* is favored over *please* is gradient linguistic knowledge shared by (American)[3] English speakers. It has no apparent functional explanation in terms of knowledge of the world or processing factors. It is hence a candidate for gradient grammatical information.

### 4.3 English NP-PP Order

In English verb phrases that immediately dominate both a noun phrase and a prepositional phrase, over 90% occur with the NP preceding the PP[4]. But the grammar of English must license both the NP-PP and the PP-NP order, to accommodate both variants in (8).

(8) a. That brings Barry Bonds to the plate.
b. That brings to the plate Barry Bonds.

The strong preference for the NP-PP ordering is a non-categorical grammatical generalization that is part of speakers' knowledge. It is arguably anti-functional, since categorical PP-NP ordering would eliminate many PP attachment ambiguities. That is, if the word order for English VPs obligatorily

---

[3] I do not know whether this holds for other varieties of English.

[4] Wasow (2002) investigated this phenomenon in the Brown corpus, using parses provided by the Linguistic Data Consortium. In a sample of over 10,000 VPs with NP and PP daughters, fewer than 700 had the PP-NP order, suggesting that the 90% figure in the text is conservative.

placed PP before NP, then an example like (9) would be unambiguous, with the PP attached to the NP.

(9) I saw the man with the telescope.

Given that the grammar of English allows both the NP-PP and the PP-NP orderings, however, the preference for NP-PP order is arguably functionally motivated. Hawkins (1994, 2004) argues that the architecture of the human parser leads to more efficient parsing when shorter phrases precede longer phrases in right-branching languages. Since PPs always contain NPs, but not vice-versa, PPs are naturally longer, on average, than NPs. Hence, the strong preference for ordering postverbal NP before PP appears to have a functional explanation.

But let us consider this argument a bit more carefully. If we assume that grammars must be categorical, then the grammar of English must either specify a fixed ordering between NP and PP, or it says nothing about the ordering. The former option would render either (8a) or (8b) ungrammatical; since both are perfectly acceptable, an account would be required of how ungrammatical strings can sound so good. So the obvious choice is to say that the grammar of English is silent on the ordering of NP and PP within the VP. An advocate of this position would then invoke the sort of performance factors Hawkins discussed to explain the statistical preference for NP-PP ordering.

Hawkins's explanation, however, makes very specific predictions. His principle of Early Immediate Constituents (EIC) says, in effect, that the preferred ordering is the one that allows the earliest possible identification of the labels on the daughters of the node being constructed. For example, the categories of the VP daughters in (10a) can be identified after the five underlined words, whereas it takes the eight underlined words in (10b) to do the same thing.

(10) a. put [on the shelf][the book we were talking about]

b. put [the book we were talking about][on the shelf]

In this way, EIC predicts a preference for the shorter constituent to precede its longer sister, unless the grammar stipulates a different ordering. So, if the grammar of English says nothing about the relative order of NP and PP within the VP, then we would predict that, in general, the shorter constituent would come first.

But this is not how these constituents are actually distributed in English. Figure 1 shows the distribution of the ordering of NP and PP daughters to VP in the Brown corpus.

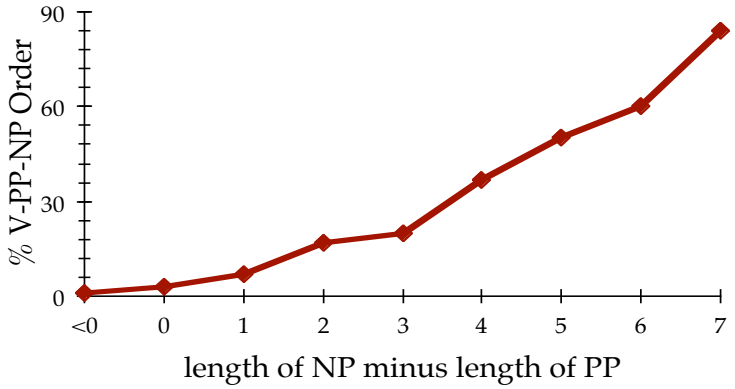


**Figure 1:** NP and PP Ordering in VPs from the Brown Corpus

The vast majority of the 10,000-plus VPs represented in this graph fall into the first two points on the graph – that is, those representing PPs that are at least as long as their NP sisters. And these are almost categorically in the canonical NP-PP order. But if the ordering were fully determined by EIC (with the grammar saying nothing about it), the curve in Figure 1 would be much higher. In particular, when the NP and PP are of equal length (0 on the horizontal axis in Figure 1), the ordering should be completely free – that is, about 50% on the vertical axis. In fact, however, the data exhibit a clear preference for the NP-PP ordering until the NP is at least five words longer than the PP.

What Figure 1 shows is that the canonical ordering of NP and PP within the English VP is NP-PP, but that this ordering is often reversed when the NP is much longer than the PP. This canonical, but violable ordering, is an obvious candidate for inclusion in the grammar of English. But it is gradient information: the preference for NP-PP ordering has a magnitude. Thus, this is a plausible case of gradience in grammar.

Of course, this constraint has a functional motivation: by adopting the NP-PP ordering as a default, the grammar usually gets the ordering preferred by the EIC (or some similar processing principle). The fact that it tends to be overridden when the length difference reaches a certain threshhold is also motivated by processing considerations. But the question I am posing is what the grammar of English says about the ordering. If the grammar must be categorical, then it either stipulates one ordering, incorrectly ruling out the other, or it says nothing about the ordering, incorrectly predicting that shorter constituents will consistently precede longer ones. The obvious solution of putting a gradient preference for short-before-long ordering into the grammar is not available to the advocate of purely categorical grammars.

## 4 Conclusion

The three examples in the previous section do not show that the grammar of English **must** include gradient information. It is unlikely that anything could,

because the location of the competence/performance boundary is so hard to pin down. The determined defender of purely categorical grammars always has the option of stipulating that any gradient information is extra-grammatical and only enters into use of language through some performance mechanism. In the absence of some principled basis for assigning information to competence or performance, however, this reduces the debate to an uninteresting terminological one.

The point of my examples was to argue that the burden of proof is on those who want to exclude gradient information from grammars. Speakers know a great deal about their languages that is not categorical, and its exclusion from grammatical theories has never been adequately justified. At a minimum, those who argue that all gradience in language should be relegated to performance should embed their grammatical theories in theories of performance that sufficiently explicit to make testable predictions.

More generally, most linguists treat the assumption that grammars are categorical as the null hypothesis; but there is no basis for this assumption. The manifestly gradient nature of so many observable linguistic phenomena naturally suggests that theories of them should include a quantitative component.

This is certainly the norm in theories of other cognitive domains, none of which (to the best of my knowledge) posits a purely categorical knowledge component interacting with gradient processing mechanisms to produce gradient observable phenomena. Rather, gradient data are taken as grounds for building gradient models. Why should language be different?

The tension between the categorical character of most generative theories and the gradience of linguistic data is one that just won't go away. Different positions on how to deal with it are linked to different empirical interests and different methodological preferences. Advocates of excluding gradience from grammar have tradition on their side, but the burden of proof is nevertheless on them.

## References


Anttila, A. 2006. Variation and opacity. *Natural Language and Linguistic Theory*. 24.893-944

Bard, E. G., D. Robertson, and A. Sorace1996. Magnitude estimation of linguistic acceptability. *Language* 72.32-68

Boersma, P. & Hayes, B. 2001. Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry* 32,45–86.

Bolinger, D. 1968. Entailment and the Meaning of Structures. *Glossa* 2.119-127

Bresnan, J. 2006. Is syntactic knowledge probabilistic? Experiments with the English dative alternation. Invited paper to appear in the proceedings of the International Conference on Linguistic Evidence, Tübingen, 2-4 February 2006, *Roots: Linguistics in search of its evidential base.* In *Studies in Generative Grammar* series, ed. by S. Featherston and W. Sternefeld. Berlin: Mouton de Gruyter

Bresnan, J, A. Cueni, T. Nikitina, and H. Baayen 2005. Predicting the Dative Alternation. KNAW Academy Colloquium: *Cognitive Foundations of Interpretation*. Amsterdam.

Bresnan, J and T. Nikitina. 2007. The Gradience of the Dative Alternation. In *Reality Exploration and Discovery: Pattern Interaction in Language and Life*, ed. by L. Uyechi and L. Hee Wee. Stanford: CSLI Publications

Chomsky, N. 1955/1975. *The Logical Structure of Linguistic Theory*. Chicago: The University of Chicago Press.

Chomsky, N. 1956. Three Models for the Description of Language. *IRE Transactions on Information Theory* 2.113-124

Chomsky, N. 1957. *Syntactic Structures*. The Hague: Mouton.

Chomsky, N. 1961. Some Methodological Remarks on Generative Grammar. *Word* 17.219-239. Reprinted as Degrees of Grammaticalness in *The Structure of Language* ed. by J. Fodor & J.J. Katz. 384-389. Englewood Cliffs, NJ: Prentice-Hall

Chomsky, N. 1966. *Syntactic Structures*. The Hague: Mouton.

Chomsky, N. 1986. *Knowledge of Language: Its Nature, Origin, and Use*. New York: Praeger

Clark, H.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.

Coetzee, A.W. 2004. *What It Means to be a Loser: Non-Optimal Candidates in Optimality Theory*. Amherst, MA: UMass dissertation.

Cowart, W. 1997. *Experimental Syntax*. Thousand Oaks, CA.: Sage Publications.

Dor, D. 2005. Toward a Semantic Account of *that*-Deletion in English. *Linguistics* 43.345-382

Garnsey, S. M., N. P. Pearlmutter, E. Myers, and M. Lotocky 1997. The Contributions of Verb Bias and Plausibility to the Comprehension of Temporarily Ambiguous Sentences. *Journal of Memory and Language*, 37.58-93.

Hawkins, J. A. 1994. *A Performance Theory of Order and Constituency*. Cambridge: Cambridge University Press.

Hawkins, J.A. 2004. *Efficiency and Complexity in Grammars*. Oxford: Oxford University Press.

Jaeger, T. F. 2006. *Redundancy and Syntactic Reduction in Spontaneous Speech*. Unpublished Stanford dissertation.

Jaeger, T. F. 2007. Usage or Gramar: Comprehension and Production Access the Same Probabilities. Talk at the Linguistic Society of America annual meeting, Anaheim.

Katz, J. J. 1964. Semi-sentences. In *The Structure of Language.* ed. by J. Fodor & J.J. Katz. 400-416. Englewood Cliffs, NJ: Prentice-Hall.

Kinsey, R., F. Jaeger, and T. Wasow 2007. What Does THAT Mean? Experimental Evidence against the Principle of No Synonymy. Talk at the Linguistic Society of America annual meeting, Anaheim.

Lavandera, B. R. 1978. Where Does the Sociolinguistic Variable Stop? *Language in Society* 7.171 – 82.

Labov, W. 1969. Contraction, Deletion, and Inherent Variability of the English Copula.. In *Language in the Inner City: Studies in Black Vernacular English*. Philadelphia: 65-129. University of Pennsylvania Press. (1972). Reprinted from *Language*. 45.715-762.

Newmeyer, F.J. 1986. *Linguistic Theory in America*, second edition. New York: Academic Press.

Newmeyer, F.J. 2003. Grammar is Grammar and Usage is Usage. *Language* 79.682-707.

Pereira, F. 2002. Formal Grammar and Information Theory: Together Again? In,*The Legacy of Zellig Harris Language and Information into the 21st century. Volume 2: Mathematics and computability of language*, ed by B.E. Nevin and S. B. Johnson. 13-32. Amsterdam: John Benjamins.

Prince, A. and P. Smolensky. 2004. *Optimality Theory: Constraint Interaction in Generative Grammar*. Oxford: Blackwell

Rickford, J.R., T. Wasow, N. Mendoza-Denton, & J. Espinoza. 1995 Syntactic variation and change in progress: Loss of the verbal coda in topic-restricting *as far as* constructions. *Language* 71.102-131

Ross, J.R. 1972. The Category Squish: Enstation Hauptwort. *Papers from the 8th Regional Meeting of the Chicago Linguistic Society*. 316-28. Department of Linguistics, University of Chicago

Ross, J.R. 1973. A Fake NP Squish. In *New Ways of Analyzing Variation in English*. ed by Bailey, C-J. & Shuy, R.W. 96-140 Washington, D.C.: Georgetown University Press.

Schütze, C.T. 1996 *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology.* Chicago: University of Chicago Press

Smolensky, P. and G. Legendre. 2006. *The Harmonic Mind: From Neural Computation ot Optimality-Theoretic Grammar*. Cambridge, MA: MIT Press.

Wasow, T. 2002 *Postverbal Behavior.* Stanford, CA: CSLI Publications.

Yaguchi, M. 2001. The Function of the Non-Deictic that in English." *Journal of Pragmatics.* 33.1125-1155.

Zipf, G. K. 1949. *Human Behaviour and the Principle of Least Effort*. Cambridge, MA: Addison-Wesley Press