Probabilities without Paradigm-Shifting: Recognizing Gradience in Natural Language Syntax

Ross Kirsling
University of Wisconsin-Madison

May 10, 2012

1 Introduction

It is no secret that natural language is gradient in nature. Indeed, the topic of gradience has been receiving a growing amount of attention within linguistics in recent years. An example of particular note is Bod et al. (2003), a collection that illustrates how gradient phenomena may be approached in various subfields of linguistics—not only psycholinguistics, sociolinguistics, and language change, but also the so-called "core" subfields of phonology, morphology, syntax, and semantics. In spite of this increased awareness, it is also quite clear that the penetration of gradience into the theoretical mainstream has been met with substantial resistance, and categorical methodology has merely begun to loosen its grasp. Yet recognizing gradience does not imply that a syntactician already working in the generative paradigm need abandon their present framework. The aim of this squib is to show that the adoption of probability theory into syntax allows for a natural incorporation of this fundamental aspect of human language, irrespective of paradigm.

2 Gradient Grammaticality

Early on in the history of transformational-generative grammar, Chomsky (1961) gave a possible account of the issue of gradient grammaticality in terms of a categorial hierarchy. At the top of this hierarchy is the all-encompassing category which might be called Word. A second level would contain preterminal categories such as N and V, while a third could contain subcategories like N_{acc} or V_{trans}. According to Chomsky's account, a sentence's "degree of grammaticalness" corresponds to the lowest level in this hierarchy at which the sentence could be represented (where full grammaticality would be the bottom-most level and full ungrammaticality the top-most). For instance, consider the following sentences with their corresponding levels of representation (assuming a hierarchy like that described here):

(1) a. She loves him. [N_{nom} V_{trans} N_{acc}]
b. *She loves he. [N V N]
c. **She loves the. [Word Word Word]

(1a) is a perfectly normal English sentence, in which a transitive verb is preceded by a nominative-marked noun and followed by an accusative-marked noun. When the accusative noun is replaced by a nominative one in (1b), however, we find an unacceptable word order for an English sentence, [N_{nom} V_{trans} N_{nom}], and need to abstract up one level in the hierarchy, giving the representation [N V N]. Since [N V N] *is* an acceptable pattern for an English sentence, this sentence is representable at the second level of the hierarchy and might be said to have a single degree of ungrammaticality (indicated here by one star). With a determiner in place of the latter noun, (1c) is still worse; even its second-level representation, [N V D], is unacceptable. Thus the only acceptable way of representing this sentence is at the top-most level, as [Word Word Word].

Unfortunately, this approach not only demands a questionable extension to the grammatical formalism, but it is also unsatisfying. With only the simple sentences in (1), one can already think of variations at levels other than that described above: 'Her love he.' is worse than (1b) but still representable as [N V N], 'The the the.' is worst of all but shares the [Word Word Word] representation of (1c), and so on. While Chomsky suggests that the hierarchy could be expanded to include a greater number of levels, it is regardless only capable of handling a few coarse-grained degrees of ungrammaticality, not arbitrarily fine-grained ones.

Pullum & Scholz (2005) point out this inadequacy as part of their proposal for a shift from generative syntax to model-theoretic syntax; however, their larger argument notwithstanding,[1] simply accounting for gradient grammaticality alone does not in itself necessitate a radical change in framework.

What is necessitated in recognizing gradience is to abandon the pursuit of generating a putative infinite set of all and only the grammatical sentences of a language.[2] As it turns out, this is not such a preposterous change. It merely requires one to expand their horizons beyond black-and-white set membership into a greyscale range of acceptability. In other words, all that is needed is to revise one's definition of grammaticality from being in the set $\{0, 1\}$ to being in the range $[0, 1]$.

In a sense, linguists have already been doing this for decades, making do with ad hoc stars and question marks (as well as hash marks) to indicate shades of ungrammaticality. Yet these symbols are a discretization of what is really a continuum. This may be good enough when considering phenomena individually, but what is truly desirable is a consistent, well-motivated way of incorporating the notion of gradience into a given grammatical formalism. Probabilities are a natural choice for this task, as the next section demonstrates.

3 Probabilistic Syntax

Given the broad usage of probability theory to address variable and uncertain phenomena in other scientific disciplines, it is not surprising that it should also find a home in theoretical

[1] Namely, they claim that the proof-theoretic production system of Emil Post which Chomsky took as an inspiration for his original syntactic theory is a well-suited approach for formal languages, but misleading when applied to natural language.

[2] Pullum & Scholz give well-reasoned independent arguments for this change in perspective which I shall not recapitulate here.

linguistics. This section makes use of probabilistic context-free grammars (PCFGs) as an illustration of how probabilities may be added into a syntactic framework. Note that the intention here is not to advocate the use of any particular framework; probabilistic variants exist of many major frameworks, including but not limited to TAG, CCG, HPSG, LFG, and OT. There are two main reasons for choosing PCFGs here. The first is their simplicity, as similar techniques can be used to add probabilities into more complex formalisms. The second is that PCFGs are a standard tool in statistical natural language processing (NLP), where their efficiency in parsing has led to their use in dealing with a wide range of syntactic phenomena (see Jurafsky & Martin (2009) ch. 14, Manning & Schütze (1999) ch. 11).[3]

Whereas a context-free (phrase structure) grammar is composed of a start node, a set of nonterminal nodes, a vocabulary, and a set of productions (phrase structure rules), in a PCFG, the productions are each augmented with probabilities.[4] For example, the various productions of a VP could have the probabilities below:

$$\begin{aligned} P(\text{V}|\text{VP}) &= 0.35 \\ P(\text{V NP}|\text{VP}) &= 0.20 \\ P(\text{V PP}|\text{VP}) &= 0.15 \\ P(\text{VP PP}|\text{VP}) &= 0.15 \\ P(\text{V NP PP}|\text{VP}) &= 0.10 \\ P(\text{V NP NP}|\text{VP}) &= 0.05 \end{aligned}$$

(Jurafsky & Martin 2009:460)

The numbers shown here are purely hypothetical, but adequate for illustrative purposes. The first line above says that the conditional probability of V given VP is 0.35. Equivalently, one could also write $P(\text{VP} \rightarrow \text{V}) = 0.35$, or that the probability of VP being rewritten as V is 0.35. Either way, the assertion is that 35% of the times a VP node is seen, it will expand into a single V node. The probability of a tree is then defined as the product of the probabilities of all the rules that together produce the tree. Necessarily, the probabilities of all rules starting from a particular node add up to 1 (as in the VP example above), and in turn, the sum of the probabilities of all sentences generable from a single PCFG is also 1.[5]

While such weights can be found relative to a given corpus, this immediately raises the question of how one could ever hope to find the exact probabilities for any such syntactic phenomenon in any language. But all is not lost, as Manning (2003:312) argues:

> Practically, all we need are reasonable estimates of probabilities, which are sufficient to give a good model of the linguistic phenomenon of interest. The difficulty of producing perfect probability estimates does not mean we are better off with no probability estimates.

[3]Specifically, the Cocke-Kasami-Younger (CKY) probabilistic parsing algorithm runs in time cubic in the length of the sentence, which is well within tractable bounds.

[4]Note that within the minimalist program, this augmentation could instead apply to individual MERGE or even MOVE/COPY operations.

[5]Actual calculations of probabilities are not essential here, but Bod (2003) is recommended as an introduction to basic probability theory with linguists specifically in mind.

As with the ad hoc stars and question marks in use by practicing linguists, what is crucial is not the precise value given as the judgment of a sentence's acceptability, but rather the relative values between the sentences under consideration. Recalling the sentences in (1), 'She loves the.' can now be explained as having a lower probability than the other sentences due to the extreme unlikelihood of a rule (or rules) expanding VP into [V D]. Generally speaking, for two sentences S_1 and S_2, the value of $P(S_1)$ might be irrelevant, but the value of $P(S_1) > P(S_2)$ is not. It is the ability to make fine-grained comparisons of this nature that makes probability theory appealing for present purposes.

Manning (2003) gives a demonstration of this ability by examining grammaticality judgments expressed by Pollard & Sag (1994). Although they claim that certain verb subcategorizations are ungrammatical, Manning finds natural-sounding examples of each within New York Times newswire text. Taking the verb *regard* as an example, he estimates a probability distribution over six possible subcategorizations, reproduced here in Figure 1, finding that the "grammatical" constructions are those that occur, roughly, in at least 1% of cases. One may argue that this is an appropriate place to draw a binary distinction, as these "ungrammatical" constructions may be saliently marginal and rare. By limiting ourselves to a categorical boundary, however, we not only ignore the fact that such subcategorizations still do occur in the wild, but also overlook the fact that the "grammatical" subcategorizations are not all created equal: almost 8 in 10 cases of the verb *regard* are in the construction

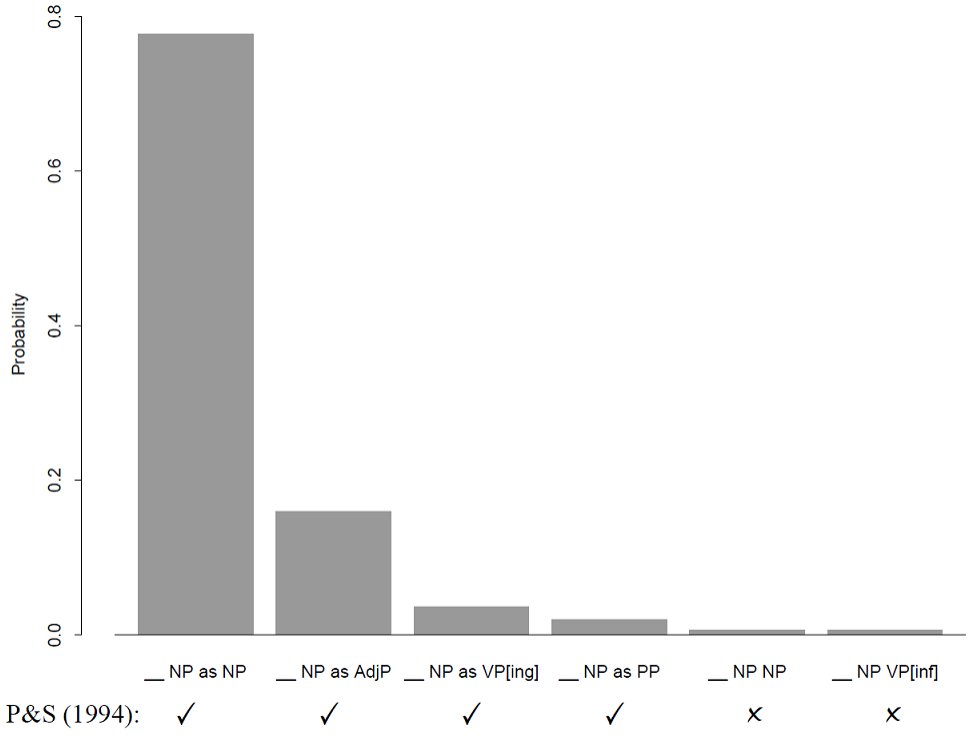

Figure 1: Estimated probability mass function (pmf) for subcategorizations of *regard* based on a 300-token New York Times corpus. The graph is annotated with grammaticality judgments of Pollard & Sag (1994). (Manning 2003:301)

regard NP *as* NP. Again, these numbers are rough estimates, but their relative values point out tendencies of potential interest.

With probabilities in tow, one is no longer limited to assertions of a language "licensing" or "prohibiting" a syntactic structure, but can also make explicit which structures the language "prefers" or "marginalizes". A weak crossover construction may be significantly more acceptable than a strong crossover construction, yet the former is still marginalized ($P(2a) < P(2b) \ll P(2c), P(2d)$).[6] Topicalization may be fully grammatical in English, but canonical word order is preferred in the absence of a suitable context, where $P(3b) > P(3a)$.

(2) a. Who$_i$ does he$_i$ love?
b. Who$_i$ does his$_i$ mother love?
c. Who$_i$ is loved by himself$_i$?
d. Who$_i$ is loved by his$_i$ mother?

(3) a. Oranges, I like.
b. I like oranges.

Statements like these help to ensure that the sentences generated (or allowed) by the grammar are faithful to the language as it is actually spoken. Most of all, having a better tool with which to distinguish various structures enables us to draw interesting conclusions from a broader spectrum of linguistic data, not merely "a shrinking subset ... increasingly removed from real usage" (Manning 2003:296).

4 Conclusion

For many decades, mainstream theoretical linguistics has been dominated by categorical methodology and unable to adequately recognize the natural gradience of human language. As this squib has shown, gradience does not have to be at odds with the existing goals of the generative (or any other) paradigm. The simple incorporation of probability theory allows for a natural means to tackle gradient phenomena. Although it requires an augmentation to the grammatical formalism, this is well justified by its ability to reflect fine-grained distinctions in grammaticality.

References

Bod, Rens. 2003. Introduction to elementary probability theory and formal stochastic language theory. In Rens Bod, Jennifer Hay & Stephanie Jannedy (eds.), *Probabilistic linguistics*, Cambridge, MA: MIT Press.

Bod, Rens, Jennifer Hay & Stephanie Jannedy (eds.). 2003. *Probabilistic linguistics*. Cambridge, MA: MIT Press.

Chomsky, Noam. 1961. Some methodological remarks on generative grammar. *Word* 17. 219–239.

[6] Such probabilities may of course differ by domain: presumably, $P(2b) > P(2d)$ in Motherese.

Jurafsky, Daniel & James H. Martin. 2009. *Speech and language processing: An introduction to natural language processing, speech recognition, and computational linguistics*. Prentice-Hall 2nd edn.

Manning, Christopher D. 2003. Probabilistic syntax. In Rens Bod, Jennifer Hay & Stephanie Jannedy (eds.), *Probabilistic linguistics*, Cambridge, MA: MIT Press.

Manning, Christopher D. & Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Pollard, Carl & Ivan Sag. 1994. *Head-driven phrase structure grammar*. Chicago: Chicago University Press.

Pullum, Geoffrey K. & Barbara C. Scholz. 2005. Contrasting applications of logic in natural language syntactic description. In Petr Hájek, Luis Valdés-Villanueva & Dag Westerståhl (eds.), *Logic, methodology and philosophy of science: Proceedings of the twelfth international congress*, 481–503. London: King's College Publications.