Draft. To appear in Bod, Hay and Jannedy (eds), Probabilistic Linguistics, MIT Press

Probabilistic Syntax

Christopher D. Manning
Departments of Linguistics and Computer Science, Stanford University
http://nlp.stanford.edu/~manning/
manning@cs.stanford.edu

January 12, 2002

1 The Tradition of Categoricity and Prospects for Stochasticity

"Everyone knows that language is variable." This is the bald sentence with which Sapir (1921:147) begins his chapter on language as an historical product. He goes on to emphasize how two speakers' usage is bound to differ "in choice of words, in sentence structure, in the relative frequency with which particular forms or combinations of words are used". I should add that much sociolinguistic and historical linguistic research has shown that the *same* speaker's usage is also variable (Labov 1966, Kroch 2001:722). However, the tradition of most syntacticians has been to ignore this thing that everyone knows.[1]

Human languages are the prototypical example of a symbolic system. From very early on, logics and logical reasoning were invented for handling natural language understanding. Logics and formal languages have a language-like form that draws from and meshes well with natural languages. It is not immediately obvious where the continuous and quantitative aspects of syntax are. The dominant answer in syntactic theory has been "nowhere" (Chomsky 1969:57; also 1956, 1957, etc.): "It must be recognized that the notion 'probability of a sentence' is an entirely useless one, under any known interpretation of this term."

In the 1950s there were prospects for probabilistic methods taking hold in linguistics, in part due to the influence of the new field of Information Theory (Shannon 1948).[2] Chomsky's influential remarks had the effect of killing off interest in probabilistic methods for syntax, just as for a long time McCarthy and Hayes (1969) discouraged exploration of probabilistic methods in Artificial Intelligence. Among his arguments were that: (i) Probabilistic models wrongly mix in world knowledge (New York occurs more in text than Dayton, Ohio, but for no linguistic reason), (ii) Probabilistic models don't model grammaticality (neither *Colorless green ideas sleep furiously* nor *Furiously sleep ideas green colorless* have previously been uttered – and hence must be estimated to have probability zero, Chomsky wrongly assumes – but the former is grammatical while the latter is not, and (iii) Use of probabilities does not meet the goal of describing the mind-internal I-language as opposed to the observed-in-the-world E-language. This chapter is not meant to be a detailed critique of Chomsky's arguments – Abney (1996) provides a survey and a rebuttal, and Pereira (2000) has further useful discussion – but some of these concerns are still important to discuss. I

[1]My thanks to Bas Aarts, Farrell Ackerman, Harald Baayen, Rens Bod, Joan Bresnan, Ariel Cohen, Shipra Dingare, Mark Johnson, and Dan Klein for comments on a draft of this chapter, or discussions of issues related to it. This chapter draws analyses and insight from joint research with Joan Bresnan and Shipra Dingare. In particular, although the presentation is different, the central example in the latter part of the paper is drawn from Bresnan et al. (2001). Taking linguistic usage seriously requires paying more attention to sociolinguistic variation than a syntactician traditionally did: I have found Chambers (1995) very helpful, and a few ideas and quotes are borrowed from that work. Parts of the research presented here were supported by the National Science Foundation under Grant No. BCS-9818077.

[2]For instance, Gleason (1961), a standard text of the period, devotes a chapter to a nontechnical introduction to information theory and suggests that it is likely to have a big impact on linguistics.

argue in section 3.2 that in retrospect none of Chomsky's objections actually damn the probabilistic syntax enterprise.

Chambers (1995:25–33) has one of the few clear discussions of this "Tradition of Categoricity" in linguistics of which I am aware. He makes a distinction between standardly used linguistic units which are *discrete* and *qualitative* versus an alternative that allows units which are *continuous* and *quantitative*. He discusses how the idea that linguistics should keep to a categorical base precedes modern generative grammar. It was standard among American structuralists, and is stated most firmly by Joos (1950:701–702):

> Ordinary mathematical techniques fall mostly into two classes, the continuous (e.g., the infinitesimal calculus) and the discrete or discontinuous (e.g., finite group theory). Now it will turn out that the mathematics called "linguistics" belongs to the second class. It does not even make any compromise with continuity as statistics does, or infinite-group theory. Linguistics is a quantum mechanics in the most extreme sense. All continuities, all possibilities of infinitesimal gradation, are shoved outside of linguistics in one direction or the other.

Modern linguistics is often viewed as a cognitive science. Within cognitive science in general, it is increasingly understood that there is a central role for the use of probabilistic methods in modeling and understanding human cognition (for vision, concept learning, and so on – inter alia, Kersten 1999, Tenenbaum 1999). Indeed, in many areas, probabilistic modeling has become so prevalent that Mumford (1999) has seen fit to declare the Dawning of an Age of Stochasticity. Human cognition has a probabilistic nature: we continually have to reason from incomplete and uncertain information about the world, and probabilities give us a well-founded tool for doing this.

Language understanding is a subcase of this: When someone says "it's cold in here", in some circumstances I'm understanding them correctly if I interpret that utterance as a request to close the window. Ambiguity and underspecification are ubiquitous in human language utterances, at all levels (lexical, syntactic, semantic, etc.), and how to resolve these ambiguities is a key communicative task for both human and computer natural language understanding. At the highest level, the probabilistic approach to natural language understanding is to view the task as trying to learn the probability distribution $P(meaning|utterance, context)$ – a mapping from form to meaning conditioned by context. In recent years, such probabilistic approaches have become dominant within computational linguistics (Manning and Schütze 1999), and are becoming increasingly used within psycholinguistics (Jurafsky, this volume; Baayen, this volume). Probabilistic methods have largely replaced earlier computational approaches because of the more powerful evidence combination and reasoning facilities that they provide. This greater power comes from the fact that knowing how likely a module thinks a certain sense, structure, or interpretation is is much more useful and powerful information than just knowing whether it is deemed possible or impossible. Language acquisition is also a likely place for probabilistic reasoning: Children are necessarily doing uncertain analytical reasoning over uncertain input. But is the same true for the core task of describing the syntax – the grammar – of a human language?

This chapter will advocate a "yes" answer. However, so far, syntax has not seen the dawn. Probabilistic syntax (or, equivalently, stochastic syntax) has been a little studied area, and there is not yet a large, coherent body of work applying sophisticated probabilistic models in syntax. Given the great promise of probabilistic techniques, the reader should see this as an invitation to get involved on the ground floor of an exciting new approach. This chapter will attempt to explain why probabilistic models have great value and promise for syntactic theory, and to give a few small examples and connections to relevant literature. Motivation for noncategorical models in syntax comes from language acquisition, historical change, and typological and sociolinguistic variation, and I touch on those motivations briefly, but only briefly since they are dealt with further in other chapters. Throughout, the aim is to examine probabilistic models for explaining language structure, as opposed to simply using techniques like significance tests to buttress empirical results.

2 The joys and perils of corpus linguistics

2.1 On the trail of a neat fact

Halfway through a long overseas plane flight, I settled down to read a novel (Russo 2001). However, I only made it to the third page before my linguist brain reacted to the sentence:

(1) By the time their son was born, though, Honus Whiting was beginning to understand and privately share his wife's opinion, as least as it pertained to Empire Falls.

Did you notice anything abnormal? It was the construction ***as** least as* that attracted my attention: For my categorical syntactician self, this construction is simply ungrammatical. I should add the rider "in my idiolect", but in just this manner, despite what is said in introductory linguistics classes, modern linguistics has become highly prescriptive, or at least has come to use "normative edited texts" – a term Chambers (1995:27) attributes to Labov.

For my corpus linguist self, this sentence is a piece of uncontrovertible primary data. This status for data is importantly different from the stance of Chomsky (1965:4) that "observed use of language ... may provide evidence ... but surely cannot constitute the subject-matter of linguistics, if this is to be a serious discipline". But there remains the question of how to interpret this datum, and, here, the distinction of Chomsky (1986) between externally visible E-language, and the internalized I-language of the mind still seems important. One possibility, which seemed the most probable one to a linguist friend sitting in the next seat, is that this is just a typo, or some similar result of editing. Or it might have been a "speech error" which somehow survived into print, but which the author would judge as ungrammatical if asked. Or it could be a matter of dialectal, sociolectal, or just idiolectal variation. In general one cannot make a highly confident decision between these and other choices. In this case one could conceivably have telephoned, but in general that option is not available, and at any rate sociolinguists have long observed that people's actual language use deviates from their reports of how they use language. This unclarity of attribution means that corpus linguistics necessarily has a statistical basis: one has to reason about the likelihood of different explanations based on both the frequency with which different forms occur and prior probabilistic models of language and other aspects of cognition and the world.

Back on land, the obvious way to address this question, within a scientific empirical linguistics, is to collect more evidence. A search of 1994 *New York Times* newswire yields no examples – is this just because they have better copy editors? – but then I find 4 examples including (2a–b) in 1995 *New York Times* newswire, and 2 further examples from 1996. It already appears less likely that this was just a typo. A search on the web (which with every passing year becomes a better medium for empirical linguistic research) yields hundreds of further examples. There are ones (apparently) from American east coast speakers (2c), from midwestern speakers (2d), and from west coast speakers (2e). There are ones from college professors (2c) and from fishing boat owners (2f). I can't even regard this as an American aberration (I'm Australian): I find examples from Australia (2g) and South Africa (2h). Finally, I find examples with *at least as* in a neighboring sentence (2i–j), showing intraspeaker variation. While much less common than *at least as* (perhaps about 175 times less common, based on this small study), *as least as* seems to have robust support and perhaps indicates a development within the language (there were several examples in discussions of the 2000 U.S. election, but there was also one citation from 1972 – more research would be needed to establish this). Unlike the initial example from Russo, many – but not all – of these additional examples use *as least as* in conjunction with an *as* Adj *as* construction, which perhaps provides a pathway for the development of this apparently new form.[3]

[3]All italics in the cited examples are mine. Sources: a. NYT newswire, 1995-09-01 article by Dave Ahearn quoting Alan Greenspan; b. NYT newswire, 1995-11-29; c. John McHale, Economics 1415, Reform of the Public Sector, Fall 1999, http:

(2) a. Indeed, the will and the means to follow through are *as least as* important *as* the initial commitment to deficit reduction.

b. Steven P. Jobs has reemerged as a high-technology captain of industry, *as least as* far *as* the stock market is concerned.

c. Alternatively, y is preferred to x if, in state x, it is not possible to carry out hypothetical lump-sum redistribution so that everyone could be made *as least as* well off *as* in y.

d. There is *as least as* much investment capital available in the Midwest *as* there is on either Coast.

e. The strip shall be of a material that is *as least as* slip-resistant *as* the other treads of the stair.

f. As many of you know he had his boat built at the same time as mine and it's *as least as* well maintained and equipped.

g. There is a history of using numbers to write music that dates *as least as* far back to Pythagoras.

h. He added: "The legislation is *as least as* strict *as* the world's strictest, if not the strictest."

i. Black voters also turned out *at least as* well *as* they did in 1996, if not better in some regions, including the South, according to exit polls. Gore was doing *as least as* well among black voters *as* President Clinton did that year.

j. Second, if the required disclosures are made by on-screen notice, the disclosure of the vendor's legal name and address must appear on one of several specified screens on the vendor's electronic site and must be *at least as* legible and set in a font *as least as* large *as* the text of the offer itself.

Thus we have discovered a neat fact about English lexicogrammar – previously unremarked on as far as I am aware – and we have at least some suspicions about its origin and range of use, which we could hope to confirm with further corpus-based analysis. And this has been quite fun to do: there was the chance discovery of a "smoking gun" followed by "text mining" to discover further instances. However, a few observations are immediately in order.[4]

2.2 The important lessons

First off, this example was easy to investigate because the phenomenon is rooted in particular lexical items. It's easy to search text corpora for *as least as* constructions; it is far harder to search corpora for something like participial relative clauses or locative inversion constructions. Such technological limitations have meant that corpus linguistic research has been largely limited to phenomena that can be accessed via searches on particular words. The average corpus linguist's main research tool remains the word concordancing program which shows a searched for keyword in context (perhaps with morphological stemming,

//icg.harvard.edu/~ec1415/lecture/lecturenote4.pdf; d. Ed Zimmer, http://tenonline.org/art/9506.html; e. State of California, Uniform Building Code, Title 24, Section 3306(r), http://www.johnsonite.com/techdata/section2/TITLE24C.HTM; f. Ron Downing (Alaskan 2nd generation halibut fishing boat captain), March 10, 1998, http://www.alaska.net/~gusto/goodbye.txt; g. Matthew Hindson (composer) catalogue notes, http://members.ozemail.com.au/~mhindson/catalogue/pnotes-pi.html; h. Sunday Business Times, South Africa, http://www.btimes.co.za/99/0425/comp/comp04.htm; i. Leigh Strope, Associated Press, 2000-11-07, http://www.theindependent.com/stories/110700/ele_turnout07.html; j. Jeffrey M. Reisner, http://www.allcities.org/Articles/Ar_Reisner.htm.

[4] See also Fillmore (1992) for an amusing account of the strengths and weaknesses of corpus linguistics.

sorting options, etc.).[5] However, a (theoretical) syntactician is usually interested in more abstract structural properties that cannot be investigated easily in this way. Fortunately for such research, recent intensive work in Statistical Natural Language Processing (Statistical NLP; Manning and Schütze 1999) has led to the availability of both a lot of more richly annotated text corpora which can support such deeper investigations, and a lot more tools capable of being used with good accuracy over unannotated text in order to recover deep syntactic relationships. The prospects for applying these corpora and tools for probabilistic syntactic analysis are bright, but it remains fair to say that these tools have not yet made the transition to the Ordinary Working Linguist without considerable computer skills.[6]

Secondly, one needs a large corpus to be able to do interesting exploration of most syntactic questions. Many syntactic phenomena, especially those commonly of interest to theoretical syntacticians, are just incredibly rare in text corpora. For instance, in the *Wall Street Journal* newswire corpus mentioned above, I searched through about 230 million words of text to find a paltry six examples of *as least as*. This is one aspect of the problem that Sinclair (1997:27) sums up as: "The linguistics of the twentieth century has been the linguistics of scarcity of evidence." All approaches to linguistics have dealt with this problem in one way or another. Generative approaches resort to inventing the primary data, on the basis of intuitions. I will not provide a detailed discussion of the possible limitations of such an approach here. Suffice it to say that even Chomsky (1986:63) admits that "the facts are often quite obscure".[7] A traditional field linguist has been constrained by a quite small collection of texts and conversations. While the major syntactic features will be clear, there is generally quite insufficient data for the kind of subtle issues that are the mainstay of theoretical syntax. Similarly for sociolinguists, the data they collect by traditional in-person techniques are rich in social and contextual information, but poor in quantity. It is probably for this reason that the vast majority of sociolinguistic work has dealt with phonetic realization (for which the data is dense), and the small amount of syntactic work has been mainly on phenomena such as copula deletion and use of modals where again the density of the use of function words makes analysis from a small corpus possible. In contrast, corpus linguists have tended to work with large amounts of data in broad pseudo-genres like "newspaper text", collapsing together the writing of many people from many places. There is no easy answer to the problem of getting sufficient data of just the right type: language changes across time, space, social class, method of elicitation, etc. There is no way that we can collect a huge quantity of data (or at least a collection dense in the phenomenon of current interest) unless we are temporarily prepared to ride roughshod over at least one of these dimensions of variation. To deal with rare syntactic phenomena, we must either work with intuitions or else be prepared to aggregate over speakers and time periods. Even then, the *sparsity* of linguistic data is nearly always a major technical challenge in probabilistic approaches to linguistics or NLP.

Finally – and most seriously – whatever their interest, the above observations on *as least as* unquestionably fit into the category of activities that Chomsky (1979:57) long ago derided as butterfly collecting: "You can also collect butterflies and make many observations. If you like butterflies, that's fine; but such work must not be confounded with research, which is concerned to discover explanatory principles of some depth and fails if it does not do so." A central question is whether probabilistic models can be used for linguistic

[5] See, for instance, the research methods discussed in McEnery and Wilson (2001). For Windows computers, Mike Scott's Wordsmith Tools http://www.liv.ac.uk/~ms2928/wordsmith/ is a leading recent example of a corpus linguistics tool based around concordances.

[6] The parsed sentences in the Penn Treebank (Marcus et al. 1993) provide one useful source of data for complex syntactic queries (used, for example, by Wasow (1997) and Bresnan et al. (2001)). But the tgrep access software supplied with some versions is limited to Unix, and has not been actively maintained (though see http://www-2.cs.cmu.edu/~dr/Tgrep2/ for an improved Tgrep2 program – also for Unix/Linux). The recently released ICE-GB corpus comes with a friendly graphical tool for Windows, ICECUP, which allows searching of parse trees (Wallis et al. 1999, Nelson et al. forthcoming). See http://nlp.stanford.edu/links/statnlp.html for a listing of corpora and NLP tools.

[7] Recently considerably more attention has been paid to the reliability of intuitive judgements, and good methodologies for gathering them (Schütze 1996, Cowart 1997), but most linguists in practice pay little attention to this area.

explanation and insight in this manner. I think this is a serious concern. To go out on a limb for a moment, let me state my view: generative grammar has produced many explanatory hypotheses of considerable depth, but is increasingly failing because its hypotheses are disconnected from verifiable linguistic data. Issues of frequency of usage are by design made external to matters of syntax, and as a result categorical judgements are overused where not appropriate, while a lack of concern for observational adequacy has meant that successive versions have tended to treat a shrinking subset of data increasingly removed from real usage. On the other side, corpus linguistics (McEnery and Wilson 2001) – or "usage-based models of grammar" (Barlow and Kemmer 2000) – has all the right rhetoric about being an objective, falsifiable empirical science interested in the totality of language use, but is failing by largely restricting itself to surface facts of language, rather than utilizing sophisticated formal models of grammar, which make extensive use of *hidden structure* (things like phrase structure trees, and other abstract representational levels). This reflects an old dichotomy: one sees it clearly in the 1960s *Handbook of Mathematical Psychology* (Luce et al. 1963) where in some chapters probabilistic finite state models (Markov chains) are being actively used to record surface-visible stimulus-response behavior, whereas in another chapter Chomsky is arguing for richer tools (then, transformational grammars) for attacking deeper analytical problems. The aim of the current chapter is to indicate a path beyond this impasse: to show how probabilistic models can be combined with sophisticated linguistic theories for the purpose of syntactic explanation.

Formal linguistics has traditionally equated structuredness with homogeneity (Weinreich et al. 1968:101), and has tried too hard to maintain categoricity by such devices as appeal to an idealized speaker/hearer. I would join Weinreich et al. (1968:99) in hoping that "a model of language which accommodates the facts of variable usage ... leads to more adequate descriptions of linguistic competence." The motivation for probabilistic models in syntax comes from two sides:

- Categorical linguistic theories claim too much. They place a hard categorical boundary of grammaticality where really there is a fuzzy edge, determined by many conflicting constraints and issues of conventionality vs. human creativity. This is illustrated with an example in section 3.1.
- Categorical linguistic theories explain too little. They say nothing at all about the soft constraints which explain how people choose to say things (or how they choose to understand them). The virtues of probabilistic models for this task are developed in section 5.

The first problem is a foundational worry, but the second is more interesting in showing the possibility for probabilistic models to give increased linguistic explanation.

3 Probabilistic syntactic models

As an example of the non-categorical nature of syntax, and the idealization that occurs in most of the syntactic literature, I will look here at verbal clausal subcategorization frames. We will first look at problems with categorical accounts, then explore how one might build a probabilistic model of subcategorization, and then finally consider some general issues in the use of probabilistic approaches within syntax.

3.1 Verbal subcategorization

It is well-known that different verbs occur with different patterns of arguments which are conventionally described in syntax by subcategorization frames (Chomsky 1965, Radford 1988). For example, some verbs must take objects while others do not:[8]

[8] In English, most verbs can be used either transitively or intransitively, and in introductory syntax classes we grope for examples of verbs that are purely transitive or intransitive, using verbs such as the examples in (3). Linguistic examples are standardly

(3) a. Kim devoured the meal.

b. *Kim devoured.

c. *Dana's fist quivered Kim's lip.

d. Kim's lip quivered.

Other verbs take various forms of sentential complements.

A central notion of all current formal theories of grammar (including Government-Binding (Chomsky 1981), the Minimalist Program (Chomsky 1995), Lexical-Functional Grammar (Bresnan 2001), Head-driven Phrase Structure Grammar (Pollard and Sag 1994), Categorial Grammar (Morrill 1994), Tree-Adjoining Grammar (Joshi and Schabes 1997), and other frameworks) is a distinction between arguments and adjuncts combining with a head. Arguments are taken to be syntactically specified and required by the head, via some kind of subcategorization frame or argument list, whereas adjuncts (of time, place, etc.) can freely modify a head, subject only to semantic compatibility constraints.

This conception of the argument/adjunct distinction is the best one can do in the categorical 0/1 world of traditional formal grammars: things have to be either selected (as arguments) or not. If they are not, they are freely licensed as adjuncts, which in theory should be able to appear with any head, and to be iterated any number of times, subject only to semantic compatibility. And there is certainly some basis for the distinction: a rich literature (reviewed in Radford 1988, Pollard and Sag 1987, Schütze 1995) has demonstrated argument/adjunct asymmetries in many syntactic domains, such as with phrasal ordering and extraction possibilities.

However, categorical models of selection have always been problematic. The general problem with this kind of model was noticed early on by Sapir (1921:38) who noted that "All grammars leak". In context, language is used more flexibly than such a model suggests. Even Chomsky in early work (1955:131) notes that "an adequate linguistic theory will have to recognize degrees of grammaticalness". Many subcategorization distinctions presented in the linguistics literature as categorical are actually counterexemplified in studies of large corpora of written language use. For example, Atkins and Levin (1995) find examples of *quiver* used transitively, such as *The bird sat, quivering its wings*. In this section, I will look at some examples of clausal complements, focusing particularly on what realizations of arguments are licensed, and then touch on the argument/adjunct distinction again at the end.

Some verbs take various kinds of sentential complements, and this can be described via subcategorization frames:

(4) *regard*: __ NP[acc] *as* {NP, AdjP}
consider: __ NP[acc] {AdjP, NP, VP[inf]}
think: { __ CP[that], __ NP[acc] NP }

Below we give the grammaticality judgements for subcategorizations given by Pollard and Sag (1994:105–108) within the context of an argument that verbs must be able to select the category of their complements.[9] We begin with *consider*, which they say appears with a noun phrase object followed by various kinds of predicative complements (nouns, adjectives, clauses, etc.), but not with *as*-complements:

informally rated for exclusively syntactic well-formedness on a scale: * (ungrammatical) > ?* > ?? > ? (questionable) > unmarked (grammatical). I use a # to indicate a sentence that is somehow deviant, without distinguishing syntactic, semantic, or contextual factors.

[9] It is certainly not my aim here to single out Pollard and Sag for reproach. I choose them as a source simply because they actually provide clear testable claims, as part of a commitment to broad observational adequacy. This has become increasingly rare in generative syntax, and I would not want to give the impression that another theory that assumes a categorical model of argument frames, but does not provide any examples of them, is superior.

(5) a. We consider Kim to be an acceptable candidate

b. We consider Kim an acceptable candidate

c. We consider Kim quite acceptable

d. We consider Kim among the most acceptable candidates

e. *We consider Kim as an acceptable candidate

f. *We consider Kim as quite acceptable

g. *We consider Kim as among the most acceptable candidates

h. ?*We consider Kim as being among the most acceptable candidates

However, this lack of *as*-complements is counterexemplified by various examples from the *New York Times*:

(6) a. The boys consider her as family and she participates in everything we do.

b. Greenspan said, “I don’t consider it as something that gives me great concern.”

c. “We consider that as part of the job,” Keep said.

d. Although the Raiders missed the playoffs for the second time in the past three seasons, he said he considers them as having championship potential.

e. Culturally, the Croats consider themselves as belonging to the “civilized” West, . . .

If this was an isolated incident (counterexemplifying (5e) and (5h) in particular), then we would merely have collected one more butterfly, and would conclude that Pollard and Sag got that one particular fact wrong. But the problem is much more endemic than that. The subcategorization facts that they give can in general be counterexemplified. Since this is an important point in the argument for probability distributions in syntax, let me give a few more examples.

According to Pollard and Sag (1994) – and generally accepted linguistic wisdom – *regard* is the opposite of *consider* in disallowing infinitival VP complements, but allowing *as* complements:

(7) a. *We regard Kim to be an acceptable candidate

b. We regard Kim as an acceptable candidate

But again we find examples in the *New York Times* where this time *regard* appears with an infinitival VP complement:

(8) a. As 70 to 80 percent of the cost of blood tests, like prescriptions, is paid for by the state, neither physicians nor patients regard expense to be a consideration.

b. Conservatives argue that the Bible regards homosexuality to be a sin.

Pollard and Sag (1994) describe *turn out* as allowing an adjectival phrase complement but not a present participle VP complement:

(9) a. Kim turned out political

b. *Kim turned out doing all the work

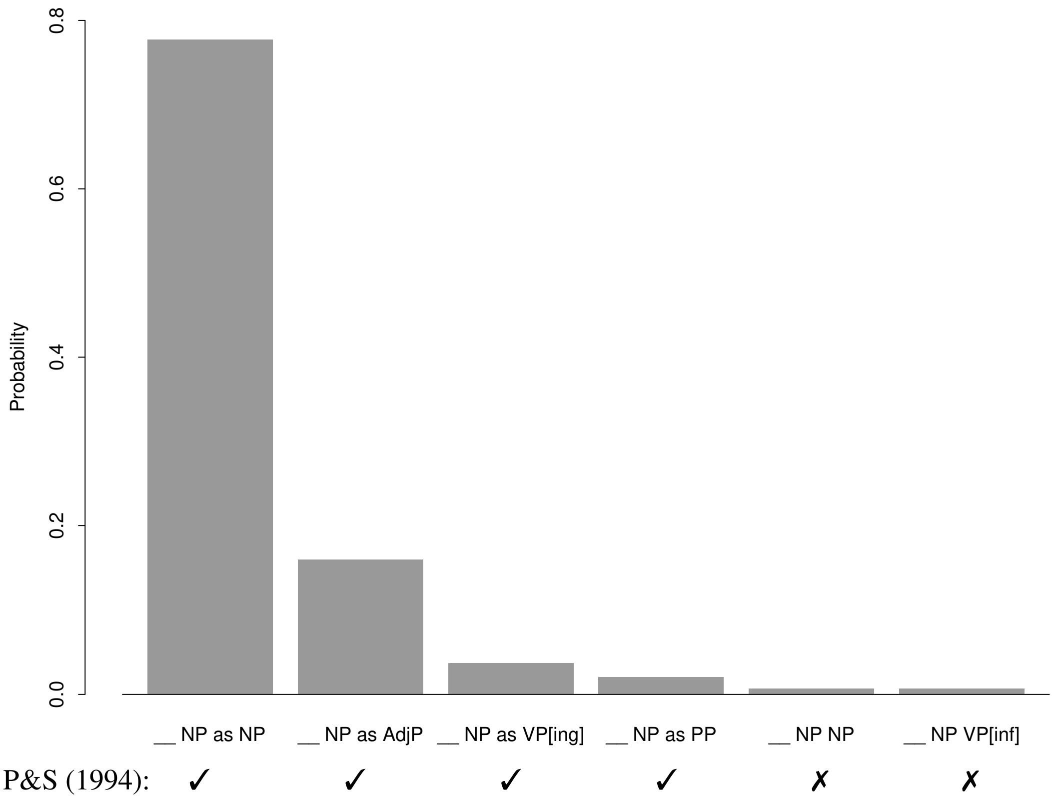

Figure 1: Estimated probability mass function (pmf) for subcategorizations of *regard* based on a 300 token *New York Times* corpus.

But again we find counterexamples in the *New York Times*:

(10) But it turned out having a greater impact than any of us dreamed.

Similarly, *end up* is predicted to disallow a perfect participle VP complement:

(11) a. Kim ended up political

b. *Kim ended up sent more and more leaflets

but an example of the sort predicted as ungrammatical again appears in *The New York Times*:

(12) On the big night, Horatio ended up flattened on the ground like a fried egg with the yolk broken.

What is going on here? Pollard and Sag's (1994) judgements seem reasonable when looking at the somewhat stilted "linguist's sentences". But with richer content and context, the *New York Times* examples sound (in my opinion) in the range between quite good and perfect. None of them would make me choke on my morning granola. They in all likelihood made it past a copy editor. While there are the usual considerations of allowing for the possibility of errors or regional or social variation, I think it is fair to say that such explanations are not terribly plausible here.

The important question is how we should solve this problem: Within a categorical linguistics there is no choice but to say that the previous model was overly restrictive, and that these other subcategorization frames should also be admitted for the verbs in question. But if we do that, we lose a lot. For we are totally failing to capture the fact that the subcategorization frames which Pollard and Sag do not recognize are extremely rare, whereas the ones they give encompass the common subcategorization frames of the verbs in question. We can get a much better picture of what is going on by estimating a probability mass function (pmf) over the subcategorization patterns for the verbs in question. A pmf over a discrete random variable (a set of disjoint choices) gives the probability of each one. Here, we will estimate a pmf for the verb *regard*. The subcategorization pmf in figure 1 was estimated from 300 tokens of *regard* from the *New York Times*. I have simply counted how often each subcategorization occurred in this sample, and then graphed these frequencies divided by the number of tokens – so as to give the maximum likelihood point estimate for each frame.[10]

I have annotated the graph with the grammaticality judgements of Pollard and Sag (1994). We note two things: first that the line of grammaticality has apparently been drawn at a fairly arbitrary point of low frequency. Roughly, things that occur at least 1 time in 100 are regarded as grammatical, while things that occur less commonly are regarded as ungrammatical. If we had set the cutoff as 1 time in 10, then only the *as* NP and *as* AdjP complementation patterns would be "grammatical"; if we had set the cutoff at 1 time in 1000, then they would all be grammatical.[11] Secondly, note how much important information is lost by simply providing categorical information. In a categorical description, there is no record whatsoever of the fact that over three-quarters of all predicative complements of *regard* are in the form of an *as* NP complement. A probability distribution captures such information in a straightforward manner.

Looking more broadly, careful consideration of the assumed categorical distinction between arguments and adjuncts further argues for new foundations for syntax. There are some very clear arguments (normally, subjects and objects), and some very clear adjuncts (of time and 'outer' location), but also a lot of stuff in the middle. Things in this middle ground are often classified back and forth as arguments or adjuncts depending on the theoretical needs and convenience of the author (Babby 1980, Fowler 1987, Maling 1989, Maling 1993, Li 1990, Wechsler and Lee 1996, Przepiórkowski 1999a, Przepiórkowski 1999b). Additionally, one sees the postulation of various in-between categories, such as argument-adjuncts (Grimshaw 1990) and pseudo-complements (Verspoor 1997).

In a probabilistic approach, in contrast, it is not necessary to categorically divide verbal dependents into subcategorized arguments and freely occurring adjuncts. Rather, we can put a probability function over the kinds of dependents to expect with a verb or class, conditioned on various features. This is especially useful in difficult in-between cases. For example, for the verb *retire*, the subcategorization frames listed in the *Oxford Advanced Learners Dictionary* (Hornby 1989) are as a simple intransitive and transitive verb, and as intransitives taking a PP[*from*] or PP[*to*] argument.[12] While prepositional phrases headed by *to* or *from* are common with *retire* (13a–b) – and are arguably arguments by traditional criteria – this does not exhaust the list of putative arguments of *retire*. While *in* most often occurs with *retire* to specify a time point (a canonical adjunct PP), it sometimes occurs expressing a destination (13c), and it seems that these examples demand the same treatment as examples with PP[*to*]. Similarly, uses of a PP[*as*] to express the position that one is retiring from (13d) seem at least as selected by the verb as those expressed by a PP[*to*] or PP[*from*].

[10] The exact numbers in this graph should be approached with caution, but I hope it is adequate to illustrate the general point. It would be necessary to do larger counts to get accurate probability estimates of rare patterns. Also, as is discussed in Roland and Jurafsky (1998), one should note that the frequency of different subcategorization patterns varies greatly with the genre of the text. In this graph forms with *as* followed by a verbal past/passive participle were collapsed together with *as* plus adjective, and passive examples were grouped with the subcategorization for the corresponding active form.

[11] Sampson (1987, 2001: ch. 11) also argues against a grammatical/ungrammatical distinction on the basis of a cline of commonness, though I think his evidence is rather less compelling (see Taylor et al. (1989) and Culy (1998)).

[12] The *OALD* is one of several dictionaries that attempts to list subcategorization information (in a traditional form). Other sources of wide coverage subcategorization information include resources constructed for NLP, such as COMLEX (Grishman et al. 1994). The discussion in this section expands the discussion of *retire* in Manning (1993).

What about the usage of *retire on* (13e), where the PP[*on*] expresses the source of monetary support?[13]

(13) a. Mr. Riley plans to retire to the $1.5 million ranch he is building in Cody, Wyo.

b. Mr. Frey, 64 years old, remains chairman but plans to retire from that post in May.

c. To all those wishing to retire in Mexico (international section, March 10 and 11), let me offer three suggestions:

d. Donald W. Tanselle, 62 years old, will retire as vice chairman of this banking concern, effective Jan. 31.

e. A worker contributing 10% of his earnings to an investment fund for 40 years will be able to retire on a pension equal to two thirds of his salary

Rather than maintaining a categorical argument/adjunct distinction and having to make in/out decisions about such cases, we might instead try to represent subcategorization information as a probability distribution over argument frames, with different verbal dependents expected to occur with a verb with a certain probability. For instance (sticking for the moment to active uses, and assuming that controlled and imperative subjects are present, etc.), we might estimate that:[14]

$$\begin{aligned} P(\mathrm{NP}[\textsc{subj}]|V = \textit{retire}) &= 1.0 \\ P(\mathrm{NP}[\textsc{obj}]|V = \textit{retire}) &= 0.52 \\ P(\mathrm{PP}[\textit{from}]|V = \textit{retire}) &= 0.05 \\ P(\mathrm{PP}[\textit{as}]|V = \textit{retire}) &= 0.06 \\ &\vdots \end{aligned}$$

By writing things like this, I am implicitly assuming independence between arguments (cf. Bod, this volume): the chance of getting a PP[*as*] is independent of whether a PP[*from*] is present or not. This is presumably not true: intuition suggests that having either a PP[*as*] or a PP[*from*] makes having the other less likely. We could choose instead to provide joint estimates of a complete subcategorization frame:

$$\begin{aligned} P(\mathrm{NP}[\textsc{subj}]\ \_\_\ |V = \textit{retire}) &= 0.25 \\ P(\mathrm{NP}[\textsc{subj}]\ \_\_\ \mathrm{NP}[\textsc{obj}]|V = \textit{retire}) &= 0.5 \\ P(\mathrm{NP}[\textsc{subj}]\ \_\_\ \mathrm{PP}[\textit{from}]|V = \textit{retire}) &= 0.04 \\ P(\mathrm{NP}[\textsc{subj}]\ \_\_\ \mathrm{PP}[\textit{from}]\ \mathrm{PP}[\textit{after}]|V = \textit{retire}) &= 0.003 \\ &\vdots \end{aligned}$$

In this latter case the sums of the probabilities of all frames would add to one (in the former it is the probabilities of whether to have a certain argument or not that add to one).

Regardless of details of the particular modeling choices, such a change of approach reshapes what questions can be asked and what results are achievable in terms of language description, learnability, production, and comprehension. Moreover, the probabilistic approach opens up possibilities for renewed productive interplay between formal syntax and various areas of applied linguistics, including computational linguistics.

[13] All italics mine. These examples are from Linguistic Data Consortium *Wall Street Journal* newswire: a. 1994/11/23, b. 1994 /11/20, c. 1987/04/07, d. 1986/12/08, e. 1994/12/12.

[14] These are rough estimates from 1987 *WSJ*. Recall footnote 10 – in particular the extremely high frequency of transitive uses of *retire* would be unusual in many genres, but companies often retire debt in the *WSJ*.

A language teacher is likely to be interested in how a certain semantic type of argument is most commonly expressed, or just with which frames a verb is most commonly used. While they incorporate a number of further notions, modern statistical NLP parsers (Collins 1997, Charniak 1997) contain generative probabilistic models of surface subcategorization, much like the above kind of lexicalized dependency information. Rather than using simple PCFG rules like VP → V NP PP of the sort discussed in Bod (this volume), these generative models use lexicalized argument frames by modeling the probability that a VP is headed by a certain verb, and then the probability of certain arguments surrounding that verb:[15]

$$P(\mathrm{VP} \rightarrow \mathrm{V}[\textit{retire}]\ \ \mathrm{PP}[\textit{from}]) = P(\mathrm{head} = \textit{retire}|\mathrm{VP}) \times P(\mathrm{VP} \rightarrow \mathrm{V}\ \ \mathrm{NP}\ \ \mathrm{PP}|\mathrm{VP}, \mathrm{head} = \textit{retire})$$

Perhaps most importantly, such models combine formal linguistic theories and quantitative data about language use, in a scientifically precise way.

However, dealing simply with surface subcategorization is well-known to be problematic. A problem that I glossed over in the discussion of *retire* above is that one also gets passives and cases where *retire* is followed by an NP, but the NP isn't an object of *retire* but a temporal NP:

(14) The SEC's chief accountant, Clarence Sampson, will retire next year and may join the FASB, which regulates the accounting profession. (*WSJ* newswire 1987/09/25)

Similarly, one might feel that use of a PP[*in*] rather than PP[*to*] (as in (13a–b)) is simply a choice from several alternate ways of expressing one thing (a goal of motion), with the choice mainly determined by the NP that occurs within the PP rather than the verb. If we accept, following much linguistic work (Grimshaw 1979, Bresnan and Moshi 1990, Grimshaw 1990), that selection is better described at a level of argument structure, with a subsequent mapping to surface subcategorization, we might instead adopt a model that uses two distributions to account for the subcategorization of a verb V. We would first propose that the verb occurs with a certain probability with a certain argument structure frame, conditioned on the verb, and presumably the context in a rich model.[16] Then we could assume that there is a *mapping* or *linking* of argument structure roles onto surface syntactic realizations. The actual surface subcategorization *Subcat* would be deterministically recoverable from knowing the argument structure *ArgStr* and the mapping *Mapping*. Then, assuming that the mapping depends on only the *ArgStr* and *Context*, and not on the particular verb, we might propose the model:

$$P(\mathit{Subcat}|V, \mathit{Context}) = \sum_{\{\mathit{Mapping}, \mathit{ArgStr}: \mathit{subcat}(\mathit{Mapping}, \mathit{ArgStr}) = \mathit{Subcat}\}} P(\mathit{ArgStr}|V, \mathit{Context}) \cdot P(\mathit{Mapping}|\mathit{Argstr}, \mathit{Context})$$

For either of these probability distributions, we might hope to usefully condition via the class of the verb (Levin 1993) rather than on particular verbs, or to do the mapping in terms of a mapping from semantic roles to grammatical functions (Bresnan and Zaenen 1990), the space of which is quite constrained.

However, such a frequency-based account is not satisfactory, because frequency of occurrence needs to be distinguished from argument status (Przepiórkowski 1999b, Grimshaw and Vikner 1993). Many arguments are optional, while some adjuncts are (almost) compulsory (e.g., a *how* adjunct for *worded* as in *He worded the proposal* ***very well***). Returning to our example of *retire* and using the same 1987 *WSJ* data, we find that the canonical temporal adjunct use of PP[*in*] (*in December*, *in 1994*) occurs with *retire* about 7 times as often as a PP[*to*] expressing a destination (and about 30 times more commonly than a PP[*in*] expressing a destination). If frequency is our only guide, a temporal PP[*in*] would have to be regarded as an

[15] Do these models estimate the entire frame jointly, or the parts of it independently? Some combination of these methods is generally best, since the joint distribution gives better information, when it can be estimated, but the sparseness of available data means that it is commonly better to work with the simpler but more robust independent estimates.

[16] A semantic selection model of roughly the required sort has been explored in NLP models by Resnik (1996).

argument. Linguists tend to regard the temporal PP as an adjunct and the destination as an argument based on other criteria, such as phrasal ordering (arguments normally appear closer to the head than adjuncts) as in (15), and iterability (adjuncts can be repeated) as in (16), or a feeling of semantic selection by the verb (any verb can have a temporal modifier).[17]

(15) a. Mr. Riley plans to retire to the Cayman Islands in December.

b. ?Mr. Riley plans to retire in December to the Cayman Islands.

(16) a. Mr. Riley plans to retire within the next five years on his birthday.

b. ?Mr. Riley plans to retire to the Cayman Islands to his ranch.

Beginning at least with Tesnière (1959), there is discussion of various criteria (morphological, syntactic, semantic, and functional) for distinguishing arguments from adjuncts. Evidence for degree of selection (or argumenthood) could thus be derived from data in more sophisticated ways, by examining such phenomena as ordering, iteration, and semantic selectivity in examples that give relevant evidence. For example, Merlo and Leybold (2001) suggest also measuring head dependence (i.e., what range of heads a particular PP appears with), which is operationalized by counting the number of different verbs which occur with a given PP (where matching is broadened to include not just exact ⟨preposition, head noun⟩ matches, but also a match over a semantic classification of the head nouns. A low number indicates argument status, while a high number indicates modifier status. We would then judge gradient argumenthood via a more complex evaluation of a variety of morphological, syntactic, semantic and functional factors, following the spirit of diagram (22) below.

However, it remains an open question whether all these phenomena are reflections of a single unified scale of argumenthood or manifestations of various underlying semantic and contextual distinctions. It has long been noted that the traditional criteria do not always converge (Vater 1978, Przepiórkowski 1999b). Such lacks of convergence and categoricity have also been seen in many other places in syntax, such as problems with the categorical unaccusative/unergative division of intransitive verbs assumed by most linguistic theories (Napoli 1981, Zaenen 1993, Sorace 2000), or the quite gradient rather than categorical ability to do extraction of or from adjuncts (Rizzi 1990, Hukari and Levine 1995).[18] There is clearly a trade-off between simplicity of the theory and factual accuracy (in either the categorical or the probabilistic case), but the presence of probabilities can make dealing with the true complexity of human language more palatable, because dominant tendencies can still be captured within the model.

3.2 On the nature of probabilistic models of syntax

In this section, I briefly address some of the questions and concerns that commonly come up about the use of probabilistic models in syntax.

3.2.1 Probabilities are not just about world knowledge

What people actually say has two parts. One part is contingent facts about the world. For example, at the time I write this, people in the Bay Area are talking a lot about electricity, housing prices, and stocks.

[17] The judgements for these examples are quite murky, and certainly wouldn't convince anyone that doesn't already believe in the argument/adjunct distinction. The distinctions are often sharper in nominalizations, which provide less freedom of ordering than verb phrases (for example, contrast *Bill Clinton's retirement from politics in 2001* with *?Bill Clinton's retirement in 2001 from politics*). See Radford (1988) for more convincing examples.

[18] The work of Ross on "squishes" is another productive source of examples (Ross 1972, Ross 1973b, Ross 1973c, Ross 1973a).

Knowledge of such facts is very useful for disambiguation in NLP, but Chomsky was right to feel that they should be excluded from syntax. But there is another part to what people say, which is the way speakers choose to express ideas using the resources of their language. For example, in English, people don't often put *that*-clauses pre-verbally, even though it is "grammatical":

(17) a. It is unlikely that the company will be able to meet this year's revenue forecasts.

b. #That the company will be able to meet this year's revenue forecasts is unlikely.

This latter statistical fact is properly to be explained as part of people's Knowledge of Language. That is, as part of syntax.[19] This is also the kind of knowledge that people outside the core community of theoretical syntacticians tend to be more interested in: sociolinguists, historical linguists, language educators, people building natural language generation and speech understanding systems, and others dealing with real language. To account for such facts, we have seen examples in this section of how one can put probability mass functions over linguistic structures. Working out an estimate of an assumed underlying distribution is referred to as *density estimation* – normally regardless of whether a discrete probability mass function or a continuous probability density function is going to be used. Probability functions can be conditionalized in various ways, or expressed as joint distributions, but for the question of how people choose to express things, our eventual goal is *density estimation* for $P(form|meaning, context)$.[20] In section 5, we will consider in more detail some methods for doing this.

One should notice – and could possibly object to the fact – that the domain of grammar has thus been expanded to include grammatical preferences, which will often reflect factors of pragmatics and discourse. Such preferences are traditionally seen as part of performance. However, I think that this move is positive. In recent decades, the scope of grammar has been expanded in various ways: people now routinely put into grammar semantic and discourse facts that would have been excluded in earlier decades. And it is not that there is nothing left in performance or contingent facts about the world: whether the speaker has short term memory limits, or is tired or drunk will influence their performance, but is not part of grammar. But principles that have a linguistic basis, and which will commonly be found to be categorical in some other language, get treated uniformly as part of grammar. This corresponds with the position of Prince and Smolensky (1993:198): "When the scalar and the gradient are recognized and brought within the purview of theory, Universal Grammar can supply the very substance from which grammars are built: a set of highly general constraints, which, through ranking, interact to produce the elaborate particularity of individual languages." Keller (2000:29) cites a study by Wolfgang Sternefeld which concludes that the bulk of instances of gradience in grammar appear to be matters of competence grammar, rather than attributable to performance, contentfully construed (i.e., things that can reasonably be attributed to the nature of online human sentence processing) or extra-linguistic factors.

3.2.2 One can put probabilities over complex hidden structure

People often have the impression that you can only put probability distributions over things that you can count, which means surface visible things like the words of a sentence, and their ordering, and not over what probabilists call "hidden structure", that is the things linguists routinely use to build theories like phrase structure trees, features with values, semantic representations, and so on. However, this is certainly no longer the case: There are now a range of techniques for putting probability distributions over hidden structure, widely used for sophisticated probabilistic modeling – we saw some simple examples above for probabilities

[19] Of course, I would not want to suggest that human grammatical knowledge stipulates exactly this. Rather it should follow from more general principles of information structure.

[20] Contrast this expression with the language understanding conditional probability from section 1.

of argument structure frames. Even if the nature or values of assumed hidden structure has never been observed, one can still build probabilistic models. In such situations, a particularly well-known and widely used technique is the Expectation Maximization or EM algorithm – (Dempster et al. 1977, McLachlan and Krishnan 1996), which is an iterative algorithm which attempts to estimate the values of hidden variables so as to make the observed data as likely as possible. This is not to say that there are not sometimes still formidable problems in successfully estimating distributions over largely hidden structure, but statisticians want to work on problems of putting probabilities over good model structures, not on denying that these structures exist. In general, there is now the ability to *add* probabilities to any existing linguistic model. The most studied case of this is for context-free grammars (Booth and Thomson 1973, Manning and Schütze 1999, Bod this volume) but distributions have also been placed over other more complex grammars such as Lexical-Functional Grammars (Johnson et al. 1999, Riezler et al. 2000, Bod and Kaplan 1998).

This is not to say that one should not rethink linguistic theories when adding probabilities, as I have tried to indicate in the beginning part of this section. Nevertheless, especially in the psycholinguistic literature, the dominant terms of debate have been that you either have to stick to good old-fashioned rule/constraint systems from linguistic theory or you have to move to connectionist modeling (Seidenberg and MacDonald 1999, Pinker 2000). Without in any way wishing to dismiss connectionist work, I think it is important to emphasize that this is a false dichotomy: 'soft' probabilistic modeling can be done over rule systems and/or symbolic structures. The numerate connectionist community has increasingly morphed into a community of sophisticated users of a variety of probabilistic models.

3.2.3 Can probabilities capture the notion of grammaticality?

There have been several recent pieces of work that have argued with varying degrees of strength for the impossibility of collapsing notions of probability and grammaticality (Abney 1996, Culy 1998, Keller 2000). This is a difficult question, but I feel that at the moment this discussion has been clouded rather than clarified by people not asking the right probabilistic questions. It is certainly true that in general a sentence can have arbitrarily low probability in a model, and yet still be perfectly good – in general this will be true of all very long sentences, as Abney (1996) notes. It is also the case that real language use (or a statistical NLP model aimed at disambiguating between meanings) builds in facts about the world. In these cases, the probability of different sentences reflects the nature of the world, and this clearly needs to be filtered out to determine a notion of grammatical acceptability independent of context. A profitable way to connect grammaticality and probability is perhaps to begin with the joint distribution $P(\mathit{form}, \mathit{meaning}, \mathit{context})$. We could then consider ways of marginalizing out the context and the meaning to be expressed.[21] We would not want to do this via the actual empirical distributions of contexts and ideas which people express, which reflect contingent facts about the world (cf. section 3.2.1), but by imposing some flat prior distribution, which is uninformative about the contingent facts of the world. For instance, in this distribution, people would be equally likely to travel to North Dakota for holidays as to Paris. Using such an uninformative prior is a rough probabilistic equivalent of considering *possible worlds* in semantics. Under this analysis, forms might be considered gradiently grammatical to the extent that they had probability mass in the resulting marginal distribution.

[21] The concept of "marginalizing out" in probability refers to removing a variable that we are not interested in, by summing (in the discrete case) a probability expression over all values of that variable. For instance, if there were just three contexts, c_1, c_2, and c_3, we could marginalize out the context by computing:

$$P(f, m) = P(f, m, c_1) + P(f, m, c_2) + P(f, m, c_3) = \sum_{c_i} P(c_i) P(f, m|c_i)$$

Computationally efficient methods for manipulating individual variables and marginalizing out ones that are not of interest are at the heart of work on using Bayesian Networks for reasoning (Jensen and Jensen 2001).

Alternatively, one might think that the above suggestion of marginalizing out context is wrong because it effectively averages over possible contexts (see footnote 21), and a closer model to human judgements would be that humans judge the grammatical acceptability of sentences by assuming a most favorable real world context. For this model, we might calculate probabilities of forms based on the probability of their most likely meaning in their most favorable context, by instead looking at:[22]

$$P(f) = \frac{1}{Z_f} \max_m \max_c P(f, m, c)$$

However, while humans clearly have some ability to consider sentences in an imagined favorable context when judging (syntactic) 'grammaticality', sociolinguistic and psycholinguistic research has shown that judgements of grammaticality are strongly codetermined by context and that people don't automatically find the best context (Labov 1972:193–198, Carroll et al. 1981). Nevertheless, Keller (2000) argues for continuing to judge grammaticality via acceptability judgements (essentially still intuitive judgements, though from a more controlled experimental setting), and modeling soft constraints by looking at relative acceptability. The issues here deserve further exploration, and there may be valuable alternative approaches, but, as exemplified at the beginning of this section, I tend to feel that there are good reasons for getting at synchronic human grammars of gradient phenomena from large data sets of actual language use rather than from human judgements. In part this is due to the unclarity of what syntactic acceptability judgements are actually measuring, as just discussed.

3.2.4 The formal learnability of stochastic grammars

One might think that adding probability functions to what are already very complex symbol systems could only make formal learnability problems worse, but this is not true. An important thing to know is that adding probability distributions to a space of grammars can actually *improve* learnability. There is not space here for a detailed discussion of language acquisition data, but I wish to address briefly the formal learnability of syntactic systems.

If linguists know anything about formal learnability, it is that Gold (1967) proved that formal languages (even finite state ones) are unlearnable from positive evidence alone. This result, together with an accepted wisdom that children do not have much access to negative evidence and do not pay much attention to it when it is given, has been used as a formal foundation for the Chomskyan assumption that an articulated and constrained universal grammar must be innate.

There are many directions from which this foundation can be undermined. Some recent work has emphasized the amount of implicit negative evidence (through lack of comprehension, etc.) that children do receive (Sokolov and Snow 1994, Pullum 1996). One can question the strictness of the criterion of identifiability in the limit used by Gold (namely, that for any language in the hypothesis space, and for any order of sentence presentation, there must be a finite length of time after which the inference device always returns the correct language). Under a weaker criterion of approachability in the limit (each incorrect grammar is rejected after a finite period of time), then context-free and context-sensitive grammars are learnable from positive evidence alone (Feldman et al. 1969).

But most importantly in this context, assuming probabilistic grammars actually makes languages easier to learn. The essential idea is that such a grammar puts probability distributions over sentences. If the corpus was produced from such a grammar, it is highly probable that any phenomenon given a high probability by the grammar will show up in the corpus. In contrast, Gold's result depends on the fact that the learner may see an unrepresentative sample of the language for any period of time. Using probabilistic grammars,

[22] The Z_f term is needed for renormalization so that a probability distribution results. See the example of renormalization in section 5.6.

absence of a sentence from the corpus thus gives implicit negative evidence: If a grammar would frequently generate things that never appear in the corpus, that grammar becomes increasingly unlikely as the size of the seen corpus data grows. In particular, Horning (1969) showed that providing one assumes a denumerable class of possible grammars with a prior distribution over their likelihood, then stochastic grammars *are* learnable from positive evidence alone.

Such formal learnability results give guidance, but they contain conditions (stationarity, independence, etc.) which mean that they are never going to exactly model "real world conditions". But more generally it is important to realize that whereas around 1960 when the linguistic orthodoxy of 'poverty of the stimulus' developed, almost nothing was known about learning, a lot has changed in the last 40 years. The fields of statistics and machine learning have made enormous advances, stemming from both new theoretical techniques and the opportunities for computer modeling. A particular result of relevance is that it is often easier to learn over continuous spaces than discontinuous categorical spaces, essentially because the presence of gradients can direct learning.[23]

3.2.5 Explanatory value and qualms about numbers

There are two final worries. One could doubt whether it will ever be possible to determine probabilities correctly for abstract but important linguistic features. What is the probability of a parasitic gap (Engdahl 1983) in a sentence given that it contains an adjunct extraction? Obviously, we can never hope to estimate all these probabilities exactly from a finite corpus, and at any rate we would be defeated by the non-stationarity of language if we tried. But this is unimportant. 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. Theoretically, our assumption is that such probabilities do exist and have exact values: there is some probability that a sentence that an English speaker utters will contain a parasitic gap. There is also a probability that the next sentence that I utter will contain a parasitic gap. This probability may differ from the average for the population, and may change when one conditions on context, the meaning to be expressed and so on, but it exists. This probability simply captures an aspect of the behavior of the wetware in our brain (and is, at this level, uncommitted even as to whether some, much or none of syntax is innate). People also sometimes just object to the use of numbers whatsoever in grammatical theory, independent of whether they can be determined. This strikes me as an invalid objection. Kroch (2001:722) argues that "There is no doubt, however, that human beings like other animals, track the frequencies of events in their environment, including the frequency of linguistic events.[24]

Secondly, one could think that there are so many numbers flying around that we can just fit (or "learn") anything, and that there is thus no explanatory value with regards to the goal of describing possible human languages. This is a genuine worry: if the model is too powerful or general, it can essentially just memorize the data, and by considering varying parameter settings, one might find that the model provides no constraint on what languages are possible.[25] This is one possible objection to the Data-Oriented Parsing models of Bod (this volume): they do essentially just memorize all the data (to provide constraints on possible linguistic systems one needs to add substantive constraints on the representations over which the model learns and on

[23] Though I should note that there has also also been considerable progress in categorical learning. There are now a number of good books on statistical and machine learning (Bishop 1995, Ripley 1996, Mitchell 1997, Hastie et al. 2001), and conferences such as CoNLL (http://ilk.kub.nl/~signll/) which focus on natural language learning.

[24] See Bod (1998) for further evidence of human sensitivity to linguistic frequencies.

[25] Concern about overly powerful learning methods is, admittedly, an issue that has been felt more keenly in linguistics than in most other areas, which have concentrated on finding sufficiently powerful learning methods. Smolensky (1999) attributes to Dave Rumelhart the notion that "linguists think backwards". Nevertheless, the goal of explaining the extensive but limited typological variation of human language is a real and valid one.

the fragment extraction rules, and even then, restrictions on the space of possible soft constraints are not present). But this just means that we should do a good job at the traditional syntactic tasks of looking for constraints that apply across syntactic constructions, and looking for model constraints that delimit possible human linguistic systems. Some starting ideas of how one can do this appear in section 5. While there are other approaches to "softness" than probabilities (such as using prototypes or fuzzy logic), the sound theoretical foundation of probability theory, and the powerful methods for evidence combination that it provides makes it the method of choice for dealing with variable phenomena.

4 Continuous categories

Earlier, we followed Chambers (1995) in emphasizing an alternative approach to linguistics that allows units which are *continuous* and *quantitative*. Any probabilistic method is quantitative, and the actual probabilities are continuous quantities, but most work using probabilities in syntax, and in particular the large amount of related work in Statistical NLP, has put probabilities over discrete structures and values, of the kind familiar from traditional linguistics. However, while language mostly acts like a discrete symbol system (making this assumption workable in much of linguistics), there is considerable evidence that the hidden structure of syntax defies discrete classification. People continually stretch the "rules" of grammar to meet new communicative needs, to better align grammar and meaning, and so on. Such leakage in grammars leads to gradual change. As a recent example, the term *e-mail* started as a mass noun like *mail* (*I get too much junk e-mail*). However, it is moving to be a count noun (filling the role of the non-existent **e-letter*): *I just got an interesting email about that.* This change happened in the last decade: I still remember when this last sentence sounded completely wrong (and ignorant (!)). It then became commonplace, but still didn't quite sound right to me. Then I started noticing myself using it.

As a more detailed example of blurring of syntactic categories during linguistic change, I'll consider what are sometimes known as "marginal prepositions" (Quirk et al. 1985:666).[26] There is a group of words in English which have the form of present participles of verbs, and which are used as verbs, but which also appear to function as prepositions. Examples include: *concerning, supposing, excepting, according, considering, regarding,* and *following*. Historically and morphologically, these words began as verbs, but are moving in at least some of their uses from being participles to prepositions. Some still clearly maintain a verbal existence, like *following, concerning, considering*; for some it is marginal, like *according, excepting*; for others their verbal character is completely lost, such as *during* (previously a verb, cf. *endure*),[27] *pending,* and *notwithstanding*.

In the remainder of this section I will illustrate with *following*, which is one of the more recent participles to make the move, and has been the subject of an earlier corpus study (Olofsson 1990), to which I am endebted for various of the references cited below. While some prepositional uses of *following* are recorded quite early:

(18) Following his ordination, the Reverend Mr Henry Edward intends to go to Rome (1851)

the prepositional uses seem to have grown considerably in frequency in the second half of the twentieth century. One manifestation of this, as so often in cases of language change, is that we begin to see prescriptive commentators on language condemning the innovation. The first edition of Fowler's well-known (British) dictionary of usage makes no mention of *following* but elsewhere shows considerably liberality about this process of change: "there is a continual change going on by which certain participles or adjectives acquire the character of prepositions or adverbs, no longer needing the prop of a noun to cling to . . . [we see] a

[26] A couple of other examples appear in chapter 1 of Manning and Schütze (1999).

[27] For example, "The wood being preserv'd dry will dure a very long time" (Evelyn 1664).

development caught in the act" (Fowler 1926). However, in Gowers (1948:56) we find: "*Following* is not a preposition. It is the participle of the verb *follow* and must have a noun to agree with". With its continued spread, the 1954 revised edition of Fowler, edited by Gowers, still generally condemns the temporal usage, but softens things by saying that it can be justified in certain circumstances.

One test for a participial use is that participial clauses require control of their subject from a noun phrase in the main clause (19a), while prepositions do not. So (19b) is possible only if we anthropomorphize the truck to control the subject of the verb *seeing*, whereas the preposition *after* in (20) does not have any controlled subject:

(19) a. Seeing the cow, he swerved violently.

b. #Seeing the car, the truck swerved violently.

(20) After the discovery, the price of gold began to drop.

One can observe the following development. From cases where we clearly have a participial verb with the basic motion sense of *follow* (21a), it became common to use *follow* to indicate place (21b) or time (21c). In these examples, it is still plausible to regard a noun phrase in the main clause (including the unexpressed subject of the imperative in (21b)) as coreferent with the subject of *following*, but such an assumption is not necessary: we could replace *following* with the preposition *after* with virtually no change of interpretation. This ambiguity of analysis allowed the temporal use to become generalized, with *following* beginning to be used as a preposition with no possible controller (21d–e):

(21) a. They moved slowly, toward the main gate, following the wall.

b. Repeat the instructions following the asterisk.

c. This continued most of the week following that ill-starred trip to church

d. He bled profusely following circumcision.

e. Following a telephone call, a little earlier, Winter had said . . . .

There is a tradition in linguistics of imposing categoricity on data. One sees this also in NLP. For example, the tagging guidelines for the Penn Treebank (Santorini 1990) declares about such cases:

> Putative prepositions ending in *-ed* or *-ing* should be tagged as past participles (VBN) or gerunds (VBG), respectively, not as prepositions (IN).
>
> According/VBG to reliable sources
> Concerning/VBG your request of last week

Maintaining this as an arbitrary rule in the face of varying linguistic usage is essentially meaningless. One can avoid accepting gradual change by stipulating categoricity. But the results of such moves are not terribly insightful: It seems that it would be useful to explore modeling words as moving in a continuous space of syntactic category, with dense groupings corresponding to traditional parts of speech (Tabor 2000).[28]

[28] Another approach is to make progressively finer discrete subdivisions of word types, as is possible within a hierarchical lexicon approach (Malouf 2000). If the subdivisions are fine enough, it becomes difficult to argue against such an approach by linguistic means (since there are normally only a finite number of linguistic tests to employ). Indeed, such a model approaches a continuous model in the limit. Such an approach is not necessarily more insightful or constrained, though; indeed, I believe that position in a continuous space may be a more natural way to approach modeling variation and change.

5 Explaining more: probabilistic models of syntactic usage

Much of the research in probabilistic natural language processing has worked with probabilities over rules, such as the PCFG models introduced in Bod (this volume). Such an approach probably seems old-fashioned to most linguists who have moved to thinking in terms of constraints on linguistic representations. In this section we adopt such a viewpoint and examine the space of probabilistic and categorical linguistic models, not from the point of view of the substantive claims of particular theories, but from the perspective of how constraints interact.

5.1 A linguistic example

I integrate with the theory discussion of a particular syntactic example drawn from the interaction of passive, person, and topicality. This example is drawn from joint work on stochastic syntax with Joan Bresnan and Shipra Dingare; see (Bresnan et al. 2001, Dingare 2001) for more details and references. The example is very much simplified for expository purposes, but nevertheless, it has enough structure to be able to compare and contrast different approaches, and to build example probabilistic models. The central aim of the example is to support the following proposition: The same categorical phenomena which are attributed to hard grammatical constraints in some languages continue to show up as soft constraints in other languages. This argues for a grammatical model that can account for constraints varying in strength from soft constraints to hard constraints within a uniform formal architecture, and for the explanatory inadequacy of a theory that cannot. Compare these remarks from Givón (1979) contrasting a categorical restriction on indefinite subjects in Krio versus the dispreference for them in English:

> But are we dealing with two different kinds of facts in English and Krio? Hardly. What we are dealing with is apparently the very same *communicative tendency*—to reserve the subject position in the sentence for the *topic*, the old-information argument, the "continuity marker." In some languages (Krio, etc.), this communicative tendency is expressed at the categorial level of 100%. In other languages (English, etc.) the very same communicative tendency is expressed "only" at the noncategorial level of 90%. And a transformational-generative linguist will then be forced to count this fact as competence in Krio and performance in English.

Ideas of stronger and weaker constraints are common in the typological and functionalist syntactic literatures, but until now there has been a dearth of formal syntactic models that can describe such a situation. Rather, the soft constraints are treated as some sort of performance effect, which bears *no* theoretical relation to the categorical rules of other languages, which would be seen as deriving, for instance, from parameter settings in Universal Grammar.

A traditional linguist could argue that the competence system is categorical, with only hard constraints, but that there is a lot of noise in the performance system, which means that ungrammatical sentences are sometimes produced. The first argument against this position was presented in section 3.1, where the problem of the non-identifiability of the competence grammar (where to draw the line between grammatical and ungrammatical sentences) was discussed. In this section, we concentrate on the stronger argument that this model is unexplanatory because one would have to put into the performance system soft constraints which should be treated uniformly with hard constraints of the competence grammar of other languages. At any rate, one needs a model of any noise attributed to the performance system; simply leaving it unspecified, as in most linguistic work, means that the model makes no empirical predictions about actual language use.

Consider an event where a policeman scolded me, where I'm part of the existing discourse (perhaps just by virtue of being a speech act participant), but mention of the policeman is new. This *input* semantic representation will be described by an extended *argument structure* as *see*⟨*policeman* **[new]**, *me* **[old]**⟩. In

English, this idea would most commonly be expressed as an *output* by either a simple active or passive sentence, as in (22). An equivalent pair of choices is available in many languages, but in others, one of these choices is excluded.

(22)

Constraints = Features = Properties	Linking	Discourse	Person
Input: *scold*⟨*policeman* **[new]**, *me* **[old]**⟩	f_1 ✓Subj/Ag *NonSubj/Ag	f_2 ✓Subj/Older *Subj/Newer	f_3 ✓$1/2 > 3$ $*3 > 1/2$
a. *A policeman scolded me*	1	0	0
b. *I was scolded by a policeman*	0	1	1

What determines a speaker's choice here? There are many factors, but in the simplified example in (22), we will consider just three. The constraints can be stated either positively or negatively, and I provide mnemonic names for each polarity. Use of negative constraints is de rigeur in Optimality Theory (Prince and Smolensky 1993), but the choice is arbitrary, and for the generalized linear models which I discuss later, it may be more helpful to think of them as positive soft constraints ("the subject should be old information", etc.).

1. *A linking constraint:* ✓Subj/Ag = *NonSubj/Ag. It is preferable for the subject of the sentence to express the entity who is performing the action of the sentence, that is, its *agent*.

2. *A discourse constraint:* ✓Subj/Older = *Subj/Newer. For purposes of topic continuity, it is better for the previously mentioned (i.e., older) entity to be the subject.

3. *A person constraint:* ✓$1/2 > 3 = *3 > 1/2$. Because the 1st/2nd person speech act participants are the central foci of discourse empathy, it is better for one of them to be the subject, the thing which the sentence is perceived to be about, than some other third person entity.

These three constraints (also referred to as features or properties) are shown in (22), both via their mnemonic names, and labeled as f_1, f_2, and f_3. The first would favor use of the active, while the other two would favor use of the passive. In the table, a 1 means a constraint is satisfied, and a 0 means that it is violated (i.e., a 0 corresponds to where a * would appear in an Optimality Theory tableau).

The constraints introduced here are all well supported crosslinguistically. Within Optimality Theory, their basis can be captured by making use of the concept of *harmonic alignment* (Prince and Smolensky 1993). When there are two scales of prominence, one structural (e.g., syntactic positions) and the other substantive (e.g., semantic features), harmonic alignment specifies a way of building an (asymmetric) partially ordered set of constraints in the form of linear scales, referred to as *constraint subhierarchies*. These give putatively universal limits on constraint ordering/strength. For instance, following Aissen (1999) we could perform harmonic alignment between a surface grammatical relations hierarchy and a thematic hierarchy to derive the importance of the Linking constraint (Aissen 1999):

(23)

Hierarchies:	Harmonic alignment:	Constraint subhierarchies:
Agent $\succ$ Patient	Subj/Ag $>$ Subj/Pt	*Subj/Pt $\gg$ *Subj/Ag
Subj $\succ$ NonSubj	NonSubj/Pt $>$ NonSubj/Ag	*NonSubj/Ag $\gg$ *NonSubj/Pt

The second subhierarchy, restated positively, gives the result that universally a feature rewarding agents that are subjects must be ranked higher than a feature rewarding patients that are subjects. For simplicity, in the remainder, we only use the higher ranked constraint, ✓Subj/Ag = *NonSubj/Ag. Similarly, it is well supported in the typological literature that there is a person hierarchy, where 1st and 2nd person outrank 3rd, that is that *local* person outranks *nonlocal* (Silverstein 1976, Kuno and Kaburaki 1977, Givón 1994). This constraint could be modeled by harmonic alignment of the person hierarchy with the grammatical relations hierarchy, but I have expressed it as a simple relational constraint – just to stress that the models allow the inclusion of any well-defined property, and are not restricted to one particular proposal for defining constraint families. At any rate, in this manner, although our constraints are possibly soft, not categorical, we can nevertheless build in and work with substantive linguistic hypotheses.

5.2 Categorical (constraint-based) syntactic theories

The standard but limiting case of a syntactic theory is the categorical case, where outputs are either grammatical or ungrammatical. The grammar assumes some kind of representation, commonly involving tree structures and/or attribute-value representations of grammatical features. The grammar explains how the representation is used to describe an infinite set of sentences, and defines a number of hard (categorical) constraints on those representations. Generative approaches assume a common set of representations and principles, Universal Grammar (UG), underlying all languages, and then either or both language-specific constraints, which are conjoined with the ones given by UG, or else parameterized constraints, which are part of UG, but which can vary on a language particular basis. A grammatical sentence must satisfy all the constraints. Sentences that do not are regarded as syntactically ill-formed or ungrammatical. For instance, there is an *agreement constraint*, which ensures that, for the passive sentence, (24a) is grammatical, while (24b) is not:

(24) a. I was scolded by a policeman.

b. *I were scolded by a policeman.

A grammar of a language is the conjunction of a large set of such constraints over the representations. One can determine the grammatical sentences of the language by solving the resulting large constraint satisfaction problem (Carpenter 1992, Stabler 2001). In conventional, categorical NLP systems, ungrammatical sentences do not parse, and so cannot be processed.[29]

For all the main forms of formal/generative syntax, the grammar of English will do no more than say that both the active and the passive outputs in (22) are possible grammatical sentences of English. Since none of the constraints $f_1, \ldots, f_3$ are categorical in English, none would be part of the grammar of English. In other words, the grammar says nothing at all about why a speaker would choose one output or the other.

However, in many languages of the world, one or other of these constraints is categorical, and one would not be able to use sentences corresponding to both of our English outputs in the situation specified in (22).[30] In general, passives are marked (Greenberg 1966, Trask 1979): Many languages lack a passive construction, passives are normally more restricted language internally, and normally morphologically marked. While this could be partly historical happenstance, I see this as reflecting the constraint f_3. If the constraint f_3 is categorical in a language, then passive forms will never occur:[31]

(25)

	scold⟨*policeman* **[new]**, *me* **[old]**⟩	*NonSubj/Ag
☞	active: S_{ag}, O_{pt} *A policeman scolded me*	
	passive: S_{pt}, Obl_{ag} *I was scolded by a policeman*	*!

[29] Although, in practical NLP systems, people have commonly added some kind of heuristic repair strategy.

[30] Below, I discuss the Linking and Person constraints. The presented Discourse constraint is somewhat oversimplified, but nevertheless represents the essentials of another widespread phenomenon. Not only is it a soft constraint in English, but in many Philippine languages, such as Tagalog, there is a categorical specificity constraint on subject position, which resembles (but is more complex than) the Discourse constraint given here (Kroeger 1993).

[31] Here, adopting Optimality Theory notations, the * shows a constraint violation, the ! means it is fatal, and the pointing hand indicates a valid form (which is unique for Optimality Theory, but there might be many for a categorical grammar).

The effects of the person hierarchy on grammar are categorical in some languages, most famously in languages with inverse systems, but also in languages with person restrictions on passivization. In Lummi (Coast Salishan, U.S.A. and Canada), for example, if one argument is 1st/2nd person and the other is 3rd person, then the 1st/2nd person argument must be the subject. The appropriate choice of passive or active is obligatory to achieve this (26–27). If both arguments are 3rd person (28), then either active or passive is possible (Jelinek and Demers 1983, Jelinek and Demers 1994):[32]

(26) a. *'The man knows me/you'

b. x̣či-t-ŋ=sən/=sxʷ ə cə swəyʔqəʔ
know-TR-PASS=1/2.NOM.SUBJ by the man
'I am/you are known by the man'

(27) a. x̣či-t=sən/=sxʷ cə swəyʔqəʔ
know-TR-3.ERG.SUBJ the man
'I/you know the man'

b. *'The man is known by me/you'

(28) a. x̣či-t-s cə swəyʔqəʔ cə swiʔqoʔəł
know-TR-3.ERG.SUBJ the man the boy
'The man knows the boy'

b. x̣či-t-ŋ cə swiʔqoʔəł ə cə swəyʔqəʔ
know-TR-PASS the boy by the man
'The boy is known by the man'

Such interactions of person and voice (or inversion) occur in a considerable number of Native American Languages, and also more widely (Bresnan et al. 2001). To account for this, in Lummi we can say that the person constraint f_2 is categorical (i.e., part of the grammar of Lummi), but the others are not. Some of the Lummi cases then look like this:

(29)

	scold⟨*policeman* **[new]**, *me* **[old]**⟩			*3 > 1/2
	active:	S_3, O_1	*A policeman scolded me*	*!
☞	passive:	S_1, Obl_3	*I was scolded by a policeman*	

(30)

	scold⟨*policeman* **[new]**, *Fred* **[old]**⟩			*3 > 1/2
☞	active:	S_3, O_3	*A policeman scolded Fred*	
☞	passive:	S_3, Obl_3	*Fred was scolded by a policeman*	

In a categorical model, if constraints come into conflict, then the form is ungrammatical. Therefore if a language had a categorical version of both the linking constraint and of the person constraint, then there would be no way to express the idea *scold*⟨*policeman*, *me*⟩. Expressive gaps of this sort may occasionally occur in language (perhaps, an example is expressing ideas that violate relative clause *wh*-extraction constraints – (Pesetsky 1998)), but are extremely rare. In general, languages seem to conspire to avoid such happenings. In a categorical grammar, the linguist has to build such conspiracies into the grammar by restricting the basic constraints, either by adding by hand complex negated conditions on constraints, or by making use of ideas like the elsewhere principle (Kiparsky 1973).

[32] Note that the Lummi pattern holds for bound pronouns; full pronouns designating speaker and hearer are focussed and formally 3rd person expressions (Jelinek and Demers 1994:714). If both arguments are 1st/2nd person, an active form is required.

The typological factors of person and linking are present in the analysis of passive and voice in Kuno and Kaburaki (1977). Again, we see in categorical work a tendency to deem ungrammatical structures that are simply very marked, though perhaps licensed in special discourse circumstances. For example, Kuno and Kaburaki (1977) star as ungrammatical sentences such as (31):

(31) *Mary was hit by me

On their account, it violates the Ban on Conflicting Empathy Foci: on the grammatical relations hierarchy, the choice of Mary as subject implies that the speaker empathizes with Mary, while the speech-act participant hierarchy dictates that the speaker must empathize most with himself. A conflict is possible in an unmarked active sentence, but not in a marked sentence type. However, Kato (1979) takes them to task for such claims, providing corpus evidence of passives with first person agents, such as:

(32) Gore [Vidal] never lacked love, nor was he abandoned by me. (*Time*)

It is somewhat unclear what Kuno and Kaburaki (1977) were claiming in the first place: while in some cases they explain a * ungrammaticality mark by invoking the Ban on Conflicting Empathy Foci, in others they suggest that special contexts can make violations of empathy constraints possible. Their conclusion points to the kind of approach to optimization over soft constraints that I am advocating here: "Violations of empathy constraints sometimes yield totally unacceptable sentences; at other times, especially when other factors make up for the violations, they yield only awkward or even acceptable sentences." Such situations can only be written about informally when working within a categorical formal framework. When working with a probabilistic syntactic framework, they can be *formally modeled*.

5.3 Optimality Theory

Standard Optimality Theory (OT) is not a probabilistic framework,[33] but it is a useful in-between point as we proceed from categorical to probabilistic grammars. OT differs from standard categorical grammars by assuming that optimization over discrete symbol structures via ranked, violable constraints is fundamental to the cognitive architecture of language. Human language is described by a set of universal constraints, which are hypothesized to be present in all grammars, but they are more or less active depending on their ranking relative to other constraints. The (unique) grammatical output for an input is the one that optimally satisfies the ranked constraints of a language, where satisfying a higher ranked constraint is judged superior to satisfying any number of lower ranked constraints. Consider again (22). If all three constraints were present and categorical, there would be no output: the theory would simply be inconsistent. However, under the OT conception of ranked violable constraints, we can maintain all the constraints, and simply rank the Person constraint highest. As the person constraint is categorical for Lummi (*undominated* in OT terminology), it will determine the passive output for (22), as shown in (33a). When there are two 3rd person arguments, Lummi falls back on other constraints. Since passivization is still possible here, we conclude that the linking constraint is ranked low, and other information structure constraints, such as our Discourse constraint determine the optimal output. For example, if the agent is the topic, that is, the older information, then the active will be chosen over the passive (33a), and vice versa (33a).

(33) a.

	input: v⟨ag/3/old, pt/3/new⟩	*3 > 1/2	*Subj/Newer	*NonSubj/Ag
☞	active: S_{ag},O_{pt}			
	passive: S_{pt},Obl_{ag}		*!	*

b.

	input: v⟨ag/3/new, pt/1/old⟩	*3 > 1/2	*Subj/Newer	*NonSubj/Ag
	active: S_{ag},O_{pt}	*!	*	
☞	passive: S_{pt},Obl_{ag}			*

[33]Except in the trivial sense where all probabilities are 0 or 1.

The introduction of a constraint ranking can potentially give a better description of Lummi syntax. It shows how secondary constraints come into play when the most important constraints are satisfied (a phenomenon referred to as "emergence of the unmarked"). The output form does not have to obey all these constraints: it is most important that Person is observed, but if this constraint does not differentiate forms, then lower ranked constraints show through and determine the grammatical output. This automatically gives us the kind of *elsewhere hierarchy* that has been widely observed in linguistic systems (Kiparsky 1973). This ranking of constraints, and the ability for less highly ranked constraints to show through is the hallmark of OT.

A problem with OT is that it flies in the face of the words of Sapir with which this chapter began: OT predicts that there should always be a unique output for every input. The highest ranked differentiating constraint always determines things, giving a unique output. This is tenable in some areas of linguistics, but it goes against widespread *variation* in the use of language, not just across members of a community, but usually also by each individual (Guy 1981, Labov 1994, Kroch 2001). Because of the generally problematic nature of this theoretical assumption, there has arisen a small industry trying to work variable outputs into Optimality Theory (Nagy and Reynolds 1997, Anttila 2001, Boersma 1997, Pesetsky 1998, Müller 1999). For the particular case here, there are existing OT accounts of diathesis alternations based on typological evidence (Legendre et al. 1993, Aissen 1999, Lødrup 1999, Ackema and Neeleman 1998), but once one incorporates the role of discourse and information structure into diathesis prediction, the consequent variation suggests the use of a stochastic account. Below we will discuss Boersma's Stochastic OT, the best motivated and most thoroughly probabilistic extension to OT.

It is interesting that, historically (for at least Paul Smolensky, one of the primary architects of OT), OT is a retreat from a quantitative framework which does harmony maximization over numerical soft constraints. The earlier approach of Harmony Grammar (Smolensky 1986, Legendre et al. 1990), although couched in a connectionist network, is mathematically extremely close to the loglinear models that we discuss below (Smolensky and Legendre 2000). Prince and Smolensky (1993:198) motivate OT by noting that in practice "Order, not quantity (or counting), is the key in Harmony-based theories. In Optimality Theory, constraints are ranked, not weighted; harmonic evaluation involves the abstract algebra of order relations rather than numerical adjudication between quantities." This strict domination ordering of constraints is argued by Prince and Smolensky to be the basic, fundamental, ubiquitous, universal, and unmarked means of constraint interaction within linguistics.

It is true that such a ranking is often sufficient (just as it is true for some purposes that categorical constraints are sufficient), but the need to handle variability is one key reason for believing that sometimes more is needed. The other reason is to be able to handle the phenomenon of "ganging up", where multiple lesser constraint violations are deemed to make an output worse than another with just one more serious constraint violation. There is partial recognition of and a partial treatment for this phenomenon within OT via the theory of Local Conjunction (Smolensky 1993, Smolensky 1997), whereby a conjunction of two constraints can be given a position in the constraint ordering separate from (i.e., higher than) the two individual constraints. But as soon as one wishes to allow a general conception of constraints having a combined effect (that is, them "ganging up"), then one needs to bring back the numbers. An ordering alone is insufficient to assess when multiple lesser constraints will or will not overrule higher ranked constraints, or (in terms of the local conjunction perspective) where a local conjunction should be placed within the overall constraint ranking. While Smolensky motivates local conjunction as a limited but necessary deviation from the basic method of linguistic constraint interaction, it goes somewhat against the spirit of OT, and suggests the model is not quite right: at the end of the day, a combination of constraint violations is deemed worse than a violation of just the highest ranked individual constraint.

5.4 The linguistic example, continued

Considering again (22), none of the three constraints shown are categorical in the grammar of English, but all of them play a role. All else being equal, old information is more commonly the subject, local persons are more commonly the subject, and agents are more commonly the subject.

Quantitative data can demonstrate that a language exhibits soft generalizations corresponding to what are categorical generalizations in other languages. A probabilistic model can then model the strength of these preferences, their interaction with each other, and their interaction with other principles of grammar. By giving variable outputs for the same input, it can predict the statistical patterning of the data. Beyond this, the model allows us to connect such soft constraints with the categorical restrictions that exist in other languages, naturally capturing that they are reflections of the same underlying principles. This serves to effectively link typological and quantitative evidence.

Bresnan et al. (2001) collected counts over transitive verbs from parsed portions of the Switchboard corpus of conversational American English (Godfrey et al. 1992), analyzing for person and active vs. passive. Switchboard is a database of spontaneous telephone conversations between anonymous callers spread across the United States. We chose Switchboard because it is a large, parsed, spoken corpus (about 1 million words are parsed). The parsing made our data collection significantly easier. Being spoken data not only made it more natural, but meant there was a high use of 1st and 2nd person, whereas in many of the (mainly written) corpora available to us, 1st and 2nd person are extremely rare. On the other hand, full passives (ones with *by* phrases) turn out to be very rare, and so despite counting around 10,000 clauses, the results table below still has a couple of zeroes in it.[34]

In English, there is not a categorical constraint of person on passivization, but we can still see the same phenomenon at work as a *soft constraint*. We found that the same disharmonic person/argument associations which are avoided categorically in languages like Lummi also depress or elevate the relative frequency of passives in the Switchboard corpus. Compared to the rate of passivization for inputs of third persons acting on third persons (1.2%), the rate of passivization for first or second person acting on third is substantially depressed (0%) while that for third acting on first or second (2.9%) is substantially elevated.[35]

(34)

Event roles:	# Act:	# Pass:	% Act:	% Pass:
v⟨ag/1,2; pt/1,2⟩	179	0	100.00	0.00
v⟨ag/1,2; pt/3⟩	6246	0	100.00	0.00
v⟨ag/3; pt/3⟩	3110	39	98.76	1.24
v⟨ag/3; pt/1,2⟩	472	14	97.11	2.89
Total/Mean	10007	53	99.47	0.53

The leftmost column gives the four types of inputs (local person acting on local, local acting on nonlocal, etc.). For each input, we calculate the rate of passivization from the number of times that input was realized as passive. The percentage of full passives in spoken English is very small: most passives involve suppression of the agent (a further 114 examples with a local person patient, and 348 examples with a 3rd

[34] Again, though, these are not "structural" zeroes representing impossible occurrences: an example of a sentence that would be counted in the top right cell was given in (32).

[35] Previous studies also show evidence of person/voice interactions in English (Svartvik 1966, Estival and Myhill 1988). However, they only provide various marginals (e.g., counts of how many actives and passives there are in texts, etc.), which gives insufficient information to reconstruct the full joint distribution of all the variables of interest. Estival and Myhill (1988) *do* provide the kind of information needed for animacy and definiteness, but they provide person frequencies only for the patient role. We want to be able to predict the overall rate of the different systemic choices (Halliday 1994) which can be made for a certain input. That is, we want to know, for instance:

$$P(form = passive|ag = 1, pt = 3)$$

We can determine the conditional frequencies needed from the full joint distribution (Bod, this volume).

person patient).[36] Person is only a small part of the picture in determining the choice of active/passive in English (information structure, genre, etc. are more important – but we left to further research consideration of information structure because it is much more difficult to classify). Nevertheless, there is a highly significant effect of person on active/passive choice.[37] The exact same hard constraint of Lummi appears as a soft constraint in English.

5.5 Stochastic Optimality Theory

Stochastic OT (Boersma 1997, Boersma 1998, Boersma and Hayes 2001) basically follows Optimality Theory, but differs in two essential ways. Firstly, constraints are not simply ordered, but they have a value on the continuous scale of real numbers. Constraints are specific distances apart, and these distances are relevant to what the theory predicts. Secondly, there is stochastic evaluation, which leads to variation, and hence to a probability distribution over outputs from the grammar for a certain input. At each evaluation the value of each constraint is perturbed by temporarily adding to its ranking value a random value drawn from a normal distribution. For example, a constraint with a rank of 99.6 could be evaluated at 97.1 or 105. It is the constraint ranking that results from this perturbation that is used in evaluation. Hence the grammar constrains but underdetermines the output. One could think that this model of random perturbation is rather strange: Does a speaker roll dice before deciding how to express him or herself? There may be some inherent randomness in human behavior, but principally the randomness simply represents the incompleteness of our model and our uncertainty about the world (Bresnan and Deo 2000). Linguistic production is influenced by many factors that we would not wish to put into a syntactic model.[38] We cannot know or model all of these, so we are predicting simply that if one averages over all such effects, then certain outputs will occur a certain proportion of the time.

For instance, for the constraints in figure 2, *NonSubj/Ag and *Subj/Newer would sometimes rerank, while a reranking of *NonSubj/Ag and $*3 > 1/2$ would be quite rare, but still occasionally noticeable.[39] In other words, the situation would resemble that observed for the English active/passive choice.[40] One usually gets active forms, as the constraint to disprefer passives usually wins, but sometimes the Discourse constraint would determine the optimal output, with passives being used to get old information into the subject position, and just occasionally the Person constraint would win out, causing a preference for passivization when the patient is local person, even when the outputs differ in how well they meet the Discourse constraint. In some of these last rare cases, the evaluation constraint ranking for English would look just like the regular constraint ranking in Lummi that we saw in (33). This exemplifies the overarching point that variable outputs within grammars can reflect the variation between grammars in a way that theories of soft constraints can illuminate. Indeed, intra-speaker variation appears to be constrained by the same typological markedness factors that play a role in inter-speaker variation (Ihalainen 1991, Cheshire 1991, Cheshire 1996, Cheshire et al. 1989, Cheshire et al. 1993, Schilling-Estes and Wolfram 1994, Anderwald 1999, Bresnan and Deo

[36] It seems reasonable to put aside the short passives (without *by*-phrases) as a separate systemic choice, and at any rate the person of the unexpressed agent cannot be automatically determined with certainty. If we include them, since they overwhelming have 3rd person agents, the results are not materially affected: the Person constraint shines through at least as clearly.

[37] $\chi^2 = 116$, $p < 0.001$, though this test is not necessarily appropriate given the zero cell entries. The more appropriate Fisher's exact test gives $p < 10^{-8}$.

[38] For example, famously Carroll et al. (1981) showed that people's judgement of active versus passive sentences is affected by whether they are sitting in front of a mirror at the time.

[39] Boersma assumes normal curves with a fixed standard deviation of 2 (1 is more standardly used, but the choice is arbitrary). At 2 standard deviations apart (4 points on the scale) reranking occurs about 8% of the time. Note that according to this model nothing is actually categorical, but if constraints are far enough apart, reversals in constraint ranking become vanishingly rare (perhaps one time in a billion a Lummi speaker does utter a sentence with a 3rd person subject and a local person object).

[40] This is just an example grammar with my simple constraint system. See Bresnan et al. (2001) for constructing actual grammars of the English passive using a more developed set of constraints.

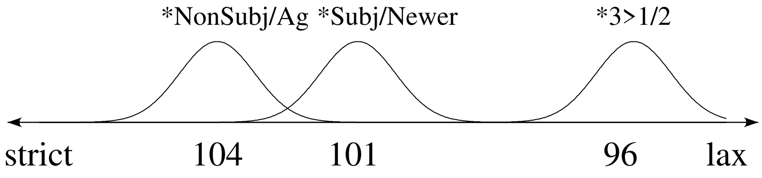

Figure 2: A Stochastic OT model of the English passive. This is a constructed model, not based on real data. Note that the scale is reversed so the highest ranked constraint is to the left, as in standard OT.

2000).

The advantages of the generalization to stochastic evaluation over standard OT include a robust learning algorithm and the capability of learning frequency distributions, which permits a unified account of both variable and categorical data (Boersma and Hayes 2001). In syntax and semantics, this approach has been adopted to explain problems of optionality and ambiguity which arise with non-stochastic OT (Boersma to appear, Bresnan and Deo 2000, Bresnan et al. 2001, Koontz-Garboden 2001, Clark 2001). Like OT, stochastic OT can explain both cross-linguistic implicational generalizations (i.e., typological asymmetries) and within-language "emergence of the unmarked" effects. Stochastic OT thus preserves the core typological predictions of ordinal OT arising from universal subhierarchies of the kind briefly mentioned in section 5.1. This has the consequence that substantive linguistic claims (such as that making an agent a non-subject must be more marked than making a patient a non-subject) are still embodied in the theory. As a consequence, many types of languages (those that categorically violate hierarchies) are not learnable under the theory, thereby accounting for the typological gaps. One cannot get a stochastic OT grammar with the constraint set shown to learn an "anti-Lummi" where passivization occurs when there is a local person agent and a 3rd person patient.

Although this has only just begun to be explored (Bresnan et al. 2001, Clark 2001), Stochastic OT also seems quite promising for modeling historical change. If a process of historical change is modeled by the movement in strength of a constraint along the ranking scale, as implied by the stochastic OT model, then (all else being equal) smooth changes in the relative frequencies of usage are predicted, just as is generally observed in practice. In particular, assuming that the choice between output A or B for an input I is crucially dependent on the difference in ranking of two constraints α and β, and that α is moving at a constant rate to meet and then cross β in ranking, then the Stochastic OT model predicts that the ratio between outputs A and B over time will be a logistic curve, of the sort shown in figure 4. That is, we would expect to see the kind of 'S' curve between the proportion of occurrences of the two outputs that has been widely remarked on in historical and sociolinguistics (Weinreich et al. 1968, Bailey 1973, Kroch 2001). A stochastic grammatical model is in many ways a more plausible model for syntactic change than the competing grammars model prevalent in generative grammar (Kroch 2001). By design (reflecting the orthodoxy of generative grammar), the competing grammars model excludes variation from within grammars, and places it as a choice between several competing grammars. Such a model can only easily generate variable outputs characterized by covariation between phenomena, captured by the choice of whether grammar A or grammar B was selected to generate a sentence. Where parameters vary independently or change at different rates, the competing grammars model requires an exponential number of competing grammars (Guy 1997, Bresnan and Deo 2000), and within-sentence variation (as in (2j)) requires intrasentential switching among them. A stochastic grammar is both more economical, by localizing points of difference, and more expressive, by allowing variation within a single grammar.

A potential limitation of Stochastic OT is that it still does not provide for doing optimization over the combination of all the constraint values – some number of highly ranked constraints at evaluation time will determine the winner, and the rest will be ignored. In particular lower ranked constraint violations cannot "gang up" to defeat a higher ranked constraint. Each will individually occasionally be ranked over the higher constraint, and having multiple violations will cause this to happen more often, but there is no true "ganging up". Providing that the constraints are well spaced, a form that violates 10 lesser ranked constraints will still almost always lose to one that violates one high-ranked constraint. This is potentially problematic as there is some evidence for ganging up effects in syntax (Keller 2000) and phonetics/phonology (Guy 1997, Flemming 2001). Further research is needed to see whether the constrained model of constraint interaction provided by Stochastic OT is adequate for all linguistic systems. Kuhn (2001a, 2001b) suggests that a stochastic OT model is quite promising in generation, when choosing on linguistic grounds between a fairly limited set of candidates, but seems less plausible as a parsing/interpretation model where in general most of the readings of an ambiguous sentence can be made plausible by varying context and lexical items in a way not easily modeled by an OT approach (the decisive evidence can come from many places). This would fit with the facts on the ground, where OT models (stochastic or otherwise) have been mainly employed for generation (though see Kuhn (2001a) for discussion of bidirectional OT), whereas work in NLP, which is mainly on parsing, has tended to use more general feature interaction models.

5.6 Loglinear models

Within speech and natural language processing, there has been a large movement away from categorical linguistic models to statistical models. Most such models are based on either Markov chain models or branching process models (Harris 1963), such as PCFGs (Bod, this volume). An important recent advance in this area has been the application of loglinear models (Agresti 1990) to modeling linguistic systems (Rosenfeld 1994, Ratnaparkhi 1998). In particular, such statistical models can deal with the many interacting dependencies and the structural complexity found in modern syntactic theories, such as constraint-based theories of syntax by allowing arbitrary features to be placed over linguistic representations and then combined into a probability model (Abney 1997, Johnson et al. 1999, Riezler et al. 2000).

In the above OT and stochastic OT frameworks, the output ("grammatical") linguistic structures are those which are optimal among the subset of possible linguistic structures that can be generated to correspond to a certain input. For example, in OT generation, the grammatical linguistic structures are precisely those which are optimal with respect to all other possible structures with the same semantic interpretation or meaning. This corresponds to a *conditional probability distribution* in the case of loglinear models, in which the probability of an output is conditioned on the input semantic interpretation. For example (22), the probability of different outputs o_k for an input i would be modeled as a conditional distribution in terms of weights w_j given to the different features f_j as follows:

$$P(o_k|i) = \frac{1}{Z(i)} e^{w_1 \cdot f_1(o_k,i) + w_2 \cdot f_2(o_k,i) + w_3 \cdot f_3(o_k,i)}$$

Here $Z(i)$ is just a normalization constant – a technical way of making sure a probability distribution results, by scaling the exponential terms to make sure that the sums of the probabilities for all the o_k add to one. Such *exponential models* are also called *maximum entropy models* because an exponential distribution maximizes the entropy of the probability distribution subject to the given constraints, and *loglinear models* because if we take the log of both sides, we have the linear model:

$$\log P(o_k|i) = w_1 \cdot f_1(o_k, i) + w_2 \cdot f_2(o_k, i) + w_3 \cdot f_3(o_k, i) - \log Z(i)$$

While the features used can have arbitrary values, in our example (22), the features are just binary. Especially in this case, the above formula has an easy interpretation: the log of the probability of an output is straightforwardly related to the sum of the weights for the features that are satisfied (i.e., in OT terms, unviolated).

For example if we assume the weights $w_1 = 4$, $w_2 = 3$, and $w_3 = 2$ then we have that:

$$\begin{aligned} \log P(\text{active}) &= w_1 - \log Z = 4 - \log Z \\ \log P(\text{passive}) &= w_2 + w_3 - \log Z = 5 - \log Z \end{aligned}$$

Assuming that these are the only possible outputs for this input, we can easily calculate the normalization term Z (using $\frac{1}{Z}(e^5 + e^4) = 1$ to obtain Z), and get the result that:

$$\begin{aligned} P(\text{active}) &= 0.27 \\ P(\text{passive}) &= 0.73 \end{aligned}$$

There are actually two ways we could use this result. By optimizing over loglinear models – that is by taking the most probable structure – $\arg\max_k P(o_k|i)$ – one can determine a unique output that best satisfies multiple conflicting constraints, in a manner similar to OT, but allowing for arbitrary ganging up of features. That is, for this particular input and these weights, the model would always choose an output of passive. But note crucially that it would do this because constraints f_2 and f_3 gang up to beat out constraint f_1, which is the highest weighted constraint. Alternatively, by using the model as a probability distribution over outputs, we would get variable outputs in the style of Stochastic OT, but again with more flexibility in constraint interaction than the systems currently employed in linguistics. Our prediction would be that one would get a passive output about 3/4 of the time, and an active output 1/4 of the time for this configuration. Given actual data on how often actives and passives occur for various input feature specifications, we can use fitting algorithms to automatically find the weights w_i which best fit the data. We still get the effect of "emergence of the unmarked": if all outputs share the same constraint violations or satisfactions for highly weighted constraints, then the optimal candidate (under the first interpretation) or the candidate that is most commonly output (under the second interpretation) will be determined by low ranked constraints on which the various output candidates differ. However, there has at present not been much research into how typological restrictions on the space of possible grammars can be built into loglinear models.

5.7 Generalized Linear Models

In the last couple of decades, there has been a revolution in the way statisticians standardly model categorical count/proportion data of the kind that is most commonly found in linguistics (that is, under the assumption that one's categories are discrete). The traditional tools were some simple parametric and non-parametric test statistics, of which by far the best known is the (Pearson) chi-squared (χ^2) test. These approaches sat as an underdeveloped sibling beside the sophisticated theory of linear models for continuous variables. Most people will have seen at least simple linear regression, where we have a set of points (x_1, y_1), (x_2, y_2), …, (x_n, y_n), as in figure 3. We would like to find the line:

$$f(x) = mx + b$$

with parameters m and b that fits these points best (in the sense of minimizing the squared error indicated by the arrow lengths in the figure).

This can easily be extended to a multiple linear regression model, where the response variable (y_i) is predicted from a variety of explanatory x_{ij} variables, or even suitable functions of them. For example, we might have multiple linear regression models that look like one of these:

$$\begin{aligned} f(y|x_1, x_2) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 \\ f(y|x_1, x_2) &= \beta_0 + \beta_1 \log(x_1) + \beta_2 x_2 + \beta_3 x_2^2 \end{aligned}$$

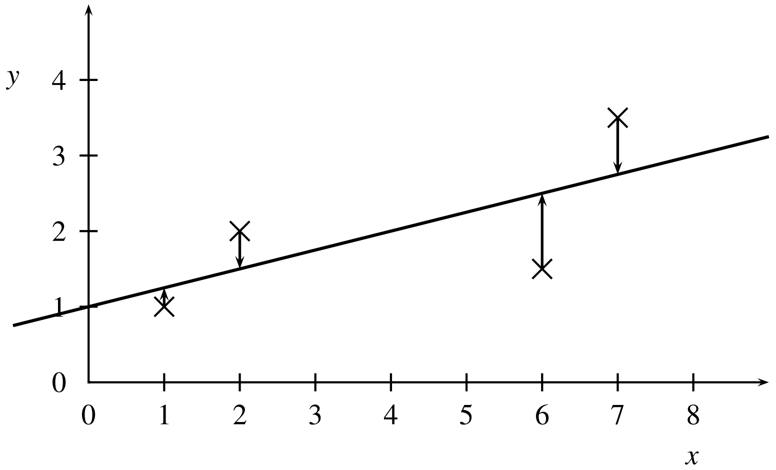

Figure 3: An example of linear regression. The line $y = 0.25x + 1$ is the best least-squares fit for the points (1,1), (2,2), (6,1.5), (7,3.5). Arrows show the y values for each x value given by the model.

Indeed such a multiple linear regression model was suggested for variable rules in Labov's earliest sociolinguistic work (Labov 1969):[41]

$$p = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

Here the x_i are indicator variables associated with a certain environmental feature, the β_i are weights, and p is the probability of a rule applying. However, this model is inappropriate when the aim is for the response variable to be a probability of rule application. Suppose that there are three explanatory variables which individually strongly suggest a certain result, say with a probability of 0.8. If we set the corresponding β_i weights to around 0.8, the problem is that if all three of these features were present, then, when linearly combined, the value for the response variable would exceed 1: an invalid value for a probability. Since this cannot be allowed, the best fit values in this model would be to set the β_i to a value around 0.3, which would mean that when only one of the features is present, the predictions of the model would be very poor (predicting only a 30% rather than 80% chance of the result). The immediate response to this was to consider multiplicative models of application and non-application (Cedergren and Sankoff 1974), but these have generally been replaced by logistic regression models (Rousseau and Sankoff 1978, Sankoff 1988).

The logistic regression model is one common case of a *generalized linear model* (GLM), suitable for modeling binary response variables. The general approach of generalized linear models has transformed the study of categorical data by bringing all the tools of traditional linear regression to the categorical case. The way this is done is by generalizing the linear model by allowing the combined value of the explanatory variables to be equal to some function of the response variable, termed the *link function* g. So we have:

$$g(y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

For traditional linear regression (with a normally distributed response), g is the identity function. By choosing a log function for g, we obtain the loglinear models of the preceding section – a product of factors

[41] Also, Keller (2000) uses multiple linear regression to model gradient grammaticality in his Linear Optimality Theory model. (Noticing that the magnitude estimation judgements of speakers are log-transformed, one could say he is using a loglinear model, as below, but since the output values are arbitrary, rather than being probabilities, no normalization is necessary, and hence the fitting problem remains standard least squares, rather than requiring more complex techniques, as for either Harmony Grammar or the loglinear models discussed above.)

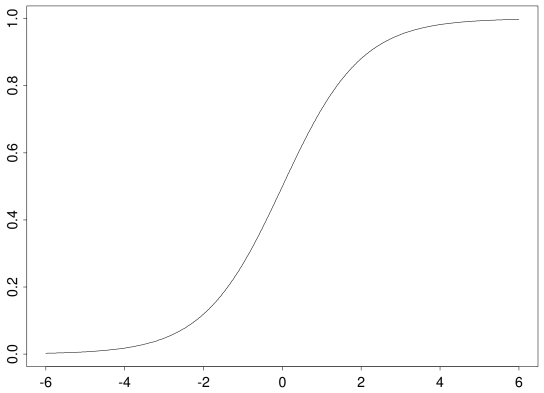

Figure 4: The logistic function

gives a sum of log factors. For a binary response variable, the appropriate link function is the logit function $\text{logit}(\pi) = \log[\pi/(1-\pi)]$. The inverse of the logit function is the logistic function. If $\text{logit}(\pi) = z$, then

$$\pi = \frac{e^z}{1+e^z}$$

The logistic function will map any value of the right hand side (z) to a proportion value between 0 and 1, as shown in figure 4. It is this logit model that has been the basis of VARBRUL sociolinguistic modeling since the late 1970s. An example with relevance to the study in this section is Weiner and Labov (1983), which presents a study of factors predictive of active and passive in 'agentless' sentences.

There is not space here to give a detailed treatment of generalized linear models. There are many good books that cover generalized linear models in detail, though not always in a way approachable to linguists (Bishop et al. 1977 is the classic, Agresti 1990 and McCullagh and Nelder 1990 are standard references, Lloyd 1999 is more recent, Fienberg 1980 and Agresti 1996 are more approachable, Powers and Xie 1999 is directed towards social scientists, and some recent general statistical methods books such as Ramsey and Schafer 1997 give elementary coverage). However, since they are likely to continue to be so important to and useful in the development of probabilistic models over linguistic data, it is useful to understand at least the main idea, and how other models relate to them.

The logistic regression model is appropriate for binary response variables or proportions based on binomial counts, that is when the number of trials is fixed. It is thus the natural model to use in the Variable Rules approach, where one is predicting the application or non-application of a particular rule. However, another very natural thing to do is to simply collect a contingency table of counts for how often various linguistic tokens appear, organized into dimensions according to how they do or do not satisfy various constraints. For such a model, there is no natural limit to the counts in one cell, and the loglinear model, with a log link function is appropriate. This model can be fit as a multinomial model by using iterative proportional fitting methods (Darroch and Ratcliff 1972) or general minimization techniques, such as conjugate gradient descent (Johnson et al. 1999).

Within a generalized linear model approach, an OT grammar is a special case, where the weight of each successive constraint is so much smaller than the preceding constraints that there is no opportunity for ganging up – the highest-weighted differentiating feature always determines things. The details differ slightly between the logistic and loglinear cases, but this connection has been noted multiple times. Rousseau and Sankoff (1978:66–67) discuss the possibility of this as a "whole hierarchy of successively weaker knockout features". Prince and Smolensky (1993:200) point out, "Optimality Theory … represents a very specialized kind of Harmonic Grammar, with exponential weighting of the constraints". Johnson (1998) shows how standard OT (with a bound on the number of constraint violations) corresponds to loglinear models with the constraint weights being far enough apart that constraints do not interact. The last two are equivalent in that Harmonic connectionist nets approximate loglinear models (Smolensky and Legendre 2000).

The example presented here is excessively simple (for expository purposes): there are just three constraints, designed to work over simple transitive sentences. I should therefore stress that these models can scale up. This has not been much demonstrated for linguistic goals, but within NLP, loglinear models with well over 100,000 features are regularly being deployed for tasks of parsing and disambiguation (Ratnaparkhi 1999, Toutanova and Manning 2000). Essentially, the models are based on *sufficient statistics* which are counts of how often certain things happen or fail to happen in a sentence. These constraints can be any evaluable function, which can probe arbitrary aspects of sentence structure. For instance, Johnson et al. (1999) place features over grammatical relations, argument/adjunct status, low vs. high attachment, etc. Essentially, we are left with the customary linguistic task of finding the right constraints on linguistic goodness. Once we have them, they can be automatically weighted within a loglinear model.

6 Conclusion

There are many phenomena in syntax that cry out for non-categorical and probabilistic modeling and explanation. The opportunity to leave behind ill-fitting categorical assumptions, and to better model probabilities of use in syntax is exciting. The existence of 'soft' constraints within the variable output of an individual speaker, of exactly the same kind as the typological syntactic constraints found across languages, makes exploration of probabilistic grammar models compelling. We saw that one is not limited to simple surface representations: I have tried to outline how probabilistic models can be applied on top of one's favorite sophisticated linguistic representations. The frequency evidence needed for parameter estimation in probabilistic models requires a lot more data collection, and a lot more careful evaluation and model building than traditional syntax, where one example can be the basis of a new theory, but the results can enrich linguistic theory by revealing the soft constraints at work in language use. This is an area ripe for exploration by the next generation of syntacticians.

Bibliography

Abney, S. 1996. Statistical methods and linguistics. In J. L. Klavans and P. Resnik (Eds.), *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, 1–26. Cambridge, MA: MIT Press.

Abney, S. P. 1997. Stochastic attribute-value grammars. *Computational Linguistics* 23(4):597–618.

Ackema, P., and A. Neeleman. 1998. Conflict resolution in passive formation. *Lingua* 104:13–29.

Agresti, A. 1990. *Categorical Data Analysis*. New York: John Wiley & Sons.

Agresti, A. 1996. *An Introduction to Categorical Data Analysis*. New York: John Wiley & Sons.

Aissen, J. 1999. Markedness and subject choice in Optimality Theory. *Natural Language and Linguistic Theory* 17:673–711.

Anderwald, L. 1999. *Negation in non-standard British English*. PhD thesis, University of Freiburg.

Anttila, A. 2001. Variation and phonological theory. In J. K. Chambers, P. Trudgill, and N. Schilling-Estes (Eds.), *Handbook of Language Variation and Change*. Oxford: Blackwell.

Atkins, B. T. S., and B. C. Levin. 1995. Building on a corpus: a linguistic and lexicographical look at some near-synonyms. *International Journal of Lexicography* 8:85–114.

Babby, L. 1980. The syntax of surface case marking. In W. Harbert and J. Herschensohn (Eds.), *Cornell Working Papers in Linguistics*, 1–32. Department of Modern Languages and Linguistics, Cornell University.

Bailey, C.-J. N. 1973. *Variation and Linguistic Theory*. Arlington, VA: Center for Applied Linguistics.

Barlow, M., and S. Kemmer (Eds.). 2000. *Usage-Based Models of Language*. Stanford, CA: CSLI Publications.

Bishop, C. M. 1995. *Neural Networks for Pattern Recognition*. Oxford: Oxford University Press.

Bishop, Y., S. E. Fienberg, and P. W. Holland. 1977. *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press.

Bod, R. 1998. *Beyond Grammar: An experience-based theory of language*. Stanford, CA: CSLI Publications.

Bod, R., and R. Kaplan. 1998. A probabilistic corpus-driven model for lexical-functional analysis. In *ACL 36/COLING 17*, 145–151.

Boersma, P. 1997. How we learn variation, optionality, and probability. *Proceedings of the Institute of Phonetic Sciences* 21:43–58.

Boersma, P. 1998. *Functional Phonology. Formalizing the interactions between articulatory and perceptual drives*. PhD thesis, University of Amsterdam.

Boersma, P. to appear. Phonology-semantics interaction in OT, and its acquisition. In R. Kirchner, W. Wikeley, and J. Pater (Eds.), *Papers in Experimental and Theoretical Linguistics*, Vol. 6. Edmonton: University of Alberta.

Boersma, P., and B. Hayes. 2001. Empirical tests of the gradual learning algorithm. *Linguistic Inquiry* 32(1):45–86.

Booth, T. L., and R. A. Thomson. 1973. Applying probability measures to abstract languages. *IEEE Transactions on Computers* C-22:442–450.

Bresnan, J. 2001. *Lexical-Functional Syntax*. Oxford: Blackwell.

Bresnan, J., and A. Deo. 2000. 'Be' in the *Survey of English Dialects*: A stochastic OT account. Paper presented at the Symposium on Optimality Theory, English Linguistic Society of Japan, November 18, 2000, Kobe, Japan.

Bresnan, J., S. Dingare, and C. D. Manning. 2001. Soft constraints mirror hard constraints: Voice and person in English and Lummi. In *Proceedings of the LFG 01 Conference*. CSLI Publications.

Bresnan, J., and L. Moshi. 1990. Object asymmetries in comparative Bantu syntax. *Linguistic Inquiry* 21(2):147–186.

Bresnan, J., and A. Zaenen. 1990. Deep unaccusativity in LFG. In K. Dziwirek, P. Farrell, and E. Mejías-Bikandi (Eds.), *Grammatical Relations: A Cross-Theoretical Perspective*, 45–57. Stanford, CA: CSLI Publications.

Carpenter, B. 1992. *The Logic of Typed Feature Structures*. Cambridge: Cambridge University Press.

Carroll, J. M., T. G. Bever, and C. R. Pollack. 1981. The non-uniqueness of linguistic intuitions. *Language* 57:368–383.

Cedergren, H. J., and D. Sankoff. 1974. Variable rules: Performance as a statistical reflection of competence. *Language* 50:333–355.

Chambers, J. K. 1995. *Sociolinguistic Theory: Linguistic Variation and its Social Significance*. Cambridge, MA: Blackwell.

Charniak, E. 1997. Statistical parsing with a context-free grammar and word statistics. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI '97)*, 598–603.

Cheshire, J. 1991. Variation in the use of *ain't* in an urban British English dialect. In P. Trudgill and J. K. Chambers (Eds.), *Dialects of English. Studies in Grammatical Variation*, 54–73. London: Longman.

Cheshire, J. 1996. Syntactic variation and the concept of prominence. In *Speech Past and Present. Essays in English Dialectology in Memory of Ossi Ihalainen*, 1–17. Frankfurt: Peter Lang.

Cheshire, J., V. Edwards, and P. Whittle. 1989. Urban British dialect grammar: the question of dialect levelling. *English World-Wide* 10:185–225.

Cheshire, J., V. Edwards, and P. Whittle. 1993. Non-standard English and dialect levelling. In J. Milroy and L. Milroy (Eds.), *Real English. The Grammar of English Dialects in the British Isles*, 53–96. London: Longman.

Chomsky, N. 1955. The logical structure of linguistic theory. Published by Plenum Press, New York, 1975.

Chomsky, N. 1956. Three models for the description of language. *IRE Transactions on Information Theory* IT-2:113–124.

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

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

Chomsky, N. 1969. Quine's empirical assumptions. In D. Davidson and J. Hintikka (Eds.), *Words and Objections: Essays on the Work of W.V. Quine*, 53–68. Dordrecht: D. Reidel.

Chomsky, N. 1979. *Language and Responsibility: Based on Conversations with Mitsou Ronat*. New York: Pantheon. Translated by John Viertel.

Chomsky, N. 1981. *Lectures on Government and Binding*. Dordrecht: Foris.

Chomsky, N. 1986. *Barriers*. Cambridge, MA: MIT Press.

Chomsky, N. 1995. *The Minimalist Program*. Cambridge, MA: MIT Press.

Clark, B. Z. 2001. A stochastic optimality theory approach to phrase structure variation and change in early English. Paper presented at the Berkeley Historical Syntax Workshop. MS, Stanford University.

Collins, M. J. 1997. Three generative, lexicalised models for statistical parsing. In *ACL 35/EACL 8*, 16–23.

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

Culy, C. 1998. Statistical distribution and the grammatical/ungrammatical distinction. *Grammars* 1:1–13.

Darroch, J. N., and D. Ratcliff. 1972. Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics* 43(3):1470–1480.

Dempster, A., N. Laird, and D. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statistical Society Series B* 39:1–38.

Dingare, S. 2001. The effect of feature hierarchies on frequencies of passivization in English. Master's thesis, Department of Linguistics, Stanford University.

Engdahl, E. 1983. Parasitic gaps. *Linguistic Inquiry* 6:5–34.

Estival, D., and J. Myhill. 1988. Formal and functional aspects of the development from passive to ergative systems. In M. Shibatani (Ed.), *Passive and Voice*, 441–491. Amsterdam: John Benjamins.

Evelyn, J. 1664. *Sylva, Or a discourse of Forest-Trees, and the Propagation of Timber in His Majesties Dominions, &c.* London: Printed by Jo. Martyn, and Ja. Allestry, for the Royal Society.

Feldman, J. A., J. Gips, J. J. Horning, and S. Reder. 1969. Grammatical complexity and inference. Technical Report CS 125, Stanford University, Computer Science Dept, Stanford, CA.

Fienberg, S. E. 1980. *The Analysis of Cross-Classified Categorical Data*. Cambridge, MA: MIT Press. 2nd edition.

Fillmore, C. J. 1992. 'Corpus linguistics' or 'computer-aided armchair linguistics'. In J. Svartvik (Ed.), *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82: Stockholm, 4-8 August 1991*, 35–60. Berlin: Mouton de Gruyter.

Flemming, E. 2001. Scalar and categorical phenomena in a unified model of phonetics and phonology. *Phonology* 18(1). To appear.

Fowler, G. H. 1987. *The Syntax of Genitive Case in Russian*. PhD thesis, University of Chicago.

Fowler, H. W. 1926. *A dictionary of modern English usage*. Oxford: Clarendon Press. [1st ed.; revised ed. 1948, 1954, etc.].

Givón, T. 1979. *On Understanding Grammar*. New York: Academic Press.

Givón, T. 1994. The pragmatics of de-transitive voice: Functional and typological aspects of inversion. In T. Givón (Ed.), *Voice and Inversion*, 3–44. Amsterdam: John Benjamins.

Gleason, H. A. 1961. *An introduction to descriptive linguistics*. New York: Holt, Rinehart and Winston. revised edition.

Godfrey, J., E. Holliman, and J. McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Proceedings of ICASSP-92*, 517–520.

Gold, E. M. 1967. Language identification in the limit. *Information and Control* 10:447–474.

Gowers, E. 1948. *Plain words: a guide to the use of English*. London: Her Majesty's Stationery Office.

Greenberg, J. 1966. *Language Universals: With Special Reference to Feature Hierarchies*. The Hague: Mouton.

Grimshaw, J. 1979. Complement selection and the lexicon. *Linguistic Inquiry* 10:279–326.

Grimshaw, J. 1990. *Argument Structure*. Cambridge, MA: MIT Press.

Grimshaw, J., and S. Vikner. 1993. Obligatory adjuncts and the structure of events. In E. Reuland and W. Abraham (Eds.), *Knowledge and Language*, Vol. II, 143–155. Dordrecht: Kluwer.

Grishman, R., C. Macleod, and A. Meyers. 1994. COMLEX syntax: Building a computational lexicon. In *COLING 15*.

Guy, G. 1981. *Linguistic variation in Brazilian Portuguese: aspects of the phonology, syntax, and language history*. PhD thesis, University of Pennsylvania.

Guy, G. 1997. Violable is variable: OT and linguistic variation. *Language Variation and Change* 9:333–348.

Halliday, M. A. K. 1994. *An introduction to functional grammar*. London: Edward Arnold. 2nd edition.

Harris, T. E. 1963. *The Theory of Branching Processes*. Berlin: Springer.

Hastie, T., R. Tibshirani, and J. H. Friedman. 2001. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer Verlag.

Hornby, A. S. 1989. *Oxford Advanced Learner's Dictionary of Current English*. Oxford: Oxford University Press. 4th edition.

Horning, J. J. 1969. *A study of grammatical inference*. PhD thesis, Stanford.

Hukari, T. E., and R. D. Levine. 1995. Adjunct extraction. *Journal of Linguistics* 31:195–226.

Ihalainen, O. 1991. On grammatical diffusion in Somerset folk speech. In P. Trudgill and J. K. Chambers (Eds.), *Dialects of English. Studies in Grammatical Variation*, 104–19. London: Longman.

Jelinek, E., and R. Demers. 1983. The agent hierarchy and voice in some Coast Salish languages. *International Journal of American Linguistics* 49:167–185.

Jelinek, E., and R. Demers. 1994. Predicates and pronominal arguments in Straits Salish. *Language* 70:697–736.

Jensen, F. V., and F. B. Jensen. 2001. *Bayesian Networks and Decision Graphs*. Berlin: Springer Verlag.

Johnson, M. 1998. Optimality-theoretic lexical functional grammar. Commentary on Joan Bresnan's presentation at the 1998 CUNY Sentence Processing conference.

Johnson, M., S. Geman, S. Canon, Z. Chi, and S. Riezler. 1999. Estimators for stochastic "unification-based" grammars. In *ACL 37*, 535–541.

Joos, M. 1950. Description of language design. *The Journal of the Acoustical Society of America* 22(6):701–708.

Joshi, A. K., and Y. Schabes. 1997. Tree-adjoning grammars. In G. Rozenberg and A. Salomaa (Eds.), *Handbook of Formal Languages*, 69–123. Berlin: Springer-Verlag.

Kato, K. 1979. Empathy and passive resistance. *Linguistic Inquiry* 10:149–152.

Keller, F. 2000. *Gradience in Grammar: Experimental and Computational Aspects of Degrees of Grammaticality*. PhD thesis, University of Edinburgh.

Kersten, D. 1999. High-level vision as statistical inference. In M. S. Gazzaniga (Ed.), *The New Cognitive Neurosciences*. Cambridge, MA: MIT Press.

Kiparsky, P. 1973. "Elsewhere" in phonology. In S. R. Anderson (Ed.), *Festschrift for Morris Halle*, 93–106. New York: Holt, Rinehart and Winston.

Koontz-Garboden, A. J. 2001. A stochastic OT approach to word order variation in Korlai Portuguese. In *Proceedings of the 37th annual meeting of the Chicago Linguistic Society*, Chicago, IL.

Kroch, A. S. 2001. Syntactic change. In M. Baltin and C. Collins (Eds.), *Handbook of Contemporary Syntactic Theory*, 699–729. Oxford: Blackwell.

Kroeger, P. 1993. *Phrase Structure and Grammatical Relations in Tagalog*. Stanford, CA: CSLI Publications.

Kuhn, J. 2001a. *Formal and Computational Aspects of Optimality-theoretic Syntax*. PhD thesis, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart.

Kuhn, J. 2001b. Sketch of a hybrid optimization architecture of grammar in context. MS, Stanford University.

Kuno, S., and E. Kaburaki. 1977. Empathy and syntax. *Linguistic Inquiry* 8:627–672.

Labov, W. 1966. *The Social Stratification of English in New York City*. Washington, DC: Center for Applied Lingusitics.

Labov, W. 1969. Contraction, deletion, and inherent variability of the English copula. *Language* 45:715–762.

Labov, W. 1972. *Sociolinguistic Patterns*. Philadelphia, PA: University of Pennsylvania Press.

Labov, W. 1994. *Principles of linguistic change: Volume 1, Internal factors*. Cambridge, MA: Blackwell.

Legendre, G., Y. Miyata, and P. Smolensky. 1990. Can connectionism contribute to syntax? Harmonic grammar, with an application. In *Proceedings of the 26th Meeting of the Chicago Linguistic Society*.

Legendre, G., W. Raymond, and P. Smolensky. 1993. An optimality-theoretic typology of case and grammatical voice systems. In *Proceedings of the 19th Annual Meeting of the Berkeley Linguistics Society*, 464–478.

Levin, B. 1993. *English Verb Classes and Alternations*. Chicago: University of Chicago Press.

Li, A. Y.-H. 1990. *Order and Constituency in Mandarin Chinese*. Dordrecht: Kluwer Academic Publishers.

Lloyd, C. J. 1999. *Statistical Analysis of Categorical Data*. John Wiley & Sons.

Lødrup, H. 1999. Linking and optimality in the Norwegian presentational focus construction. *Nordic Journal of Linguistics* 22:205–230.

Luce, R. D., R. R. Bush, and E. Galanter (Eds.). 1963. *Handbook of Mathematical Psychology*. Vol. II. New York: John Wiley and Sons.

Maling, J. 1989. Adverbials and structural case in Korean. In S. Kuno, I. Lee, J. Whitman, S.-Y. Bak, Y.-S. Kang, and Y.-J. Kim (Eds.), *Harvard Studies in Korean Linguistics*, Vol. 3, 297–308, Cambridge, MA.

Maling, J. 1993. Of nominative and accusative: The hierarchical assignment of grammatical case in Finnish. In A. Holmberg and U. Nikanne (Eds.), *Case and Other Functional Categories in Finnish Syntax*, 51–76. Dordrecht: Mouton de Gruyter.

Malouf, R. P. 2000. *Mixed Categories in the Hierarchical Lexicon*. Stanford, CA: CSLI Publications.

Manning, C. D. 1993. Automatic acquisition of a large subcategorization dictionary from corpora. In *ACL 31*, 235–242.

Manning, C. D., and H. Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Boston, MA: MIT Press.

Marcus, M. P., B. Santorini, and M. A. Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn treebank. *Computational Linguistics* 19:313–330.

McCarthy, J., and P. Hayes. 1969. Some philosophical problems from the standpoint of artificial intelligence. In B. Meltzer and D. Michie (Eds.), *Machine Intelligence*, Vol. 4, 463–502. Edinburgh: Edinburgh University Press.

McCullagh, P., and J. A. Nelder. 1990. *Generalized Linear Models*. CRC Press. 2nd edition.

McEnery, T., and A. Wilson. 2001. *Corpus Linguistics*. Edinburgh: Edinburgh University Press. 2nd edition.

McLachlan, G. J., and T. Krishnan. 1996. *The EM Algorithm and Extensions*. John Wiley & Sons.

Merlo, P., and M. Leybold. 2001. Automatic distinction of arguments and modifiers: the case of prepositional phrases. In *Proceedings of the Workshop on Computational Natural Language Learning (CoNLL-2001)*, 121–128.

Mitchell, T. M. (Ed.). 1997. *Machine Learning*. New York: McGraw-Hill.

Morrill, G. 1994. *Type-Logical Grammar*. Dordrecht: Kluwer Academic.

Müller, G. 1999. Optimality, markedness, and word order in German. *Linguistics* 37:777–818.

Mumford, D. 1999. The dawning of the age of stochasticity. Based on a lecture at the Accademia Nazionale dei Lincei. Available at: http://www.dam.brown.edu/people/mumford/Papers/Dawning.ps.

Nagy, N., and B. Reynolds. 1997. Optimality theory and variable word-final deletion in Faetar. *Language Variation and Change* 9(1):37–56.

Napoli, D. J. 1981. Semantic interpretation vs. lexical governance. *Language* 57:841–887.

Nelson, G., S. Wallis, and B. Aarts. forthcoming. *Exploring Natural Language: Working with the British Component of the International Corpus of English*. Amsterdam: John Benjamins.

Olofsson, A. 1990. A participle caught in the act. on the prepositional use of *following*. *Studia Neophilologica* 62:23–35.

Pereira, F. 2000. Formal grammar and information theory: Together again. *Philosophical Transactions of the Royal Society* 358:1239–1253.

Pesetsky, D. 1998. Principles of sentence pronunciation. In P. Barbosa, D. Fox, P. Hagstrom, M. McGinnis, and D. Pesetsky (Eds.), *Is the Best Good Enough*. Cambridge, MA: MIT Press.

Pinker, S. 2000. *Words and Rules: The Ingredients of Language*. Harper Perennial.

Pollard, C., and I. A. Sag. 1987. *Information-Based Syntax and Semantics*. Vol. 1. Stanford, CA: Center for the Study of Language and Information.

Pollard, C., and I. A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. Chicago, IL: University of Chicago Press.

Powers, D. A., and Y. Xie. 1999. *Statistical Methods for Categorical Data Analysis*. Academic Press.

Prince, A., and P. Smolensky. 1993. Optimality theory: Constraint interaction in generative grammar. Technical Report TR-2, Rutgers University Center for Cognitive Science.

Przepiórkowski, A. 1999a. On case assignment and "adjuncts as complements". In G. Webelhuth, J.-P. Koenig, and A. Kathol (Eds.), *Lexical and Constructional Aspects of Linguistic Explanation*, 231–245. Stanford, CA: CSLI Publications.

Przepiórkowski, A. 1999b. On complements and adjuncts in Polish. In R. D. Borsley and A. Przepiórkowski (Eds.), *Slavic in Head-Driven Phrase Structure Grammar*, 183–210. Stanford, CA: CSLI Publications.

Pullum, G. K. 1996. Learnability, hyperlearning, and the poverty of the stimulus. In J. Johnson, M. L. Juge, and J. L. Moxley (Eds.), *Proceedings of the 22nd Annual Meeting of the Berkeley Linguistics Society: General Session and Parasession on the Role of Learnability in Grammatical Theory*, 498–513. Berkeley, CA: Berkeley Linguistics Society.

Quirk, R., S. Greenbaum, G. Leech, and J. Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Radford, A. 1988. *Transformational Grammar*. Cambridge: Cambridge University Press.

Ramsey, F. L., and D. W. Schafer. 1997. *The Statistical Sleuth: A Course in Methods of Data Analysis*. Belmont, CA: Duxbury Press.

Ratnaparkhi, A. 1998. *Maximum Entropy Models for Natural Language Ambiguity Resolution*. PhD thesis, University of Pennsylvania.

Ratnaparkhi, A. 1999. Learning to parse natural language with maximum entropy models. *Machine Learning* 34:151–175.

Resnik, P. 1996. Selectional constraints: an information-theoretic model and its computational realization. *Cognition* 61:127–159.

Riezler, S., D. Prescher, J. Kuhn, and M. Johnson. 2000. Lexicalized stochastic modeling of constraint-based grammars using log-linear measures and EM. In *ACL 38*.

Ripley, B. D. 1996. *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.

Rizzi, L. 1990. *Relativized minimality*. Cambridge, MA: MIT Press.

Roland, D., and D. Jurafsky. 1998. How verb subcategorization frequencies are affected by corpus choice. In *ACL 36/COLING 17*, 1122–1128.

Rosenfeld, R. 1994. *Adaptive Statistical Language Modeling: A Maximum Entropy Approach*. PhD thesis, CMU. Technical report CMU-CS-94-138.

Ross, J. R. 1972. The category squish: Endstation Hauptwort. In *Papers from the Eighth Regional Meeting*, 316–328, Chicago. Chicago Linguistic Society.

Ross, J. R. 1973a. Clausematiness. In E. L. Keenan (Ed.), *Formal Semantics of Natural Language*. Cambridge UK: Cambridge University Press.

Ross, J. R. 1973b. A fake NP squish. In C.-J. N. Bailey (Ed.), *New ways of analyzing variation in English*. Washington: Georgetown University Press.

Ross, J. R. 1973c. Nouniness. In O. Fujimura (Ed.), *Three dimensions of linguistic theory*, 137–257. Tokyo: The Tokyo English Corporation.

Rousseau, P., and D. Sankoff. 1978. Advances in variable rule methodology. In D. Sankoff (Ed.), *Linguistic Variation: Models and Methods*. New York: Academic Press.

Russo, R. 2001. *Empire Falls*. New York: Alfred A. Knopf.

Sampson, G. 1987. Evidence against the "grammatical"/"ungrammatical" distinction. In W. Meijs (Ed.), *Corpus Linguistics and Beyond*. Amsterdam: Rodopi.

Sampson, G. 2001. *Empirical Linguistics*. London: Continuum International.

Sankoff, D. 1988. Variable rules. In U. Ammon, N. Dittmar, and K. J. Mattheier (Eds.), *Sociolinguistics: An International Handbook of the Science of Language and Society*, Vol. 2, 984–997. Berlin: Walter de Gruyter.

Santorini, B. 1990. Part-of-speech tagging guidelines for the Penn treebank project. 3rd Revision, 2nd printing, Feb. 1995. University of Pennsylvania.

Sapir, E. 1921. *Language: an introduction to the study of speech*. New York: Harcourt Brace.

Schilling-Estes, N., and W. Wolfram. 1994. Convergent explanation and alternative regularization patterns: *Were/weren't* leveling in a vernacular English variety. *Language Variation and Change* 6:273–302.

Schütze, C. T. 1995. PP attachment and argumenthood. In *MIT Working Papers in Linguistics: Papers on Language Processing and Acquisition*, Vol. 26, 95–151. Cambridge, MA: MITWPL.

Schütze, C. T. 1996. *The empirical base of linguistics: grammaticality judgments and linguistic methodology*. Chicago, IL: University of Chicago Press.

Seidenberg, M. S., and M. C. MacDonald. 1999. A probabilistic constraints approach to language acquisition and processing. *Cognitive Science* 23:569–588.

Shannon, C. E. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27:379–423, 623–656.

Silverstein, M. 1976. Hierarchy of features and ergativity. In R. M. W. Dixon (Ed.), *Grammatical Categories in Australian Languages*, 112–171. Canberra: Australian Institute of Aboriginal Studies.

Sinclair, J. M. 1997. Corpus evidence in language description. In A. Wichmann, S. Fligelstone, T. McEnery, and G. Knowles (Eds.), *Teaching and Language Corpora*. Longman.

Smolensky, P. 1986. Information processing in dynamical systems: Foundations of harmony theory. In D. E. Rumelhart, J. L. McClelland, and the PDP Research Group (Eds.), *Parallel Distributed Processing. Explorations in the Microstructure of Cognition. Volume 1: Foundations*, 194–281. Cambridge, MA: The MIT Press.

Smolensky, P. 1993. Harmony, markedness, and phonological activity. Handout of keynote address, Rutgers Optimality Workshop-1; ROA-87, October.

Smolensky, P. 1997. Constraint interaction in generative grammar II: Local conjunction, or random rules in universal grammar. Presented at the Hopkins Optimality Theory Workshop/University of Maryland Mayfest, May.

Smolensky, P. 1999. Principles of Dave's philosophy. Contribution to the David Rumelhart Celebration at Carnegie-Mellon University, October.

Smolensky, P., and G. Legendre. 2000. Architecture of the mind/brain: Neural computation, optimality, and universal grammar in cognitive science. Draft ms., Dec 1, 2000, Johns Hopkins University.

Sokolov, J. L., and C. E. Snow. 1994. The changing role of negative evidence in theories of language development. In C. Gallaway and B. J. Richards (Eds.), *Input and interaction in language acquisition*, 38–55. New York: Cambridge University Press.

Sorace, A. 2000. Gradients in auxiliary selection with intransitive verbs. *Language* 76:859–890.

Stabler, E. P. 2001. *Computational Minimalism: Acquiring and Parsing Languages With Movement*. Blackwell.

Svartvik, J. 1966. *On voice in the English verb*. The Hague: Mouton.

Tabor, W. 2000. Lexical categories as basins of curvature. MS, University of Connecticut.

Taylor, L., C. Grover, and E. Briscoe. 1989. The syntactic regularity of English noun phrases. In *EACL 4*, 256–263.

Tenenbaum, J. B. 1999. Bayesian modeling of human concept learning. In M. S. Kearns, S. A. Solla, and D. A. Cohn (Eds.), *Advances in Neural Information Processing Systems*, Vol. 11, Cambridge, MA. MIT Press.

Tesnière, L. 1959. *Éléments de Syntaxe Structurale*. Paris: Librairie C. Klincksieck.

Toutanova, K., and C. D. Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *EMNLP/VLC 2000*, 63–70.

Trask, R. L. 1979. On the origins of ergativity. In F. Plank (Ed.), *Ergativity: Towards a theory of grammatical relations*, 385–404. London: Academic Press.

Vater, H. 1978. On the possibility of distinguishing between complements and adjuncts. In W. Abraham (Ed.), *Valence, Semantic Case and Grammatical Relations*, 21–45. Amsterdam: John Benjamins.

Verspoor, C. M. . 1997. *Contextually-Dependent Lexical Semantics*. PhD thesis, Edinburgh.

Wallis, S. A., B. Aarts, and G. Nelson. 1999. Parsing in reverse – exploring ICE-GB with fuzzy tree fragments and ICECUP. In J. M. Kirk (Ed.), *Corpora Galore: papers from the 19th International Conference on English Language Research on Computerised Corpora, ICAME-98*, 335–344. Amsterdam: Rodopi.

Wasow, T. 1997. Remarks on grammatical weight. *Language Variation and Change* 9:81–105.

Wechsler, S., and Y.-S. Lee. 1996. The domain of direct case assignment. *Natural Language and Linguistic Theory* 14:629–664.

Weiner, E. J., and W. Labov. 1983. Constraints on the agentless passive. *Journal of Linguistics* 19:29–58.

Weinreich, U., W. Labov, and M. Herzog. 1968. Empirical foundations for a theory of language change. In W. Lehmann and Y. Malkiel (Eds.), *Directions for Historical Linguistics*, 95–188. Austin, TX: University of Texas Press.

Zaenen, A. 1993. Unaccusativity in Dutch: Integrating syntax and lexical semantics. In J. Pustejovsky (Ed.), *Semantics and the Lexicon*, 129–161. London: Kluwer.