Statistical Methods and Linguistics

Steven Abney
University of Tübingen

In: Judith Klavans and Philip Resnik, eds., *The Balancing Act.* MIT Press, Cambridge, MA. 1996.

abney@sfs.nphil.uni-tuebingen.de
http://www.sfs.nphil.uni-tuebingen.de/~abney/
Wilhelmstr. 113, 72074 Tübingen, Germany

Statistical Methods and Linguistics

Steven Abney
University of Tübingen

In the space of the last ten years, statistical methods have gone from being virtually unknown in computational linguistics to being a fundamental given. In 1996, no one can profess to be a computational linguist without a passing knowledge of statistical methods. HMM's are as de rigeur as LR tables, and anyone who cannot at least use the terminology persuasively risks being mistaken for kitchen help at the ACL banquet.

More seriously, statistical techniques have brought significant advances in broad-coverage language processing. Statistical methods have made real progress possible on a number of issues that had previously stymied attempts to liberate systems from toy domains; issues that include disambiguation, error correction, and the induction of the sheer volume of information requisite for handling unrestricted text. And the sense of progress has generated a great deal of enthusiasm for statistical methods in computational linguistics.

However, this enthusiasm has not been catching in linguistics proper. It is always dangerous to generalize about linguists, but I think it is fair to say that most linguists are either unaware (and unconcerned) about trends in computational linguistics, or hostile to current developments. The gulf in basic assumptions is simply too wide, with the result that research on the other side can only seem naive, ill-conceived, and a complete waste of time and money.

In part the difference is a difference of goals. A large part of computational linguistics focuses on practical applications, and is little concerned with human language processing. Nonetheless, at least some computational linguists aim to advance our scientific understanding of the human language faculty by better understanding the computational properties of language. One of the most interesting and challenging questions about human language computation is just how people are able to deal so effortlessly with the very issues that make processing unrestricted text so difficult. Statistical methods provide the most promising current answers, and as a result the excitement about statistical methods is also shared by those in the cognitive reaches of computational linguistics.

In this paper, I would like to communicate some of that excitement to fellow linguists, or at least, perhaps, to make it comprehensible. There is no denying that there is a culture clash between theoretical and computational linguistics that serves to reinforce mutual prejudices. In charicature, computational linguists believe that by throwing more cycles and more raw text into their statistical black box, they can dispense with linguists altogether, along with their fanciful Rube Goldberg theories about exotic linguistic phenomena. The

I would like to thank Tilman Hoehle, Graham Katz, Marc Light, and Wolfgang Sternefeld for their comments on an earlier draft of this paper. All errors and outrageous opinions are of course my own.

linguist objects that, even if those black boxes make you oodles of money on speech recognizers and machine-translation programs (which they don't), they fail to advance our understanding. I will try to explain how statistical methods just might contribute to understanding of the sort that linguists are after.

This paper, then, is essentially an apology, in the old sense of *apology*. I wish to explain why we would do such a thing as to use statistical methods, and why they are not really such a bad thing, maybe not even for linguistics proper.

1 Language Acquisition, Language Variation, and Language Change

I think the most compelling, though least well-developed, arguments for statistical methods in linguistics come from the areas of language acquisition, language variation, and language change.

Language acquisition. Under standard assumptions about the grammar, we would expect the course of language development to be characterized by abrupt changes, each time the child learns or alters a rule or parameter of the grammar. If, as seems to be the case, changes in child grammar are actually reflected in changes in relative frequencies of structures that extend over months or more, it is hard to avoid the conclusion that the child has a probabilistic or weighted grammar in some form. The form that would perhaps be least offensive to mainstream sensibilities is a grammar in which the child "tries out" rules for a time. During the trial period, both the new and old versions of a rule co-exist, and the probability of using one or the other changes with time, until the probability of using the old rule finally drops to zero. At any given point, in this picture, a child's grammar is a stochastic (i.e., probabilistic) grammar.

An aspect of this little illustration that bears emphasizing is that the probabilities are added to a grammar of the usual sort. A large part of what is meant by "statistical methods" in computational linguistics is the study of stochastic grammars of this form: grammars obtained by adding probabilities in a fairly transparent way to "algebraic" (i.e., non-probabilistic) grammars. Stochastic grammars of this sort do not constitute a rejection of the underlying algebraic grammars, but a supplementation. This is quite different from some uses to which statistical models (most prominently, neural networks) are put, in which attempts are made to model some approximation of linguistic behavior with an undifferentiated network, with the result that it is difficult or impossible to relate the network's behavior to a linguistic understanding of the sort embodied in an algebraic grammar. (It should, however, be pointed out that the problem with such applications does not lie with neural nets, but with the unenlightening way they are put to use.)

Language change. Similar comments apply, on a larger scale, to language change. If the units of change are as algebraic grammars lead us to expect—rules or parameters or the like—we would expect abrupt changes. We might expect some poor bloke to go down to the local pub one evening, order "Ale!", and be served an eel instead, because the Great Vowel Shift happened to him a day too early.[1] In fact, linguistic changes that are attributed to rule changes or changes of parameter settings take place gradually, over considerable stretches of time, measured in decades or centuries. It is more realistic to assume that the language of a speech community is a stochastic composite of the languages of the individual speakers, described by a stochastic grammar. In the stochastic "community" grammar, the probability of a given construction reflects the relative proportion of speakers who use the construction in question. Language change consists in shifts in relative frequency of constructions (rules, parameter settings, etc.) in the community. If we think of speech communities as populations of grammars that vary within certain bounds, and if we think of language change as involving gradual shifts in the center of balance of the grammar population, then statistical models are of immediate applicability [25].

In this picture, we might still continue to assume that an adult monolingual speaker possesses a particular algebraic grammar, and that stochastic grammars are only relevant for the description of communities of varying grammars. However, we must at least make allowance for the fact that individuals routinely comprehend the language of their community, with all its variance. This rather suggests that at least the grammar used in language comprehension is stochastic. I return to this issue below.

Language variation. There are two senses of language variation I have in mind here: dialectology, on the one hand, and typology, on the other. It is clear that some languages consist of a collection of dialects that blend smoothly one into the other, to the point that the dialects are more or less arbitrary points in a continuum. For example, Tait describes Inuit as "a fairly unbroken chain of dialects, with mutual intelligibility limited to proximity of contact, the furthest extremes of the continuum being unintelligible to each other" [26, p.3]. To describe the distribution of Latin American native languages, Kaufman defines a *language complex* as "a geographically continuous zone that contains linguistic diversity greater than that found wthin a single language . . ., but where internal linguistic boundaries similar to those that separate clearly discrete languages are lacking" [14, p.31]. The continuousness of changes with geographic distance is consistent with the picture of a speech community with grammatical variance, as sketched above. With geographic distance, the mix of frequency of usage of various constructions changes, and a stochastic grammar of some sort is an appropriate model [15].

Similar comments apply in the area of typology, with a twist. Many of the universals of language that have been identified are statistical rather than abso-

[1] I have read this anecdote somewhere before, but have been unable to find the citation. My apologies to the unknown author.

lute, including rough statements about the probability distribution of language features ("head-initial and head-final languages are about equally frequent") or conditional probability distributions ("postpositions in verb-initial languages are more common than prepositions in verb-final languages") [11, 12]. There is as yet no model of how this probability distribution comes about, that is, how it arises from the statistical properties of language change. Which aspects of the distribution are stable, and which would be different if we took a sample of the world's languages 10,000 years ago or 10,000 years hence? There is now a vast body of mathematical work on stochastic processes and the dynamics of complex systems (which includes, but is not exhausted by, work on neural nets), much of which is of immediate relevance to these questions.

In short, it is plausible to think of all of these issues—language acquisition, language change, and language variation—in terms of populations of grammars, whether those populations consist of grammars of different speakers or sets of hypotheses a language learner entertains. When we examine populations of grammars varying within bounds, it is natural to expect statistical models to provide useful tools.

2 Adult Monolingual Speakers

But what about an adult monolingual speaker? Ever since Chomsky, linguistics has been firmly committed to the idealization to an adult monolingual speaker in a homogeneous speech community. Do statistical models have anything to say about language under that idealization?

In a narrow sense, I think the answer is probably not. Statistical methods bear mostly on all the issues that are outside the scope of interest of current mainstream linguistics. In a broader sense, though, I think that says more about the narrowness of the current scope of interest than about the linguistic importance of statistical methods. Statistical methods are of great linguistic interest because the issues they bear on are linguistic issues, and essential to an understanding of what human language is and what makes it tick. We must not forget that the idealizations that Chomsky made were an expedient, a way of managing the vastness of our ignorance. One aspect of language is its algebraic properties, but that is only one aspect of language, and certainly not the only important aspect. Also important are the statistical properties of language communities. And stochastic models are also essential for understanding language production and comprehension, particularly in the presence of variation and noise. (I will focus here on comprehension, though considerations of language production have also provided an important impetus for statistical methods in computational linguistics [22, 23].)

To a significant degree, I think linguistics has lost sight of its original goal, and turned Chomsky's expedient into an end in itself. Current theoretical syntax gives a systematic account of a very narrow class of data, judgments about the well-formedness of sentences for which the intended structure is specified, where the judgments are adjusted to eliminate gradations of goodness and other complications. Linguistic data other than structure judgments are classified as "performance" data, and the adjustments that are performed on structure-judgment data are deemed to be corrections for "performance effects". Performance is considered the domain of psychologists, or at least, not of concern to linguistics.

The term *performance* suggests that the things that the standard theory abstracts away from or ignores are a natural class; they are data that bear on language processing but not language structure. But in fact a good deal that is labelled "performance" is not computational in any essential way. It is more accurate to consider performance to be negatively defined: it is whatever the grammar does not account for. It includes genuinely computational issues, but a good deal more that is not. One issue I would like to discuss in some detail is the issue of grammaticality and ambiguity judgments about sentences as opposed to structures. These judgments are no more or less computational than judgments about structures, but it is difficult to give a good account of them with grammars of the usual sort; they seem to call for stochastic, or at least weighted, grammars.

2.1 Grammaticality and ambiguity

Consider the following:

(1) a. the a are of I

b. the cows are grazing in the meadow

c. John saw Mary

The question is the status of these examples with respect to grammaticality and ambiguity. The judgments here, I think, are crystal clear: (1a) is word salad, and (1b) and (c) are unambiguous sentences.

In point of fact, (1a) is a grammatical noun phrase, and (1b) and (c) are at least two ways ambiguous, the non-obvious reading being as a noun phrase. Consider: an *are* is a measure of area, as in *a hectare is a hundred ares*, and letters of the alphabet may be used as nouns in English (*"Written on the sheet was a single lowercase a," "As described in section 2 paragraph b ..."*). Thus (1a) has a structure in which *are* and *I* are head nouns, and *a* is a modifier of *are*. This analysis even becomes perfectly natural in the following scenario. Imagine we are surveyors, and that we have mapped out a piece of land into large segments, designated with capital letters, and subdivided into one-are sub-segments, designated with lower-case letters. Then *the a are of I* is a perfectly natural description for a particular parcel on our map.

As for (1b), *are* is again the head noun, *cows* is a premodifier, and *grazing in the meadow* is a postmodifier. It might be objected that plural nouns cannot be nominal premodifiers, but in fact they often are: consider *the bonds market, a securities exchange, he is vice president and media director, an in-home health care services provider, Hartford's claims division, the financial-services industry, its line of systems management software.* (Several of these examples are extracted from the Wall Street Journal.)

It may seem that examples (1a) and (b) are illustrative only of a trivial and artificial problem that arises because of a rare usage of a common word. But the problem is not trivial: without an account of 'rare usage', we have no way of distinguishing between genuine ambiguities and these spurious ambiguities. Alternatively, one might object that if one does not know that *are* has a reading as a noun, then *are* is actually unambiguous in one's idiolect, and (1a) is genuinely ungrammatical. But in that case the question becomes why *a hectare is a hundred ares* is not judged equally ungrammatical by speakers of the idiolect in question.

Further, (1c) illustrates that the rare usage is not an essential feature of examples (a) and (b). *Saw* has a reading as a noun, which may be less frequent than the verb reading, but is hardly a rare usage. Proper nouns can modify (*Gatling gun*) and be modified by (*Typhoid Mary*) common nouns. Hence, *John saw Mary* has a reading as a noun phrase, referring to the Mary who is associated with a kind of saw called a John saw.

It may be objected that constructions like *Gatling gun* and *Typhoid Mary* belong to the lexicon, not the grammar, but however that may be, they are completely productive. I may not know what *Cohen equations, the Russia house,* or *Abney sentences* are, but if not, then the denotata of *Cohen's equations, the Russian house,* or *those sentences of Abney's* are surely equally unfamiliar.[2] Likewise I may not know who *pegleg Pete* refers to, or *riverboat Sally*, but that does not make the constructions any less grammatical or productive.

The problem is epidemic, and it snowballs as sentences grow longer. One often hears in computational linguistics about completely unremarkable sentences with hundreds of parses, and that is in fact no exaggeration. Nor is it merely a consequence of having a poor grammar. If one examines the undesired analyses, one generally finds that they are extremely implausible, and often do considerable violence to 'soft' constraints like heaviness constraints or the number and sequence of modifiers, but no one piece of the structure is outright ungrammatical.

To illustrate, consider this sentence, drawn more or less at random from a book (Quine's *Word and Object*) drawn more or less at random from my shelf:

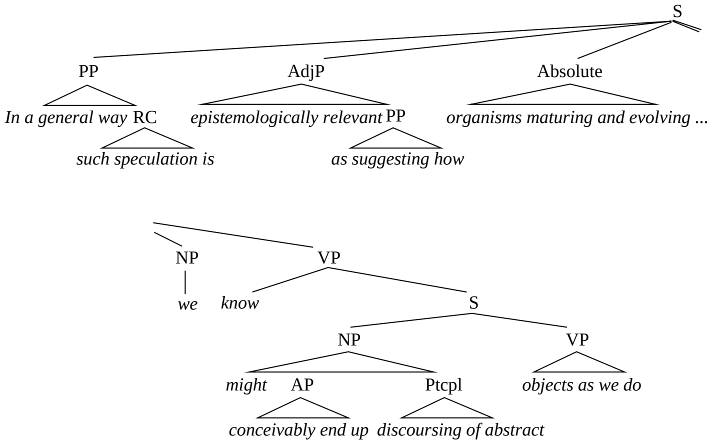
(2) In a general way such speculation is epistemologically relevant, as suggesting how organisms maturing and evolving in the physical environ-

[2] There are also syntactic grounds for doubt about the assumption that noun-noun modification belongs to the lexicon. Namely, adjectives can intervene between the modifying noun and the head noun. (Examples are given later in this section.) If adjective modification belongs to the syntax, and if there are no discontinuous words or movement of pieces of lexical items, then at least some modification of nouns by nouns must take place in the syntax.

> ment we know might conceivably end up discoursing of abstract objects as we do. [28, p. 123]

One of the many spurious structures this sentence might receive is the following:

(3)



There are any number of criticisms one can direct at this structure, but I believe none of them are fatal. It might be objected that the PP-AdjP-Absolute sequence of sentential premodifiers is illegitimate, but each is individually fine, and there is no hard limit on stacking them. One can even come up with relatively good examples with all three modifiers, e.g.: [$_{\text{PP}}$ *on the beach*] [$_{\text{AdjP}}$ *naked as jaybirds*] [$_{\text{Absolute}}$ *waves lapping against the shore*] *the wild boys carried out their bizarre rituals.*

Another point of potential criticism is the question of licensing the elided sentence after *how*. In fact its content could either be provided from preceding context or from the rest of the sentence, as in *though as yet unable to explain how, astronomers now know that stars develop from specks of grit in giant oysters.*

Might is taken here as a noun, as in *might and right.* The AP *conceivably end up* may be a bit mysterious: *end up* is here an adjectival, as in *we turned the box end up. Abstract* is unusual as a mass noun, but can in fact be used as one, as for example in *the article consisted of three pages of abstract and only two pages of actual text.*

One might object that the NP headed by *might* is bad because of the multiple postmodifiers, but in fact there is no absolute constraint against stacking nominal postmodifiers, and good examples can be constructed with the same structure: *marlinespikes, business end up, sprinkled with tabasco sauce, can be a powerful deterrent against pigeons.* Even the commas are not absolutely required. The strength of preference for them depends on how heavy the modifiers are: cf. *strength judicially applied increases the effectiveness of diplomacy, a cup* *of peanuts unshelled in the stock adds character.*[3]

In short, the structure (3) seems to be best characterized as grammatical, though it violates any number of parsing preferences and is completely absurd.

One might think that one could eliminate ambiguities by turning some of the dispreferences into absolute constraints. But attempting to eliminate unwanted readings that way is like squeezing a balloon: every dispreference that is turned into an absolute constraint to eliminate undesired structures has the unfortunate side effect of eliminating the desired structure for some other sentence. No matter how difficult it is to think up a plausible example that violates the constraint, some writer has probably already thought one up by accident, and we will improperly reject his sentence as ungrammatical if we turn the dispreference into an absolute constraint. To illustrate: if a noun is premodified by both an adjective and another noun, standard grammars require the adjective to come first, inasmuch as the noun adjoins to N^0 but the adjective adjoins to $\overline{\mathrm{N}}$. It is not easy to think up good examples that violate this constraint. Perhaps the reader would care to try before reading the examples in the footnote.[4]

Not only is my absurd analysis (3) arguably grammatical, there are many, many equally absurd analyses to be found. For example, *general* could be a noun (the army officer) instead of an adjective, or *evolving in* could be analyzed as a particle verb, or *the physical* could be a noun phrase (a physical exam)—not to mention various attachment ambiguities for coordination and modifiers, giving a multiplicative effect. The consequence is considerable ambiguity for a sentence that is perceived to be completely unambiguous.

Now perhaps it seems I am being perverse, and I suppose I am. But it is a perversity that is implicit in grammatical descriptions of the usual sort, and it emerges unavoidably as soon as we systematically examine the structures that the grammar assigns to sentences. Either the grammar assigns too many structures to sentences like (2), or it incorrectly predicts that examples like *three pages of abstract* or *a cup of peanuts unshelled in the stock* have no well-formed structure.

To sum up, there is a problem with grammars of the usual sort: their predictions about grammaticality and ambiguity are simply not in accord with human perceptions. The problem of how to identify the correct structure from among the in-principle possible structures provides one of the central motivations for the use of weighted grammars in computational linguistics. A weight is assigned to each aspect of structure permitted by the grammar, and the weight of a particular analysis is the combined weight of the structural features that make it up. The analysis with the greatest weight is predicted to be the perceived analysis for a given sentence.

[3] Cf. this passage from Tolkien: "Their clothes were mended as well as their bruises their tempers and their hopes. Their bags were filled with food and provisions light to carry but strong to bring them over the mountain passes." [27, p.61]

[4] *Maunder climatic cycles, ice-core climatalogical records, a Kleene-star transitive closure, Precambrian era solar activity, highland igneous formations.*

Before describing in more detail how weighted grammars contribute to a solution to the problem, though, let me address an even more urgent issue: is this even a linguistic problem?

2.2 Is this linguistics?

Under usual assumptions, the fact that the grammar predicts grammaticality and ambiguity where none is perceived is not a linguistic problem. The usual opinion is that perception is a matter of performance, and that grammaticality alone does not predict performance; we must also include non-linguistic factors like plausibility and parsing preferences and maybe even probabilities.

Grammaticality and acceptability. The implication is that *perceptions* of grammaticality and ambiguity are not linguistic data, but performance data. This stance is a bit odd—aren't grammaticality judgments perceptions? And what do we mean by "performance data"? It would be one thing if we were talking about data that clearly has to do with the course of linguistic computation, data like response times and reading times, or regressive eye movement frequencies, or even more outlandish things like PET scans or ERP traces. But human perceptions (judgments, intuitions) about grammaticality and ambiguity are classic linguistic data. What makes the judgments concerning examples (1a-c) performance data? All linguistic data is the result of little informal psycholinguistic experiments that linguists perform on themselves, and the experimental materials are questions of the form "Can you say this?" "Does this mean this?" "Is this ambiguous?" "Are these synonymous?"

Part of the answer is that the judgments about examples (1a-c) are judgments about sentences alone rather than about sentences with specified structures. The usual sort of linguistic judgment is a judgment about the goodness of a particular structure, and example sentences are only significant as bearers of the structure in question. If any choice of words and any choice of context can be found that makes for a good sentence, the structure is deemed to be good. The basic data are judgments about structured sentences in context—that is, sentences plus a specification of the intended structure and intended context—but this basic data is used only grouped in sets of structured contextualized sentences having the same (possibly partial) structure. Such a set is defined to be good just in case any structured contextualized sentence it contains is judged to be good. Hence a great deal of linguists' time is spent in trying to find some choice of words and some context to get a clear positive judgment, in order to show that a structure of interest is good.

As a result, there is actually no intent that the grammar *predict*—that is, generate—individual structured sentence judgments. For a given structured sentence, the grammar only predicts whether there is some sentence with the same structure that is judged to be good.

For the examples (1), then, we should say that the structure

$[_{\text{NP}}$ the $[_{\text{N}}$ a$]$ $[_{\text{N}}$ are$]$ $[_{\text{PP}}$ of $[_{\text{N}}$ I$]]]$

is indeed grammatical in the technical sense, since it is acceptable in at least one context, and since every piece of the structure is attested in acceptable sentences.

The grouping of data by structure is not the only way that standard grammars fail to predict acceptability and ambiguity judgments. Judgments are rather smoothly graded, but goodness according to the grammar is all or nothing. Discrepancies between grammar and data are ignored if they involve sentences containing center embedding, parsing preference violations, garden path effects, or in general if their badness can be ascribed to "processing complexity".[5]

Grammar and computation. The difference between structure judgments and string judgments is not that the former is "competence data" in some sense and the latter is "performance data". Rather, the distinction rests on a working assumption about how the data are to be explained, namely, that the data is a result of the interaction of grammatical constraints with computational constraints. Certain aspects of the data are assumed to be reflections of grammatical constraints, and everything else is ascribed to failures of the processor to translate grammatical constraints transparently into behavior, whether because of memory limits or heuristic parsing strategies or whatever obscure mechanisms create gradedness of judgments. We are justified in ignoring those aspects of the data that we ascribe to the idiosyncracies of the processor.

But this distinction does not hold up under scrutiny. Dividing the human language capacity into grammar and processor is only a manner of speaking, a way of dividing things up for theoretical convenience. It is naive to expect the logical grammar/processor division to correspond to any meaningful physiological division—say, two physically separate neuronal assemblies, one functioning as a store of grammar rules and the other as an active device that accesses the grammar-rule store in the course of its operation. And even if we *did* believe in a physiological division between grammar and processor, we have no evidence at all to support that belief; it is not a distinction with any empirical content.

A couple of examples might clarify why I say that the grammar/processor distinction is only for theoretical convenience. Grammars and syntactic structures are used to describe computer languages as well as human languages, but typical compilers do not access grammar-rules or construct parse-trees. At

[5] In addition, there are properties of grammaticality judgments of a different sort that are not being modelled, properties that are poorly understood and somewhat worrisome. Disagreements arise not infrequently among judges—it is more often the case than not that I disagree with at least some of the judgments reported in syntax papers, and I think my experience is not unusual. Judgments seem to change with changing theoretical assumptions: a sentence that sounds "not too good" when one expects it to be bad may sound "not too bad" if a change in the grammar changes one's expectations. And judgments change with exposure. Some constructions that sound terrible on a first exposure improve considerably with time.

the level of description of the operation of the compiler, grammar-rules and parse-trees exist only "virtually" as abstract descriptions of the course of the computation being performed. What is separately characterized as, say, grammar versus parsing strategy at the logical level is completely intermingled at the level of compiler operation.

At the other extreme, the constraints that probably have the strongest computational flavor are the parsing strategies that are considered to underly garden-path effects. But it is equally possible to characterize parsing preferences in grammatical terms. For example, the low attachment strategy can be characterized by assigning a cost to structures of the form $[_{X^{i+1}}\ X^i\ Y\ Z]$ proportional to the depth of the subtree Y. The optimal structure is the one with the least cost. Nothing depends on how trees are actually computed: the characterization is only in terms of the shapes of trees.

If we wish to make a distinction between competence and computation, an appropriate distinction is between *what* is computed and *how* it is computed. By this measure, most "performance" issues are not computational issues at all. Characterizing the perceptions of grammaticality and ambiguity described in the previous section does not necessarily involve any assumptions about the computations done during sentence perception. It only involves characterizing the set of structures that are perceived as belonging to a given sentence. That can be done, for example, by defining a weighted grammar that assigns costs to trees, and specifying a constant C such that only structures whose cost is within distance C of the best structure are predicted to be perceived. How the set thus defined is actually computed during perception is left completely open.

We may think of competence versus performance in terms of knowledge versus computation, but that is merely a manner of speaking. What is really at issue is an idealization of linguistic data for the sake of simplicity.

The frictionless plane, autonomy and isolation. Appeal is often made to an analogy between competence and frictionless planes in mechanics. Syntacticians focus on the data that they believe to contain the fewest complicating factors, and "clean up" the data to remove what they believe to be remaining complications that obscure simple, general principles of language.

That is proper and laudable, but it is important not to lose sight of the original problem, and not to mistake complexity for irrelevancy. The test of whether the simple principles we think we have found actually have explanatory power is how well they fare in making sense of the larger picture. There is always the danger that the simple principles we arrive at are artifacts of our data selection and data adjustment. For example, it is sometimes remarked how marvelous it is that a biological system like language should be so discrete and clean, but in fact there is abundant gradedness and variability in the original data; the evidence for the discreteness and cleanness of language seems to be mostly evidence we ourselves have planted.

It has long been emphasized that syntax is autonomous. The doctrine is older than Chomsky; for example, Tesnière writes "...la syntaxe. Il est **autonome**" (emphasis in the original). To illustrate that structure cannot be equated with meaning, he presents the sentence pair:

> le signal vert indique le voie libre
> le symbole veritable impose le vitesse lissant

The similarity to Chomsky's later but more famous pair

> revolutionary new ideas appear infrequently
> colorless green ideas sleep furiously

is striking.

But autonomy is not the same as isolation. Syntax is autonomous in the sense that it cannot be reduced to semantics; well-formedness is not identical to meaningfulness. But syntax in the sense of an algebraic grammar is only one piece in an account of language, and it stands or falls on how well it fits into the larger picture.

The holy grail. The larger picture, and the ultimate goal of linguistics, is to describe language in the sense of that which is produced in language production, comprehended in language comprehension, acquired in language acquisition, and, in aggregate, that which varies in language variation and changes in language change.

I have always taken the holy grail of generative linguistics to be to characterize a class of models, each of which represents a particular (potential or actual) human language L, and characterizes a speaker of L by defining the class of sentences a speaker of L produces, the structures that a speaker of L perceives for sentences; in short, by predicting the linguistic data that characterizes a speaker of L.

A "Turing test" for a generative model would be something like the following. If we use the model to generate sentences at random, the sentences that are produced are judged by humans to be clearly sentences of the language—to "sound natural". And in the other direction, if humans judge a sentence (or non-sentence) to have a particular structure, the model should also assign precisely that structure to the sentence.

Natural languages are such that these tests cannot be passed by an unweighted grammar. An unweighted grammar distinguishes only between grammatical and ungrammatical structures, and that is not enough. "Sounding natural" is a matter of degree. What we must mean by "randomly generating natural-sounding sentences" is that sentences are weighted by the degree to which they sound natural, and we sample sentences with a probability that accords with their weight. Moreover, the structure that people assign to a sentence is the structure they judge to have been intended by the speaker, and that judgment is also a matter of degree. It is not enough for the grammar to define the set of structures that could possibly belong to the sentence; the grammar should predict which structures humans actually perceive, and what the relative weights are in cases where humans are uncertain about which structure the speaker intended.

The long and little of it is, weighted grammars (and other species of statistical methods) characterize language in such a way as to make sense of language production, comprehension, acquisition, variation, and change. These are linguistic, and not computational issues, a fact that is obscured by labelling everything "performance" that is not accounted for by algebraic grammars. What is really at stake with "competence" is a provisional simplifying assumption, or an expression of interest in certain subproblems of linguistics. There is certainly no indicting an expression of interest, but it is important not to lose sight of the larger picture.

3 How Statistics Helps

Accepting that there are divergences between theory and data—for example, the divergence between predicted and perceived ambiguity—and accepting that this is a linguistic problem, and that it is symptomatic of the incompleteness of standard grammars, how does adding weights or probabilities help make up the difference?

Disambiguation. As already mentioned, the problem of identifying the correct parse—the parse that humans perceive—among the possible parses is a central application of stochastic grammars in computational linguistics. The problem of defining which analysis is correct is not a computational problem, however; the computational problem is describing an algorithm to compute the correct parse. There are a variety of approaches to the problem of defining the correct parse. A stochastic context-free grammar provides a simple illustration. Consider the sentence *John walks,* and the grammar

(4)	1.	S → NP V	.7
	2.	S → NP	.3
	3.	NP → N	.8
	4.	NP → N N	.2
	5.	N → John	.6
	6.	N → walks	.4
	7.	V → walks	1.0

According to grammar (4), *John walks* has two analyses, one as a sentence and one as a noun phrase. (The rule S → NP represents an utterance consisting of a single noun phrase.) The numbers in the rightmost column represent the weights of rules. The weight of an analysis is the product of the weights of the rules used in its derivation. In the sentential analysis of *John walks,* the derivation consists of rules 1, 3, 5, 7, so the weight is (.7)(.8)(.6)(1.0) = .336. In the noun-phrase analysis, the rules 2, 4, 5, 6 are used, so the weight is (.3)(.2)(.6)(.4) = .0144. The weight for the sentential analysis is much greater, predicting that it is the one perceived. More refined predictions can be obtained by hypothesizing that an utterance is perceived as ambiguous if the next-best case is not too much worse than the best. If "not too much worse" is interpreted as a ratio of, say, not more than 2:1, we predict that *John walks* is perceived as unambiguous, as the ratio between the weights of the parses is 23:1.[6]

Degrees of grammaticality. Gradations of acceptability are not accommodated in algebraic grammars: a structure is either grammatical or not. The idea of degrees of grammaticality has been entertained from time to time, and some classes of ungrammatical structures are informally considered to be "worse" than others (most notably, ECP violations versus subjacency violations). But such degrees of grammaticality as have been considered have not been accorded a formal place in the theory. Empirically, acceptability judgments vary widely across sentences with a given structure, depending on lexical choices and other factors. Factors that cannot be reduced to a binary grammaticality distinction are either poorly modelled or ignored in standard syntactic accounts.

Degrees of grammaticality arise as uncertainty in answering the question "Can you say X?" or perhaps more accurately, "If you said X, would you feel you had made an error?" As such, they reflect degrees of error in speech production. The null hypothesis is that the same measure of goodness is used in both speech production and speech comprehension, though it is actually an open question. At any rate, the measure of goodness that is important for speech comprehension is not degree of grammaticality alone, but a global measure that combines degrees of grammaticality with at least naturalness and structural preference (i.e., "parsing strategies").

We must also distinguish degrees of grammaticality, and indeed, global goodness, from the probability of producing a sentence. Measures of goodness and probability are mathematically similar enhancements to algebraic grammars, but goodness alone does not determine probability. For example, for an infinite language, probability must ultimately decrease with length, though arbitrarily long sentences may be perfectly good.

Perhaps one reason that degrees of grammaticality have not found a place in standard theory is the question of where the numbers come from, if we permit continuous degrees of grammaticality. The answer to where the numbers come from is *parameter estimation.* Parameter estimation is well-understood for a

[6] The hypothesis that only the best structure (or possibly, structures) are perceptible is somewhat similar to current approaches to syntax in which grammaticality is defined as optimal satisfaction of constraints or maximal economy of derivation. But I will not hazard a guess here about whether that similarity is significant or mere happenstance.

number of models of interest, and can be seen psychologically as part of what goes on during language acquisition.

Naturalness. It is a bit difficult to say precisely what I mean by naturalness. A large component is plausibility, but not plausibility in the sense of world knowledge, but rather plausiblity in the sense of selectional preferences, that is, semantic sortal preferences that predicates place on their arguments.

Another important component of naturalness is not semantic, though, but simply "how you say it". This is what has been called collocational knowledge, like the fact that one says *strong tea* and *powerful car*, but not vice versa [23], or that you say *thick accent* in English, but *starker Akzent* ("strong accent") in German.

Though it is difficult to define just what naturalness is, it is not difficult to recognize it. If one generates text at random from an explicit grammar plus lexicon, the shortcomings of the grammar are immediately obvious in the unnatural—even if not ungrammatical—sentences that are produced. It is also clear that naturalness is not at all the same thing as meaningfulness. For example, I think it is clear that *differential structure* is more natural than *differential child*, even though I could not say what a differential structure might be. Or consider the following examples, that were in fact generated at random from a grammar:

(5) a. matter-like, complete, alleged strips
a stratigraphic, dubious scattering
a far alternative shallow model

b. indirect photographic-drill sources
earlier stratigraphically precise minimums
Europe's cyclic existence

All these examples are about on a par as concerns meaningfulness, but I think the (b) examples are rather more natural than the (a) examples.

Collocations and selectional restrictions have been two important areas of application of statistical methods in computational linguistics. Questions of interest have been both how to include them in a global measure of goodness, and how to induce them distributionally [19], both as a tool for investigations, and as a model of human learning.

Structural preferences. Structural preferences, or parsing strategies, have already been mentioned. A "longest-match" preference is one example. The example

(6) the emergency crews hate most is domestic violence

is a garden-path because of a strong preference for the longest initial NP, *the emergency crews*, rather than the correct alternative, *the emergency*. (The correct interpretation is: *the emergency [that crews hate most] is domestic violence.*) The longest-match preference plays an important role in the dispreference for the structure (3) that we examined earlier.

As already mentioned, these preferences can be seen as structural preferences, rather than parsing preferences. They interact with the other factors we have been examining in a global measure of goodness. For example, in (6), an even longer match, *the emergency crews hate*, is actually possible, but it violates the dispreference for having plural nouns as nominal modifiers.

Error tolerance. A remarkable property of human language comprehension is its error tolerance. Many sentences that an algebraic grammar would simply classify as ungrammatical are actually perceived to have a particular structure. A simple example is *we sleeps*, a sentence whose intended structure is obvious, albeit ungrammatical. In fact, an erroneous structure may actually be preferred to a grammatical analysis; consider

(7) Thanks for all you help.

which I believe is preferentially interpreted as an erroneous version of *Thanks for all your help.* However, there is a perfectly grammatical analysis: *thanks for all those who you help.*

We can make sense of this phenomenon by supposing that a range of error-correction operations are available, though their application imposes a certain cost. This cost is combined with the other factors we have discussed, to determine a global goodness, and the best analysis is chosen. In (7), the cost of error correction is apparently less than the cost of the alternative in unnaturalness or structural dispreference. Generally, error detection and correction are a major selling point for statistical methods. They were primary motivations for Shannon's noisy channel model [21], which provides the foundation for many computational linguistic techniques.

Learning on the fly. Not only is the language that one is exposed to full of errors, it is produced by others whose grammars and lexica vary from one's own. Frequently, sentences that one encounters can only be analysed by adding new constructions or lexical entries. For example, when the average person hears *a hectare is a hundred ares*, they deduce that *are* is a noun, and succeed in parsing the sentence. But there are limits to learning on the fly, just as there are limits to error correction. Learning on the fly does not help one parse *the a are of I.*

Learning on the fly can be treated much like error correction. The simplest approach is to admit a space of learning operations—e.g., assigning a new part of speech to a word, adding a new subcategorization frame to a verb, etc.—and assign a cost to applications of the learning operations. In this way it is conceptually straightforward to include learning on the fly in a global optimization.

People are clearly capable of error correction and learning on the fly; they are highly desirable abilities given the noise and variance in the typical linguistic environment. They greatly exacerbate the problem of picking out the intended parse for a sentence, because they explode the candidate space even beyond the already large set of candidates that the grammar provides. To explain how it is nonetheless possible to identify the intended parse, there is no serious alternative to the use of weighted grammars.

Lexical acquisition. A final factor that exacerbates the problem of identifying the correct parse is the sheer richness of natural language grammars and lexica. A goal of earlier linguistic work, and one that is still a central goal of the linguistic work that goes on in computational linguistics, is to develop grammars that assign a reasonable syntactic structure to every sentence of English, or as nearly every sentence as possible. This is not a goal that is currently much in fashion in theoretical linguistics. Especially in GB, the development of large fragments has long since been abandoned in favor of the pursuit of deep principles of grammar.

The scope of the problem of identifying the correct parse cannot be appreciated by examining behavior on small fragments, however deeply analyzed. Large fragments are not just small fragments several times over—there is a qualitative change when one begins studying large fragments. As the range of constructions that the grammar accommodates increases, the number of undesired parses for sentences increases dramatically.

In-breadth studies also give a different perspective on the problem of language acquisition. When one attempts to give a systematic account of phrase structure, it becomes clear just how many little facts there are that do not fall out from grand principles, but just have to be learned. The simple, general principles in these cases are not principles of syntax, but principles of acquisition. Examples are the complex constraints on sequencing of prenominal elements in English, or the syntax of date expressions (*Monday June the 4th, Monday June 4, *Monday June the 4, *June 4 Monday*) or the syntax of proper names (*Greene County Sheriff's Deputy Jim Thurmond*), or the syntax of numeral expressions.

The largest piece of what must be learned is the lexicon. If parameter-setting views of syntax acquisition are correct, then learning the syntax (which in this case does not include the low-level messy bits discussed in the previous paragraph) is actually almost trivial. The really hard job is learning the lexicon.

Acquisition of the lexicon is a primary area of application for distributional and statistical approaches to acquisition. Methods have been developed for the acquisition of parts of speech [4, 20], terminological noun compounds [1], collocations [23], support verbs [10], subcategorization frames [2, 16], selectional restrictions [19], and low-level phrase structure rules [7, 24]. These distributional techniques do not so much compete with parameter setting as a model of acquisition, as much as complement it, by addressing issues that parameter-setting accounts pass over in silence. Distributional techniques are also not adequate alone as models of human acquisition—whatever the outcome of the syntactic versus semantic bootstrapping debate, children clearly do make use of situations and meaning to learn language—but the effectiveness of distributional techniques indicates at least that they might account for a component of human language learning.

4 Objections

There are a couple of general objections to statistical methods that may be lurking in the backs of readers minds, that I would like to address. First is the sentiment that, however relevant and effective statistical methods may be, they are no more than an engineer's approximation, not part of a proper scientific theory. Second is the nagging doubt: didn't Chomsky debunk all this ages ago?

4.1 Stochastic models are for engineers?

One might admit that one can account for parsing preferences by a probabilistic model, but insist that a probabilistic model is at best an approximation, suitable for engineering but not for science. On this view, we do not need to talk about degrees of grammaticality, or preferences, or degrees of plausibility. Granted, humans perceive only one of the many legal structures for a given sentence, but the perception is completely deterministic. We need only give a proper account of all the factors affecting the judgment.

Consider the example:

> Yesterday three shots were fired at Humberto Calvados, personal assistant to the famous tenor Enrique Felicidad, who was in Paris attending to unspecified personal matters.

Suppose for argument's sake that 60% of readers take the tenor to be in Paris, and 40% take the assistant to be in Paris. Or more to the point, suppose a particular informant, John Smith, chooses the low attachment 60% of the time when encountering sentences with precisely this structure (in the absence of an informative context), and low attachment 40% of the time. One could still insist that no probabilistic decision is being made, but rather that there are lexical and semantic differences that we have inappropriately conflated across sentences with 'precisely this structure', and if we take account of these other effects, we end up with a deterministic model after all. A probabilistic model is only a stopgap in absence of an account of the missing factors: semantics, pragmatics, what topics I've been talking to other people about lately, how tired I am, whether I ate breakfast this morning.

By this species of argument, stochastic models are practically always a stopgap approximation. Take stochastic queue theory, for example, by which one can give a probabilistic model of how many trucks will be arriving at given depots in a transportation system. One could argue that if we could just model everything about the state of the trucks and the conditions of the roads, the location of every nail that might cause a flat and every drunk driver that might cause an accident, then we could in principle predict deterministically how many trucks will be arriving at any depot at any time, and there is no need of stochastic queue theory. Stochastic queue theory is only an approximation in lieue of information that it is impractical to collect.

But this argument is flawed. If we have a complex deterministic system, and if we have access to the initial conditions in complete detail, so that we can compute the state of the system unerringly at every point in time, a simpler stochastic description may still be more insightful. To use a dirty word, some properties of the system are genuinely *emergent*, and a stochastic account is not just an approximation, it provides more insight than identifying every deterministic factor. Or to use a different dirty word, it is a *reductionist* error to reject a successful stochastic account and insist that only a more complex, lower-level, deterministic model advances scientific understanding.

4.2 Chomsky v. Shannon

In one's introductory linguistics course, one learns that Chomsky disabused the field once and for all of the notion that there was anything of interest to statistical models of language. But one usually comes away a little fuzzy on the question of what, precisely, he proved.

The arguments of Chomsky's that I know are from "Three Models for the Description of Language" [5] and *Syntactic Structures* [6] (essentially the same argument repeated in both places), and from the *Handbook of Mathematical Psychology*, chapter 13 [17]. I think the first argument in *Syntactic Structures* is the best known. It goes like this.

> Neither (a) 'colorless green ideas sleep furiously' nor (b) 'furiously sleep ideas green colorless', nor any of their parts, has ever occured in the past linguistic experience of an English speaker. But (a) is grammatical, while (b) is not.

This argument only goes through if we assume that if the frequency of a sentence or 'part' is zero in a training sample, its probability is zero. But in fact, there is quite a literature on how to estimate the probabilities of events that do not occur in the sample, and in particular how to distinguish real zeros from zeros that just reflect something that is missing by chance.

Chomsky also gives a more general argument:

> If we rank the sequences of a given length in order of statistical approximation to English, we will find both grammatical and ungrammatical sequences scattered throughout the list; there appears to be no particular relation between order of approximation and grammaticalness.

Because for any n, there are sentences with grammatical dependencies spanning more than n words, so that no nth-order statistical approximation can sort out the grammatical from the ungrammatical examples. In a word, you cannot define grammaticality in terms of probability.

It is clear from context that 'statistical approximation to English' is a reference to nth-order Markov models, as discussed by Shannon. Chomsky is saying that there is no way to choose n and ϵ such that

$$\text{for all sentences } s, \text{ grammatical}(s) \leftrightarrow P_n(s) > \epsilon$$

where $P_n(s)$ is the probability of s according to the 'best' nth-order approximation to English.

But Shannon himself was careful to call attention to precisely this point: that for any n, there will be some dependencies affecting the well-formedness of a sentence that an nth-order model does not capture. The point of Shannon's approximations is that, as n increases, the total mass of ungrammatical sentences that are erroneously assigned nonzero probability decreases. That is, we *can* in fact define grammaticality in terms of probability, as follows:

$$\text{grammatical}(s) \leftrightarrow \lim_{n\rightarrow\infty} P_n(s) > 0$$

A third variant of the argument appears in the *Handbook*. There Chomsky states that parameter estimation is impractical for an nth-order Markov model where n is large enough "to give a reasonable fit to ordinary usage". He emphasizes that the problem is not just an inconvenience for statisticians, but renders the model untenable as a model of human language acquisition: "we cannot seriously propose that a child learns the values of 10^9 parameters in a childhood lasting only 10^8 seconds."

This argument is also only partially valid. If it takes at least a second to estimate each parameter, and parameters are estimated sequentially, the argument is correct. But if parameters are estimated in parallel, say, by a high-dimensional iterative or gradient-pursuit method, all bets are off. Nonetheless, I think even the most hardcore statistical types are willing to admit that Markov models represent a brute force approach, and are not an adequate basis for psychological models of language processing.

However, the inadequacy of Markov models is not that they are statistical, but that they are statistical versions of finite-state automata! Each of Chomsky's arguments turns on the fact that Markov models are finite-state, not on the fact that they are stochastic. None of his criticisms are applicable to stochastic models generally. More sophisticated stochastic models do exist: stochastic context-free grammars are well understood, and stochastic versions of Tree-Adjoining Grammar [18], GB [8], and HPSG [3] have been proposed.

In fact, probabilities make Markov models more adequate than their non-probabilistic counterparts, not less adequate. Markov models are surprisingly effective, given their finite-state substrate. For example, they are the workhorse of speech recognition technology. Stochastic grammars can also be easier to learn than their non-stochastic counterparts. For example, though Gold [9] showed that the class of context-free grammars is not learnable, Horning [13] showed that the class of stochastic context-free grammars *is* learnable.

In short, Chomsky's arguments do not bear at all on the probabilistic nature of Markov models, only on the fact that they are finite-state. His arguments are not by any stretch of the imagination a sweeping condemnation of statistical methods.

5 Conclusion

In closing, let me repeat the main line of argument as concisely as I can. Statistical methods—by which I mean primarily weighted grammars and distributional induction methods—are clearly relevant to language acquisition, language change, language variation, language generation, and language comprehension. Understanding language in this broad sense is the ultimate goal of linguistics.

The issues to which weighted grammars apply, particularly as concerns perception of grammaticality and ambiguity, one may be tempted to dismiss as performance issues. However, the set of issues labelled "performance" are not essentially computational, as one is often led to believe. Rather, "competence" represents a provisional narrowing and simplification of data in order to understand the algebraic properties of language. "Performance" is a misleading term for "everything else". Algebraic methods are inadequate for understanding many important properties of human language, such as the measure of goodness that permits one to identify the correct parse out of a large candidate set in the face of considerable noise.

Many other properties of language, as well, that are mysterious given unweighted grammars, properties such as the gradualness of rule learning, the gradualness of language change, dialect continua, and statistical universals, make a great deal more sense if we assume weighted or stochastic grammars. There is a huge body of mathematical techniques that computational linguists have begun to tap, yielding tremendous progress on previously intransigent problems. The focus in computational linguistics has admittedly been on technology. But the same techniques promise progress at long last on questions about the nature of language that have been mysterious for so long. The time is ripe to apply them.

References

[1] Didier Bourigault. Surface grammatical analysis for the extraction of terminological noun phrases. In *COLING-92, Vol. III*, pages 977–981, 1992.

[2] Michael R. Brent. Automatic acquisition of subcategorization frames from untagged, free-text corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 209–214, 1991.

[3] Chris Brew. Stochastic HPSG. In *Proceedings of EACL-95*, 1995.

[4] Eric Brill. *Transformation-Based Learning.* PhD thesis, Univ. of Pennsylvania, 1993.

[5] Noam Chomsky. Three models for the description of language. *IRE Transactions on Information Theory*, IT-2(3):113–124, 1956. Institute of Radio Engineers, New York.

[6] Noam Chomsky. *Syntactic Structures.* Mouton, 1957.

[7] Steven Paul Finch. *Finding Structure in Language.* PhD thesis, University of Edinburgh, 1993.

[8] Andrew Fordham and Matthew Crocker. Parsing with principles and probabilities. In *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, 1994.

[9] E. Mark Gold. Language identification in the limit. *Information and Control*, 10(5):447–474, 1967.

[10] Gregory Grefenstette. Corpus-based method for automatic identification of support verbs for nominalizations. In *EACL-95*, 1995.

[11] John A. Hawkins. *Word Order Universals.* Academic Press, New York, 1983.

[12] John A. Hawkins. A parsing theory of word order universals. *Linguistic Inquiry*, 21(2):223–262, 1990.

[13] James Jay Horning. *A Study of Grammatical Inference.* PhD thesis, Stanford (Computer Science), 1969.

[14] Terrence Kaufman. The native languages of Latin America: general remarks. In Christopher Moseley and R.E. Asher, editors, *Atlas of the World's Languages*, pages 31–33. Routledge, London and New York, 1994.

[15] Brett Kessler. Computational dialectology in Irish Gaelic. In *EACL-95*, 1995.

[16] Christopher D. Manning. Automatic acquisition of a large subcategorization dictionary from corpora. In *31st Annual Meeting of the Association for Computational Linguistics*, pages 235–242, 1993.

[17] George A. Miller and Noam Chomsky. Finitary models of language users. In R.D. Luce, R. Bush, and E. Galanter, editors, *Handbook of Mathematical Psychology*, chapter 13. Wiley, New York, 1963.

[18] Philip Resnik. Probabilistic Tree-Adjoining Grammar as a framework for statistical natural language processing. In *COLING-92*, pages 418–424, 1992.

[19] Philip Resnik. *Selection and Information.* PhD thesis, University of Pennsylvania, Philadelphia, PA, 1993.

[20] Hinrich Schütze. Part-of-speech induction from scratch. In *31st Annual Meeting of the Association for Computational Linguistics*, pages 251–258, 1993.

[21] Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3–4):379–423, 623–656, 1948.

[22] Frank Smadja. Microcoding the lexicon for language generation. In Uri Zernik, editor, *Lexical Acquisition: Using on-line resources to build a lexicon.* MIT Press, 1989.

[23] Frank Smadja. *Extracting Collocations from Text. An Application: Language Generation.* PhD thesis, Columbia University, New York, NY, 1991.

[24] Tony C. Smith and Ian H. Witten. Language inference from function words. Manuscript, University of Calgary and University of Waikato, January 1993.

[25] Whitney Tabor. The gradualness of syntactic change: A corpus proximity model. Ms. for Berkeley Colloquium talk, CSLI, Stanford University, November 1993.

[26] Mary Tait. North America. In Christopher Moseley and R.E. Asher, editors, *Atlas of the World's Languages*, pages 3–30. Routledge, London and New York, 1994.

[27] J.R.R. Tolkien. *The Hobbit.* Houghton Mifflin Co., Boston, 1966.

[28] Willard van Orman Quine. *Word and Object.* The MIT Press, Cambridge, MA, 1960.