The Stochastic Generalization
(Part II)

JOAN BRESNAN
Stanford University

[based on Bresnan, Dingare, and Manning (2001)]

[Optimality Theory and Typology, Summer School 2002]

1

Givón:

"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."

— Givón (1979: 26–31)

2

I. Generalizing from Categorical to Frequentistic Phenomena (Review)

In a nutshell . . .

i) The generalization: The same categorical phenomena which are attributed to hard grammatical constraints in some languages continue to show up as statistical preferences in other languages, motivating a grammatical model that can account for soft constraints.

ii) A case study: The person hierarchy affects subject selection categorically in Lummi (Straits Salish, British Columbia), Picurís (Tanoan, New Mexico), and other languages. It also affects the frequency of subject selection in active/passive choices in English.

iii) A model: Stochastic optimality theory can account for the differences between Lummi (or Picurís) and English by positing different strengths for constraints within the same typologically motivated constraint system.

3

A Central Hypothesis

The same constraints are hypothesized to be present in all grammars, but are more or less active depending on their ranking relative to other constraints.

Lummi and (by hypothesis) Picurís fall back on *S_{newer} (or *$S_{nontopical}$, = Aissen's *S_t) with third person agent and patient:

	input: v(ag/3/new,pt/3)	*S_3	*S_{newer} (or *S_t)	*S_{pt}	*S_{ag}
	active: S_{ag},O_{pt}	*	*!		*
☞	passive: S_{pt},Obl_{ag}	*		*	

In English the person-avoidance constraints are overridden by discourse constraints:

	input: v(ag/3, pt/1)	*S_{newer} (or *S_t)	*S_{pt}	*S_{ag}	*S_3
☞	active: S_{ag},O_{pt}			*	*
	passive: S_{pt},Obl_{ag}		*!		

We know this because the disharmonic combinations are still grammatical in English, unlike Lummi and Picurís: *She met me*, *She'll be met by you*.

4

Why should person/role constraints be present in every grammar?

Two (broad) theories:

perspective-based: empathy or perspective-taking (Kuno and Kaburaki 1977; DeLancey 1981; Kuno 1987; MacWhinney in progress, ao) — grammar is designed to facilitate perspective shifting during communication; interlocutors share the perspectives of speech-act participants and of referents having causal roles.

pragmatics-based: accessibility of referents in the pragmatic context (Givón 1976, 1979, 1994; Ariel 1991; Warren and Gibson 2001; cf. Gordon et al. 2001) — nominal expressions are most easily processed when their referents are contextually accessible

The connection to voice: Speech-act participants, referents having causal roles, and contextually accessible referents all tend to receive more attention and are consequently more frequently the ***subjects of predication.***

How can we generalize from hard to soft constraints?

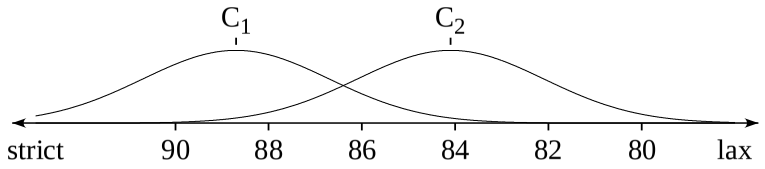
Stochastic OT[a] (Boersma 1998, 2000, Boersma and Hayes 2001) differs from standard OT in two essential ways:

(i) **ranking on a continuous scale:** Constraints are not simply ranked on a discrete ordinal scale; rather, they have a value on the continuous scale of real numbers. Thus constraints not only dominate other constraints, but are specific distances apart, and these distances are relevant to what the theory predicts.

(ii) **stochastic evaluation**: At each evaluation the real 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 the mean rank of 99 could be evaluated at 98.12 or 100.3. It is the constraint ranking that results from these new disharmonic values that is used in evaluation.

[a]—one of a family of new optimization-based theories of grammar that can provide a unified account of categorical, variable, and gradient data (see Anttila 2002, Manning to appear, and references).

Constraint ranking on a continuous scale with stochastic evaluation:[a]



An OT grammar with stochastic evaluation can generate both categorical and variable outputs.

Categorical outputs arise when crucially ranked constraints are distant. As the distance between constraints increases, interactions become vanishingly rare. (A distance of five standard deviations ensures an error rate of less than 0.02% (Boersma and Hayes 2001: 50).)[b]

Variable outputs arise when crucially ranked constraints are closer together.

[a]Note the numerical scale is reversed to show stricter constraints to left as in OT tableaux.

[b]Units of measurement are arbitrary. With standard deviation = 2.0, a ranking distance of 10 units between constraints is taken to be effectively categorical.

What is gained by the model?

Recall: Logical Entailment of Implicational Universals

The theory of harmonic alignment logically entails certain crosslinguistic generalizations, which follow from the constraint subhierarchies and the transitivity of constraint domination (≫) in ordinal ('vanilla') OT.

Comrie (1989: 128): *"…the most natural kind of transitive construction is one where the A is high in animacy and definiteness and the P is lower in animacy and definiteness; and any deviation from this pattern leads to a more marked construction."*

The spread of markedness:

Agent ↓ Patient →	Local person	Third person
Local person		
Third person		

Disregarding other constraints, if passivization is categorical for some input, then it must be categorical for any more marked input (Dingare 2001: 16–17). For example, in Lummi and Picurís, passive is obligatory for input from the lower left cell and optional for input from the lower right cell. *Prediction: In no languages does the reverse hold.*

Generalization: Predictions of Relative Frequency

Disregarding other constraints, if passivization occurs with some *frequency* for a given input, then (by Aissen's theory of harmonic alignment expressed within the Stochastic OT model) it must occur with equal or higher *frequency* for any more marked input (Dingare 2001: 18).

Agent ↓ Patient →	Local person	Third person
Local person		
Third person		

Evidence from English

(Bresnan, Dingare, and Manning's 2001 study of SWITCHBOARD)

Rate of Passivization

Agent ↓ Patient →	Local person	Third person
Local person	0.0%	0.0%
Third person	2.9%	1.2%

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.

Harmonic alignment gave us two particular hypotheses which are supported by these data: that the rate of passivization of 3 → 1,2 should be higher than for 3 → 3 (1-sided Fisher exact, $p < 0.008$); and that the rate of passivization of 1,2 → 3 should be lower than for 3 → 3 (1-sided Fisher exact, $p < 0.0001$).

Summary:

The same disharmonic person/argument associations which are avoided categorically in languages like Lummi and Picurís by making passives either impossible or obligatory, are avoided in the SWITCHBOARD *corpus of spoken English by either depressing or elevating the frequency of passives relative to actives.*

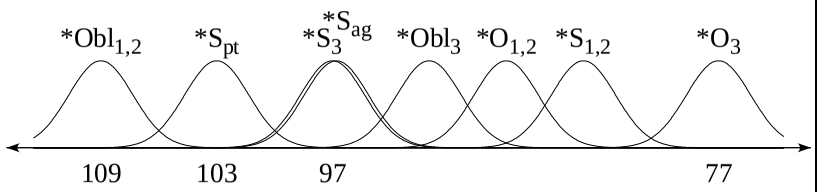
The generalization across categorical and frequentistic outputs can be captured in Stochastic Optimality Theory.

"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 tranformational-generative linguist will then be forced to count this fact as competence in Krio and performance in English."

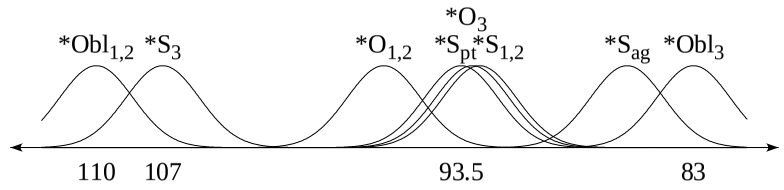
— Givón (1979: 26–31)

II. Stochastic Grammars

Partial stochastic grammar of English:



Partial stochastic grammar of Lummi:



13

Where do the real number ranking values in a stochastic grammar come from?

The input-output frequency distributions produced by two constraints whose evaluation values are each normally distributed can be exactly calculated as the differences of two Gaussians (normal distributions). But with many constraints acting on various inputs the calculations become very complicated. We therefore use computational simulations (Boersma's 1998 Gradual Learning Algorithm) to determine ranking values.

The n constraints define an n-dimensional space and each grammar can be located as a point in the space, according to its constraint ranking values $C_1, \ldots, C_n$. Because in general there are multiple grammars for each language, a language corresponds to a region in the space.

Grammars (and languages) are not evenly distributed in the constraint space. The theory embedded in the constraint set limits the space of possible grammars (for example, no grammars exist in areas which violate constraint subhierarchies). Our simulations serve to demonstrate the existence of grammars in the feasible space which do give the observed distributions.

14

The Gradual Learning Algorithm (GLA) is implemented in the Praat system (Boersma and Weenink 2000).

Starting from an initial state grammar in which all constraints have the same ranking values (arbitrarily set to be 100.0), the GLA is presented with learning data consisted of input-output pairs having the statistical distribution of (in the present case) a sample of spoken English.

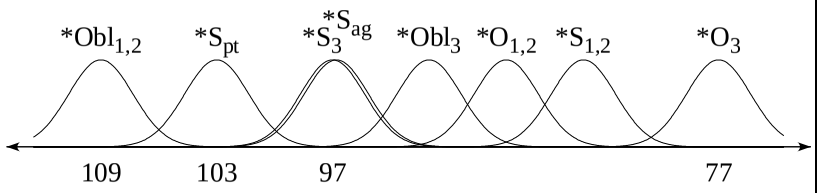
For each learning datum (a given input-output pair), the GLA compares the output of its own grammar for the same input; if its own output differs from the given output, it adjusts its grammar by moving all the constraints that differentially disfavor its own output upward on the continuous ranking scale by a small increment, and moving all constraints that differentially disfavor the given output downward along the scale by a small decrement. The adjustment process applies recursively to constraint subhierachies in order to preserve their local ordering relations.[a]

[a]The increment/decrement value is called the 'plasticity' and may be assumed to vary stochastically and to change with age (Boersma 2000).

15

English . . .

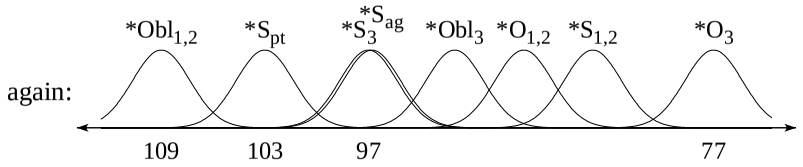
Partial stochastic grammar of English:



Output distribution of grammar:

input:			% Active:	% Passive:
1,2	→	1,2	100.00	0.00
1,2	→	3	100.00	0.00
3	→	3	98.80	1.20
3	→	1,2	97.21	2.79

16

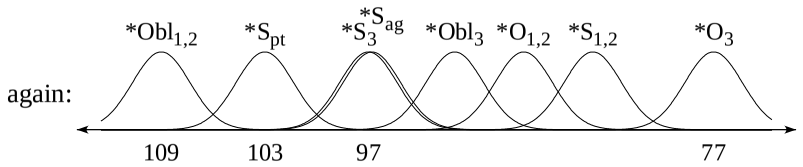


Observe: $*S_{pt} \gg *S_3$ but $|*S_{pt} - *S_3| = 6$, close enough to produce low frequency variable outputs for some inputs. For inputs where only the agent is third person, passive outputs will occasionally be favored by $*S_3$:

An (infrequent) effect of $*S_3$ on passive outputs:

	input: v(ag/3,pt/1)	$*S_3$	$*S_{pt}$	$*S_{ag}$
	active: S_{ag},O_{pt}	*!		*
☞	passive: S_{pt},Obl_{ag}		*	

When both agent and patient are third person, the $*S_3$ constraint cannot decide between active and passive, and the decision passes to other constraints.



Observe: $|*Obl_{1,2} - *O_{pers}| > 10$. ($*O_{1,2}$ disfavors an active for an input with local-person patient and $*O_3$ for an input with third-person patient.) These rankings reflect the zero frequency of local person passive agents in our data. But Kato (1979) cites (from Studs Terkel, *Working*):

> I said, "Me watch it! Fuck that! Let him watch it." He was hired by me. I could fire him if I didn't like him.
>
> When somebody says to me, "You're great, how come you're *just* a waitress?" *Just* a waitress. I'd say, "Why, don't you think you deserve to be served by me?"

Caveats:

- With more training data and a more complete constraint set which includes factors of topicality and focus, our model should learn grammars that produce passives with local person agents.[a]
- The set of constraints used in this system is motivated by broader typological considerations (Aissen 1999 and references). Some of these constraints play no necessary part in the system presented here, and a smaller constraint set is able to model the observed data equally well.
- This constraint set contains no information structure constraints which would motivate the use of passive independent of person. Because of this, the grammar models the 'background level' of passivization by keeping $*S_{ag}$ close enough to $*S_{pt}$ that one will occasionally get passives. This can be viewed as an artifact of our incomplete constraint set.

[a] If the ranking value of $*Obl_{1,2}$ in the grammar were lowered from 109 to 104, the output of local person passives would increase to one-tenth of one percent, 0.1%, while barely changing the frequency of other outputs.

In sum, stochastic OT can capture the soft influence of person on English passivization that exists beneath the level of grammaticality judgments. Disharmonic person/argument combinations are grammatical but avoided, affecting the frequency of passivization.

Lummi...

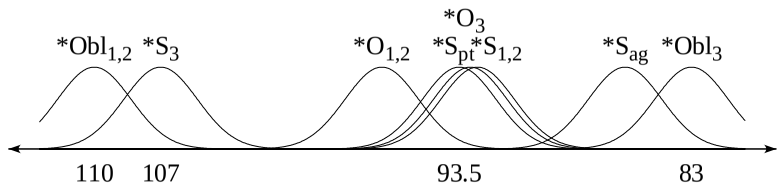
Unfortunately we lack a parsed SWITCHBOARD corpus for Lummi or Picurís. Nevertheless, it is possible to show by simulation how the descriptions of passive/voice interactions in Lummi or Picurís grammar can also be captured by a stochastic OT grammar. We interpret the descriptions of Lummi from Jelinek and Demers (1983, 1994) by means of a simple distribution. *Where a sentence type is described as ungrammatical, we assign it 0% relative frequency; where it is described as obligatory, we assign it 100%; and where it is described as optional, we assign it 50%:*

Simulated Lummi input/output distribution:

input:			% Active:	% Passive:
1,2	→	1,2	100.00	0.00
1,2	→	3	100.00	0.00
3	→	3	50.00	50.00
3	→	1,2	0.00	100.00

The simulated input/output distribution is then used to generate training data for the GLA, as before. The initial state of the grammar and the training regime are exactly the same as for English.

Partial stochastic grammar of Lummi



Observe: $|{*S_3} - {*S_{pt}}| > 10$. This ranking yields the obligatory passivization of inputs with local person patients and non-local person agents, capturing the categorical influence of person on Lummi passivization.[a]

The output distribution of the grammar matches the simulated learning distribution exactly.

[a] This analysis, deriving from our GLA simulations, differs from that of Aissen (1999), though the constraints are the same.

Isn't ranking on the continuous scale of real numbers powerful enough to learn any distribution?

No, it isn't. Under the present theory there are no stochastic OT grammars for 'anti-Lummi' or 'anti-English' distributions, which reverse the generalizations embodied in our data. Greater relative frequency of passives for first or second person acting on third would imply that third person subjects are avoided less than first or second person subjects. If so, then $*S_{1,2}$ must dominate $*S_3$ for a greater proportion of evaluations. But that ranking violates the constraint subhierarchy, which requires the *mean* ranking values of these constraints to occur in the reverse order.

Thus, the output of stochastic OT grammars are limited to subspaces of distributions that conform to the theory embodied in the constraint set – the sharing of the effect of constraint violations across inputs, and in particular, here, the constraint subhierarchies. Within that feasible space, they can match input frequencies. But they are not completely general-purpose statistical analyzers and they do not just memorize frequencies (Boersma 2000).

Why is English like Lummi and Picurís?

It is "a mainstay of functional linguistics" that "linguistic elements and patterns that are frequent in discourse become conventionalized in grammar" (from a publisher's blurb on Bybee and Hopper 2001). On this view, Lummi and Picurís are simply at an extreme point from English along the continuum of conventionalization that connects frequentistic preferences in usage to categorical grammatical constraints.

Conventionalization and Frequency

Stochastic OT grammars allow us to place the person/voice interactions in English and Lummi at points on a continuum of conventionalization that connects frequentistic preferences in usage to categorical grammatical constraints. If this general perspective is correct, then we would expect to find languages at intermediate points on this same continuum.

25

Consider Squamish:

3 → 2: passive obligatory in Lummi and Squamish
3 → 1: passive obligatory in Lummi, optional in Squamish

Analysis:[a]

Lummi:	Squamish:
*Obl$_{1,2}$ ≫ *S$_3$ ≫ *O$_2$,*O$_1$,*S$_{pt}$	*Obl$_{1,2}$ ≫ *O$_2$ ≫ *S$_3$,*O$_1$,*S$_{pt}$

[a]This analysis differs from that of Aissen (1999), reflecting our GLA simulations. Recall also that the mutual ranking of the local-person avoidance constraints is not fixed by the subhierarchy, but subject to crosslinguistic variation.

However, it is not fully informative to say, as has been customary (Jelinek and Demers 1983 ao), that passivization with third person agents and first person patients is "optional" in Squamish.

In terms of what is preferred rather than what is merely possible, Squamish is described as being much the same as Lummi, "except that third person acting on first may be active, though commonly passive" (Klokeid 1969: 11).[a]

26

Thus in Squamish as in English, passives of the type *I was fooled by her* are optional alternatives to actives with disharmonic local-person objects: *She fooled me*. But in spoken English, such passives are exceedingly infrequent, far less common than the corresponding actives, while in Squamish they are more frequent than the corresponding actives.

Why?

[a]We were unable to find quantitative measures of Squamish passives. Jacobs' (1994) corpus study of Squamish excludes first and second person because the purpose is to examine interactions of topic continuity with voice/inversion through measures of distance between pronouns and their textual antecedents.

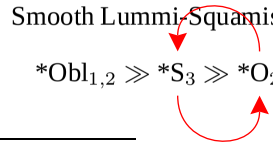
27

Lummi:	Squamish:
*Obl$_{1,2}$ ≫ *S$_3$ ≫ *O$_2$,*O$_1$,*S$_{pt}$	*Obl$_{1,2}$ ≫ *O$_2$ ≫ *S$_3$,*O$_1$,*S$_{pt}$

Squamish and Lummi are closely related Coast Salish languages. In the continuous constraint space of stochastic OT, the similarities of their grammars to each other and to the grammars of their common ancestors will appear as close distances between constraints.

In particular, different points in the changing categoricity of person effects on the passive will be reflected by gradual changes in frequency, as the relative distance between constraints shrinks and grows:[a]

Smooth Lummi-Squamish Reranking:



[a]The rankings of Aissen (1999) differ somewhat from those learned by the GLA, though the sets of possible outputs are equivalent.

28

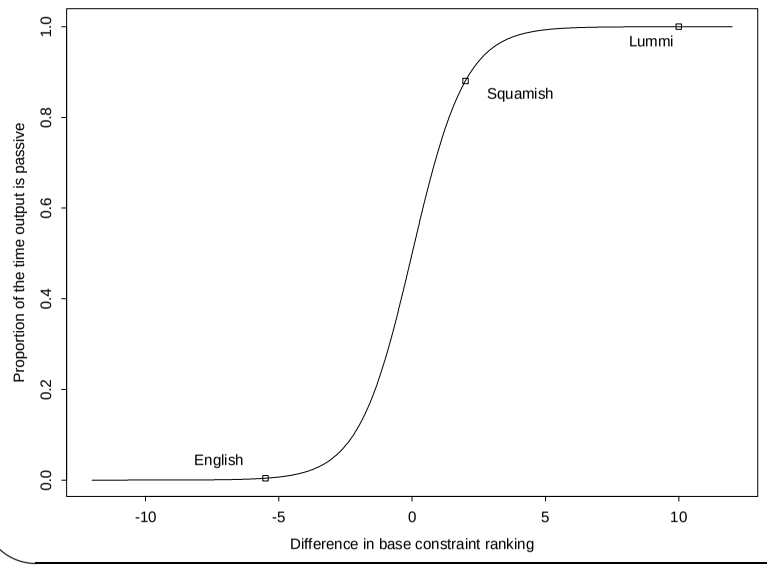
Reranking produces smooth changes in frequency—

If reranking is the movement in strength of a constraint along the continuous scale, as implied by the stochastic OT model, then (all else being equal) smooth changes in the relative frequencies of usage are predicted.

—but not linear changes:

If a constraint reranking is crucial to the choice between two outputs, and the distance between the two constraints is changing linearly, the prediction is that we should see an 'S' curve between the proportion of occurrences of the two outputs, of the sort that has been widely remarked on in historical and socio-linguistics (Weinreich, Labov, and Herzog 1968, Bailey 1973, Kroch 2001).

29



30

How can this gradual process work in a conventional generative grammar?

There, frequentistic processes (such as the conventionalization of usage preferences) must belong either to grammar-external 'performance' along with speech errors and memory limitations, or to external choices among competing dialect grammars. Yet neither of these alternatives is an adequate model of variation and change (Weinreich, Labov, and Herzog 1968).

31

Could competing conventional generative grammars explain the passive variation in English?

The competing grammars theory of variation is a model of diglossia (Kroch 2001). On the diglossic model of variation, the contact between two different populations having different grammars leads to internalization of competing grammars by individual speakers, who control two separate varieties. For example, some historical changes in English word order are attributed to the influence of Scandinavian speakers in Northern England (Kroch and Taylor 1997).

Could the diglossic model explain our passive findings? On this account, individual speakers would vary in the frequency of passive outputs because they have internalized alternative grammars which they deploy with varying frequency. The different grammars would have arisen from contact between different populations speaking varieties of English with and without the passive construction for certain person/role combinations. One population would have Lummi-like gaps in actives and passives as a hard constraint of their English grammar, perhaps as a result of some parameter setting of UG.

32

Some early studies propose that middle-class English speakers use an 'elaborated code' which has a higher proportion of passive verbs among all finite verbs than a 'restricted code' of working-class speakers, which has a lower percentage (Bernstein 1971 ao). But these studies have been criticized for failing to isolate the syntactic choice between active and passive, which shows no signficant difference between these groups (Weiner and Labov 1981: 32). (Passives should be compared to equivalent actives, rather than to all sentences. The latter can be influenced by differences in what is talked about, given that passives require fewer arguments than actives.)

Spontaneous speech shows significant stylistic and discourse effects on the choice of (agentless) passive or generalized-subject active.[a] But: "All of these conditions on the selection of active vs. passive are general features of the English language, used in much the same way by the very different sub-sections of the speech communities that we studied."(Weiner and Labov 1981: 56).

Conclusion: Diglossia is an unlikely model for our passive data. *"All sections of the population appear to treat the passive/active choice in the same way, and conversely, the same constraints are found throughout the speech community."* (Weiner and Labov 1981: 56)

[a]Generalized pronoun subjects ("they") are characteristic of colloquial English, while passives are a mark of formal scientific and literary discourse; passives are also favored by the discourse tendencies to preserve subject reference and structural parallelism.