

Generality, or How We Solved the “Splang” Problem

(also, the “Burnt” Problem)

1. Chomsky and Halle (1968, 331)

“The problem to which the linguist addresses himself [is] to account for the child’s construction of a grammar and to determine what preconditions on the form of language make it possible. Our approach to this problem is two-pronged. First we develop a system of formal devices for expressing rules and a set of general conditions on how these rules are organized and how they apply ... Secondly, we determine a procedure of evaluation that selects the highest valued of a set of hypotheses of the appropriate form, each of which meets a condition of compatibility with the primary linguistic data.

“A proposed theory that specifies formal devices and an evaluation procedure can be proven false (all too easily, in actual fact) by confronting it with empirical evidence relating to the grammar that actually underlies the speaker’s performance. There is such a grammar, and it is an empirical problem to discover it and to determine the basis for its acquisition. However difficult it may be to find relevant evidence for or against a proposed theory, there can be no doubt whatsoever about the empirical nature of the problem.”

2. One Way to Pursue This Goal

Develop an algorithm for learning phonological and morphological rules, and implement it on a machine.

3. A Difficulty

Many rules have exceptions, but are nevertheless productive.

- The exceptions aren’t labeled!
- Perhaps the real world is full of “learning traps”: data patterns that readily invite rules that fail to match what humans do.

4. Goals

- Show that there **are** such traps.
- Show that with the right learning algorithm, we needn’t fall into them.

5. Two Traps

- The *Burnt* problem

Process some data pairs: [dʒʌmp] ~ [dʒʌmpt]
[kɪk] ~ [kɪkt]
[mɪs] ~ [mɪst]
[læf] ~ [læft]

“OK! Voiceless consonants take /t/!”

[bɜːn] ~ [bɜːnt]

“Better still! Consonants take /t/!”

We will show this is not a hypothesis that can be immediately dispensed with...

- The *Splang* problem

to be explained...

SOME BACKGROUND ON OUR PHONOLOGY LEARNER

6. Entertain and Test Many Hypotheses

because you never know in advance what will pan out...

Pinker and Prince (1988, 131): “A symbolic model would solve the problem using a mechanism that can formulate, provisionally maintain, test, and selectively expunge hypotheses about rules of various degrees of generality.”

... “Candidates for rules are hypothesized by comparing base and past tense versions of a word, and factoring apart the changing portion, which serves as the rule operation, from certain morphologically-relevant phonological components of the stem, which serve to define the class of stems over which the operation can apply.”

Given: [sɪp] ~ [sɪpt]

“Add /-t/ to /Xɪp/”	(encouraged by: <i>tip, grip</i>)
“Add /-t/ to /Xp/”	(encouraged by: <i>tap, pop, hope</i>)
“Add /-t/ to /X[C,-vce]/”	(encouraged by: <i>kick, miss, laugh</i>)
“Add /-t/ to /XC/”	(encouraged by: <i>drag, rub</i> , etc. if you know some phonotactics)
“Add /-t/ to anything”	(well, not really: see later)

7. The Grammatical Shrine

Pinker and Prince (1988, 134): “[by various means], some regularities can be enshrined as permanent productive rules whereas others can be discarded or treated differently.”

8. Our Assessment of Some Subsequent Experimental Work

Discard with caution! Evidence indicates there are rules that:

- have the same structural change as the “enshrined” default rule
- apply to a subset of the cases to which the default applies
- nevertheless play a role in speaker judgments

9. Experimental Findings

- Subjects in experiments feel **more confident** in the output of a general rule when the input falls within an “**island of reliability**.”
- This seen in Wug-testing experiments by Prasada and Pinker (1993), Albright (1998).
- Example: English /Xɪp/ → /Xɪpt/ is an island (statistics of English lexicon), and people like [mɪpt] as the past tense of [mɪp] *better* than they like regular Wug pasts in general.
- More vivid effects (because more obvious islands) occur in Italian (Albright)
- See also Marcus et al. (1995) on German noun plurals, past participles.

10. Inference: Grammars are Rich

If people do Wug testing using some kind of symbolic grammar, then:

- That grammar has many overlapping rules (same structural change) that superficially appear redundant;
- The rules are annotated for how much the user can “trust” them;
- Such annotations reflect the statistics of the lexicon.

11. The Role of Generality in our Learner

- It seeks it, yet fears it...
- Given a data pair, with the same structural change, formulate the **tightest** hypothesis compatible with the pair:

[rɪŋ] ~ [rΛŋ]

[flɪŋ] ~ [flΛŋ]

$X [\text{liquid}] \text{ } \text{ɪ } \eta \rightarrow X [\text{liquid}] \wedge \eta$

- Given a highly heterogeneous data pair, the tightest hypothesis will be quite general:

[buba] [bubat] \searrow
 [mumit] [mumitt] \swarrow yield: $X \rightarrow X_t$

reason: /a/ and /t/ have nothing in common.

- More fully: compare data with existing generalized rules, to obtain more general rules.

12. Minimal Generality

Never generalize beyond the point needed to match input data.

13. Benefits

This tends to find all the islands of reliability.¹

But it also finds highly general rules that can handle completely novel forms.

Example: every rule in (6) above, except “Add /-t/ to anything”, is in fact discovered.

14. The Learner Evaluates Rules, Not Grammars (cf. SPE)

- Rules are evaluated on the basis of how well you can **trust** them (given what you know about the data) to generate correct outputs.
- Computing this: a first stab

$$\text{Raw Reliability} = \frac{\text{number of cases where rule applies successfully}}{\text{number of cases where structural description of rule is met}}$$

- When you generate new forms, the many rules will usually permit a number of different outputs. Favor those outputs generated by rules with high reliability.

15. Weighing the Testimony

- 1000/1000 is more credible than 5/5.
- We use a statistical adjustment proposed by Mikheev (1995), which uses confidence limits to take this into account. We call the result:

Confidence

- Example: 5/5 gets demoted from 1 to .825; 1000/1000 from 1 to .999

¹ It finds all of them, insofar as the schema for possible structural descriptions is complete. For contexts, our schema is:

[_{word} (Variable) (FeatureMatrix) (SegmentString) ____ (SegmentString) (FeatureMatrix) (Variable)]_{word}

which can get quite a lot.

- We find that Confidence yields a better match to Wug testing judgments than Raw Reliability.²

16. Confidence Doesn't Have To Be Near 1 to Yield a Reasonable Wug-Test Guess

Some forms from Prasada and Pinker (1993):

		Learner rating	"Production Probability" (P&P human results)
[kli:d]	➡ [kli:dəd]	.812	.83
	? [kləd]	.526	.42
[friŋ]	➡ [friŋd]	.806	.79
	? [frʌŋ]	.512	.33 (V changes reported together)
	?-?? [fræŋ]	.273	

17. ...But the Single Digits Tend to Correspond to Dreadful Outcomes

[kli:d]	* [kli:t]	.023	
	* [kli:d]	.019	
[smi:nθ]	[smi:nθt]	.806	(Note effect of phonotactically-deviant base)
	??-* [smənθ]	.059	
	* [smounθ]	.023	

18. A Tiny Bit of Phonology

- Given a set of ill-formed sequences, our system can in a rough way learn the rules that fix them.
- English: Given the knowledge that

*[bt], *[gt] *[pd], *[kd] *[tt], *[dd] (etc.)

are ill-formed, the system will find (crude versions of) Progressive Voicing Assimilation and Schwa Epenthesis, yielding:

[bd], [gd] [pt], [kt] [təd], [dəd]

² The particular version of Confidence we use is the .75 lower confidence limit for the value of Raw Reliability.

19. How to Give Phonology a Chance

- Minimal generalization can choke off phonology. E.g.

$X [-\text{voice}] \rightarrow X [-\text{voice}] t$

chokes off the possibility of letting /-d/ be the affix, with [-t] derived phonologically.

- Often, knowledge of phonology lets us find rules that work more generally than before, and are thus even more trustable.
- Also, trying out more morpheme combinations lets you explore the phonology more carefully.

20. Doppelgänger Rules

- Given $A \rightarrow B$, $A \rightarrow B'$, $A \rightarrow B''$, etc. as structural changes
- Given a rule $[(\text{context})] A \rightarrow [(\text{context})] B$
- Then also add:
 $[(\text{context})] A \rightarrow [(\text{context})] B'$
 $[(\text{context})] A \rightarrow [(\text{context})] B''$
etc.
- This gives rules a chance to improve their generality and confidence by invoking phonology.

THE “BURNT” PROBLEM

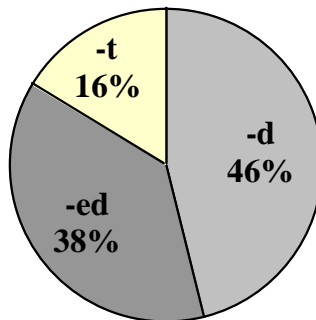
21. Background: Some Rough Statistics of the English Verb Lexicon

Taken from a representative sample of 2182 verbs ³

<u>Final segment</u>	<u>Regular past allomorph</u>	<u>Number</u>
Vowel	[d]	287
Sonorant consonants:	[d]	476
/d/:	[əd]	121
Other voiced obstruents:	[d]	296
/t/:	[əd]	459
Other voiceless obstruents:	[t]	380

³ From Web page of Brian MacWhinney, augmented to include all irregulars.

22. Relative Distribution of Regular Past Tense Allomorphs in English



23. The Course of Minimal Generalization, for /-t/

Input datum	Rules spawned
a. [dʒʌmp] ~ [dʒʌmpt]	dʒʌmp → dʒʌmp+t
b. [si:p] ~ [si:pt]	si:p → si:p+t
	Xp → Xp+t
c. [kɪk] ~ [kɪkt]	kɪk → kɪk+t
	X[-syl, -son, -cont, -cor, -voice]
	→ X[-syl, -son, -cont, -cor, -voice]+t
d. [mɪs] ~ [mɪst]	mɪs → mɪst
	X[-syl, -son, -voice] → X[-syl, -son, -voice]+t

24. Some Interesting Irregulars Found in Many English Dialects

burnt, learnt, spelt, smelt, dwelt, spilt

hit any one of these, and you spawn:

$X[-syl] \rightarrow X[-syl]+t$

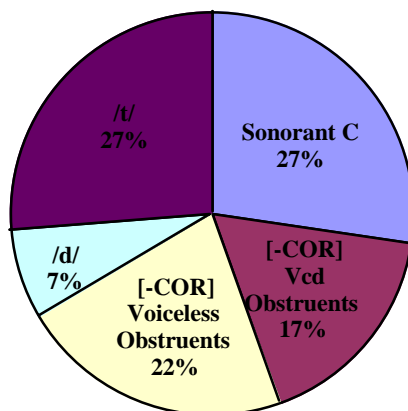
= “Attach /-t/ to any consonant-final stem.”

25. Calculating the Confidence of “X[-syl] → X[-syl]+t”

It works for:

- All regular stems ending in /p,tʃ,k,f,θ,s,ʃ/ (380).
- With rightward assimilation of [+voice], triggered by voiced obstruents, all regular stems ending in /b,dʒ, g, v, ð, z, ʒ/ (296)

- With phonology /Tt/ → /Təd/, even the stems ending in /t and d/ (580).



Result: 1256/2182 = Confidence .568 if you allow dicey phonology
 676/2182 = Confidence .303 if you don't

Either way: this is well into the zone where people give reasonably accepting Wug-test judgments (see (16) above).

26. This is Empirically Disastrous

[flæn] ~ *[flænt], [kaɪl] ~ *[kaɪlt], [spɹam] ~ *[spɹamt]

This is the “Burnt” Problem.

27. Note on Alternative Approaches

A learning procedure that values generality above all else is doomed to generate *[flænt]. Its mapping is the most general one.

28. When Will The “Burnt” Problem Arise?

whenever a few exceptional forms take the “wrong” allomorph

- French “h aspiré” words
- The ~50 roots of Hungarian that take the “wrong” suffix harmony: *híd-nak* ‘to the bridge’ vs. regular *víz-nek* ‘to the water’ (Vago 1976)
- The small set of similar cases in Turkish, such as *saat-i* ‘watch-acc.’, Clements and Sezer 1983, 242.

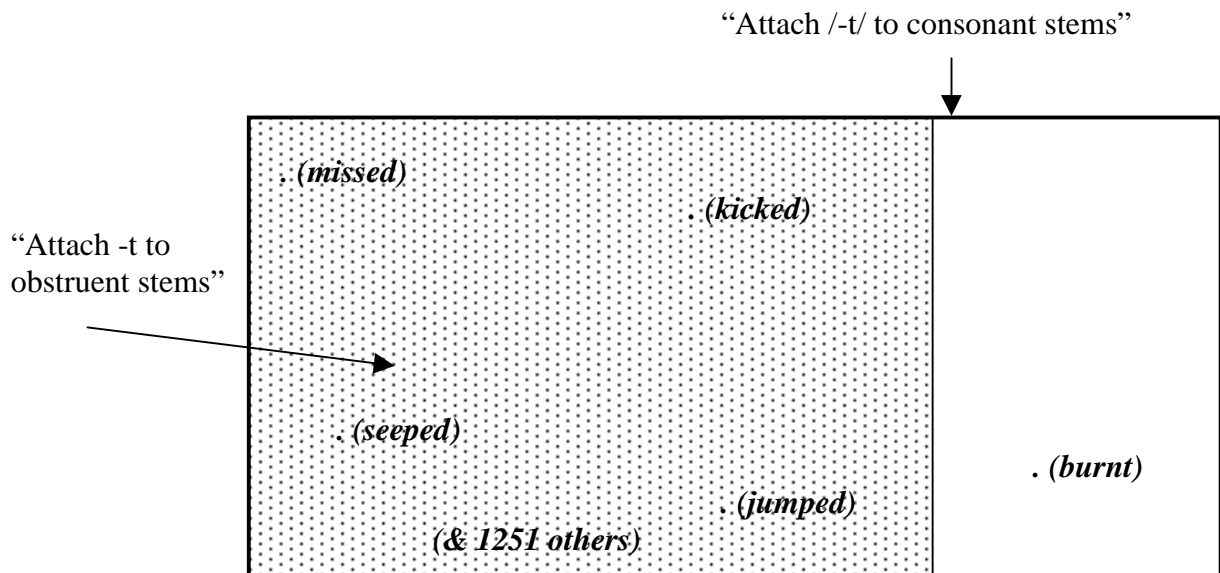
OUR DIAGNOSIS OF THE ‘BURNT’ PROBLEM

29. The False Generalization is Internally Heterogeneous

For now: ignore all irregulars other than *burnt* itself.

	Scope	Hits
“Attach /-t/ to consonant stems”	1895	1256
“Attach /-t/ to obstruent stems”	1255	1255

- In comparison to “Attach /-t/ to obstruent stems”, “Attach /-t/ to consonant stems” has added a large number to the Scope—and a pathetic single case to the Hits.
- To the extent that “Attach /-t/ to consonant stems” looks good, it does so because it has a major subset that is doing almost all the work.



30. An Informal Criterion for Rule Goodness

- If we are to take a rule seriously, it must offer us a substantial improvement over the performance of its best subset.

31. An Algorithm for Enforcing This Criterion (first version)

Apply: after some learning has already happened; or later (we do this after all rules discovered).

For each rule R:

- Consider every other rule R' in the grammar that has the same structural change.
- Does R' apply to a subset of the cases of R (= is its structural description a superset of R's?). Then:

- Find out how well R performs in the cases that aren't also covered by R'. We denote this set of cases as R_{residue}.

$$\text{Raw reliability}(R_{\text{residue}}) = \frac{\text{hits}(R) - \text{hits}(R')}{\text{scope}(R) - \text{scope}(R')}$$

- From Raw reliability(R_{residue}), calculate Confidence(R_{residue})
- If Confidence(R_{residue}) is considerably worse than Confidence(R), then we know that much of the virtue that we attribute to R would be better attributed to R'. Therefore, **impugn R**:

In Wug-testing with R, use Confidence(R_{residue}) rather than Confidence(R).

- End loop: comparisons of R with all possible R'
- End loop: consider all R.

A "BURNT" SIMULATION

32. What We Looked At

- 454 regular English verbs, in a fairly representative selection.
- Final segment: proportions matched (21) above.
- Also, 22 irregulars like *sing* ~ *sang*, to be examined later.
- Also: *burnt*

33. Wug-Testing the Grammar with an Ordinary Verb: [lɔp]

Guess	Rule Used	Confidence
[lɔpt]	X $\begin{bmatrix} -\text{son} \\ -\text{dors} \end{bmatrix} \rightarrow \text{X} \begin{bmatrix} -\text{son} \\ -\text{dors} \end{bmatrix} \text{t}$.985 not impugned

[lɔ̃pəd]	$X \begin{bmatrix} -\text{son} \\ -\text{cont} \\ -\text{voice} \\ -\text{dors} \end{bmatrix} \rightarrow X \begin{bmatrix} +\text{cor} \\ -\text{cont} \\ -\text{vce} \\ -\text{dors} \end{bmatrix} \text{əd}$ = p,t	.842 impugned to: .062
----------	--	-------------------------------------

[lɔ̃pd] Not generated at all: ruled out by phonology.

34. Wug-Testing [læn]: Guesses Made by the Grammar

(as always: each guess is assumed to be supported by the *best* rule that generates it)

Guess	Rule Used	Confidence
[lænd]	$X \begin{bmatrix} V \\ -\text{high} \end{bmatrix} n \rightarrow X \begin{bmatrix} V \\ -\text{high} \end{bmatrix} nd$.959 not impugned
[lænt]	$X \begin{bmatrix} +\text{cor} \\ -\text{cont} \\ +\text{ant} \end{bmatrix} n \rightarrow X \begin{bmatrix} +\text{cor} \\ -\text{cont} \\ +\text{ant} \end{bmatrix} nt$ = t,d,n	.834 impugned to: .060
[lænəd]	$X \begin{bmatrix} +\text{cor} \\ -\text{cont} \\ +\text{ant} \end{bmatrix} n \rightarrow X \begin{bmatrix} +\text{cor} \\ -\text{cont} \\ +\text{ant} \end{bmatrix} nəd$ = t,d,n	.750 impugned to: .025

35. Followup: A Wug Form that is Very Close to *burnt*

[mɔ̃n]

Guess	Rule Used	Confidence
[mɔ̃nd]	$X \begin{bmatrix} -\text{cont} \\ -\text{dors} \end{bmatrix} \rightarrow X \begin{bmatrix} -\text{cont} \\ -\text{dors} \end{bmatrix} d$.972 impugned to .926

$$\begin{array}{lcl}
[məːnt] & X \begin{bmatrix} +cor \\ -ant \\ -cons \end{bmatrix} n \rightarrow X \begin{bmatrix} +cor \\ -ant \\ -cons \end{bmatrix} nt & .126 \\
& = r, ə & \textbf{not impugned} \\
[məːnəd] & X \begin{bmatrix} +cor \\ +ant \\ -cont \end{bmatrix} \rightarrow X \begin{bmatrix} +cor \\ +ant \\ -cont \end{bmatrix} əd & .750 \\
& & \textbf{impugned to} \\
& & .025
\end{array}$$

36. Local Conclusion

It looks like impugning is doing the right thing in the right places.

CASES WHERE THERE ARE VERY FEW FORMS

37. We Now Have a Take on “Spling ~ Splang”

- Minimal generalization has a major virtue: it lets you find all the detailed generalizations you need, rather than settling for something extremely general and quitting prematurely.⁴
- But earlier, we found that there are ways in which it is *too* conservative: there aren’t always enough data to spawn a needed constraint.
- Specific example: from [splɪŋ] we want [splæŋ].

38. The Forms One Has to Go On (7 total)

[r ɪ ŋ] ~ [ræŋ]
 [dr ɪ ŋk] ~ [dræŋk]
 [ʃr ɪ ŋk] ~ [ʃræŋk] (dial.)
 [spr ɪ ŋ] ~ [spræŋ]
 [s ɪ ŋ] ~ [sæŋ]
 [s ɪ ŋk] ~ [sæŋk]
 [sw ɪ m] ~ [swæm]

- Note: in all, the /ɪ/ preceded by /s/ or /r/ or /w/.

39. The Most General Rule for (38) Licensed by Minimal Generalization

$$X \begin{bmatrix} +cont \\ -lat \end{bmatrix} \begin{bmatrix} +nas \\ -cor \end{bmatrix} I \rightarrow X \begin{bmatrix} +cont \\ -lat \end{bmatrix} æ \begin{bmatrix} +nas \\ -cor \end{bmatrix} Y$$

⁴ This was a problem with an earlier version of our learner; and with greedy top-down algorithms in general.

-nas

-nas

- This excludes /l/ from its left environment.
- Confidence of this rule: .439 (7 hits/13 scope)

40. The Two Tasks at Hand

- Generate a hypothesis that will produce *splang*—the rule given in (39) will not.
- Find a way to take this rule seriously.

41. The (Best) Rule of our Grammar That Generates *Splang*

$$X \begin{bmatrix} +\text{cont} \\ -\text{nas} \end{bmatrix} \text{ I } \begin{bmatrix} +\text{nas} \\ -\text{cor} \end{bmatrix} Y \rightarrow X \begin{bmatrix} +\text{cont} \\ -\text{nas} \end{bmatrix} \text{ æ } \begin{bmatrix} +\text{nas} \\ -\text{cor} \end{bmatrix} Y$$

confidence = .334

42. How We Generated Rule (41)

- Through cloning of structural changes; see (20) above.
- Constraints based on *flung/slung/clung* yield Doppelgängers that would generate *flang/slang/clang*—and also *splang*.
- This was not the purpose of introducing Doppelgängers, but it worked here.
- We are pondering alternatives for expanding the hypothesis space...

43. Why It is Hard to Take Rule (41) Seriously

- All the examples covered by (41) are already covered by the more specific (39).
- How do we keep (39) from impugning (41) into oblivion?
- N.B. this is exactly what we *want* to happen with *burnt*.

44. How Do You Know When You Should Venture Into New Phonological Territory?

burnt: Don't do it! (even if some forms tempt you to)

splang: Do it! (even though there are no forms that tempt you to)

45. Expanding the Role of Timidity in the System

- Before: be timid in generalizing beyond what the input data give you; that is, in asserting what forms are possible. Mechanism: use Confidence (lower confidence limit) instead of Raw Reliability.
- Thus, generalizations must be supported by copious data to be asserted confidently.
- Now: be timid in asserting what forms are **not** possible.
- Thus predicted *non*-occurrences must also be supported by copious data.

46. Implementation: Make Impugnment Timid

- The crucial computed value in impugnment, so far, is:

$$\text{Raw reliability}(R_{\text{residue}}) = \frac{\text{hits}(R) - \text{hits}(R')}{\text{scope}(R) - \text{scope}(R')}$$

- But when the data are few, we cannot be as sure that the reliability of the residue is really so poor (testimony of the few).
- Thus for impugnment, we must compute

$$\text{Confidence}(R_{\text{residue}})$$

as the **upper** statistical confidence limit for Raw reliability(R_{residue}), in order to be timid in impugning.

47. Confidence in General

- Use lower confidence limits in asserting existence.
- Use upper confidence limits in asserting non-existence.

48. How This Worked for *Splang*

- Impugning constraint: the specific one (39) (no [I]'s)
- (Potentially) impugned constraint: the general one (41) (with [I]'s allowed)
- Procedure:

Subtract the scope:	$17 - 13 = 4$
Subtract the hits:	$7 - 7 = 0$
Raw Reliability _{residue}	0
(upper) Confidence _{residue}	.453
(lower) Confidence of rule (39)	.438
- Rule (41) is not impugned.

49. Full Results for *Splang*

Guess	Rule Used	Confidence
[splɪŋd]	$X \begin{bmatrix} +\text{son} \\ -\text{lat} \end{bmatrix} \rightarrow X \begin{bmatrix} +\text{son} \\ -\text{lat} \end{bmatrix} d$.926 impugned to: .776
[splʌŋ]	$X \{d, l\} \cap Y \rightarrow X \{d, l\} \cap Y$.825 (5/5) not impugned

[splæŋ]	$X \begin{bmatrix} +\text{cont} \\ -\text{dors} \\ -\text{nas} \end{bmatrix} \text{ } \text{I Y} \rightarrow X \begin{bmatrix} +\text{cont} \\ -\text{dors} \\ -\text{nas} \end{bmatrix} \text{ } \text{æ Y}$.334 not impugned
[splɪŋt]	$X \begin{bmatrix} -\text{syl} \\ +\text{dors} \end{bmatrix} \rightarrow X \begin{bmatrix} -\text{syl} \\ +\text{dors} \end{bmatrix} \text{ } \text{t}$.451 impugned to: .072
[splɪŋəd]	$X \begin{bmatrix} -\text{lab} \\ +\text{vce} \\ -\text{cont} \end{bmatrix} \rightarrow X \begin{bmatrix} -\text{lab} \\ +\text{vce} \\ -\text{cont} \end{bmatrix} \text{ } \text{t}$.312 impugned to: .017

CONCLUSIONS

50. The Hierarchy of Generality

- All potential phonological generalizations exist within a hierarchy of generality.
- In phonological learning, one must explore this hierarchy to find the generalizations that best match the data.
- Explicit comparison between rules is needed to find the best rules—AND to discover which rules are essentially fraudulent.
- Simple, “one rule at a time” approaches (such as pure Minimal Generalization, or top-down greedy algorithms) are unworkable, because they can be tricked into going too far.

51. In Praise of Timidity

- Timidity (in the form of Minimal Generalization) gives you the ability to find enough constraints (Islands of Reliability; () above).
- Timidity (in the form: “don’t trust a rule until you’ve compared it with its logical neighbors) solves the Burnt Problem.
- Timidity (in the form: “don’t impugn a rule unless you’ve got solid evidence to back you up”) permits forms like *splang* to be projected, in the absence of data that would support them directly.

References

- Albright, Adam (1998) "Phonology subregularities in productive and nonproductive inflectional classes: evidence from Italian," M.A. thesis, Dept. of Linguistics, UCLA.
- Chomsky, Noam and Morris Halle (1968) *The Sound Pattern of English*, Harper and Row, New York.
- Clements, George N. and Engin Sezer (1982) "Vowel and consonant disharmony in Turkish," in Harry van der Hulst and Norval Smith, eds., *The Structure of Phonological Representations, Part II*, Foris, Dordrecht.
- Marcus, Gary, et al. (1995) German Inflection: the Exception that Proves the Rules. *Cognitive Psychology* 29, 186-256.
- Mikheev, Andrei (1995) "Automatic rule induction for unknown-word guessing," *Computational Linguistics* 23, 405-423.
- Pinker, Steven and Alan S. Prince (1988) "On language and connectionism: analysis of a parallel distributed processing model of language acquisition," *Cognition* 28, 73-193.
- Prasada, Sandeep and Steven Pinker (1993) "Generalization of regular and irregular morphological patterns," *Language and Cognitive Processes* 8, 1-56.
- Vago, Robert (1976) "Theoretical implications of Hungarian vowel harmony," *Linguistic Inquiry* 7, 243-263.