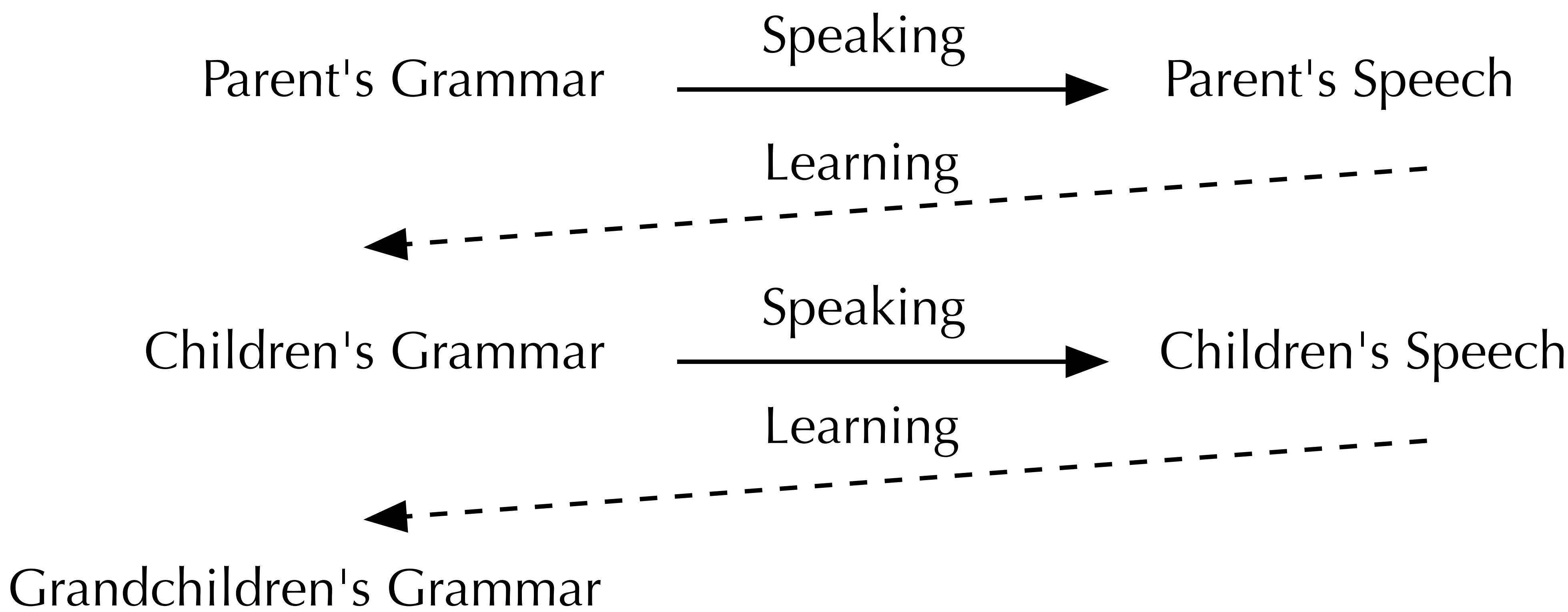
Toward a Quantitative Model of Morphological Change

Charles Yang
Dept. of Linguistics, Psychology, and Computer Science
Yale University

Penn Linguistics

2-23-2006

Language Learning and Change

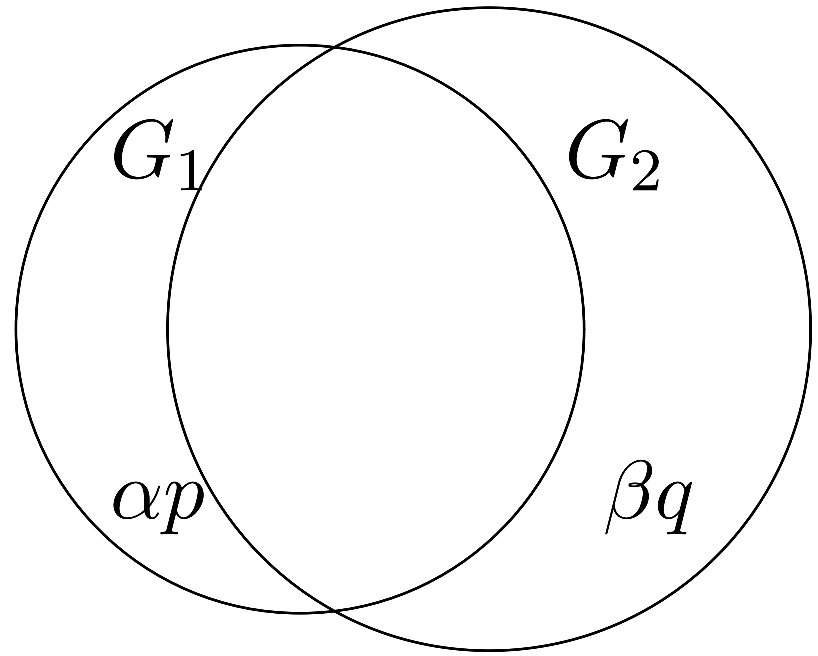


- parallel to the transmission of genetic materials under forces of evolution
  - how can we extrapolate from the synchronic study of language learning and processing to the diachronic study of language change?
- "the same mechanisms which operated to produce the large-scale changes of the past may be observed operating in the current changes taking place around us" (Labov 1973, p161)

Grammar Competition

- Yang (1998, 2002): Parameter setting as probability matching (Bush & Mosteller 1951, Herrstein 1961)
  - see also Roeper (2000), Crain & Pietroski (2002), Rizzi (2005)
- Evidence for competing parameters in child language: more on Saturday's PLC (w. J. Legate)
- Evidence for competing parameters in historical change (Kroch 1989 and much subsequent work)

Syntactic Change



- mutual incompatibility between two grammars can be quantified via corpus statistics: α and β are the **fitness** measures in grammar competition
- A model of syntactic change follows, which may predict the direction of parameter change
  - application to the loss of V2 in Old French (and potentially other Western Romance languages) and more tentatively, that in Middle English (Yang 2000, *Language Variation and Change*)

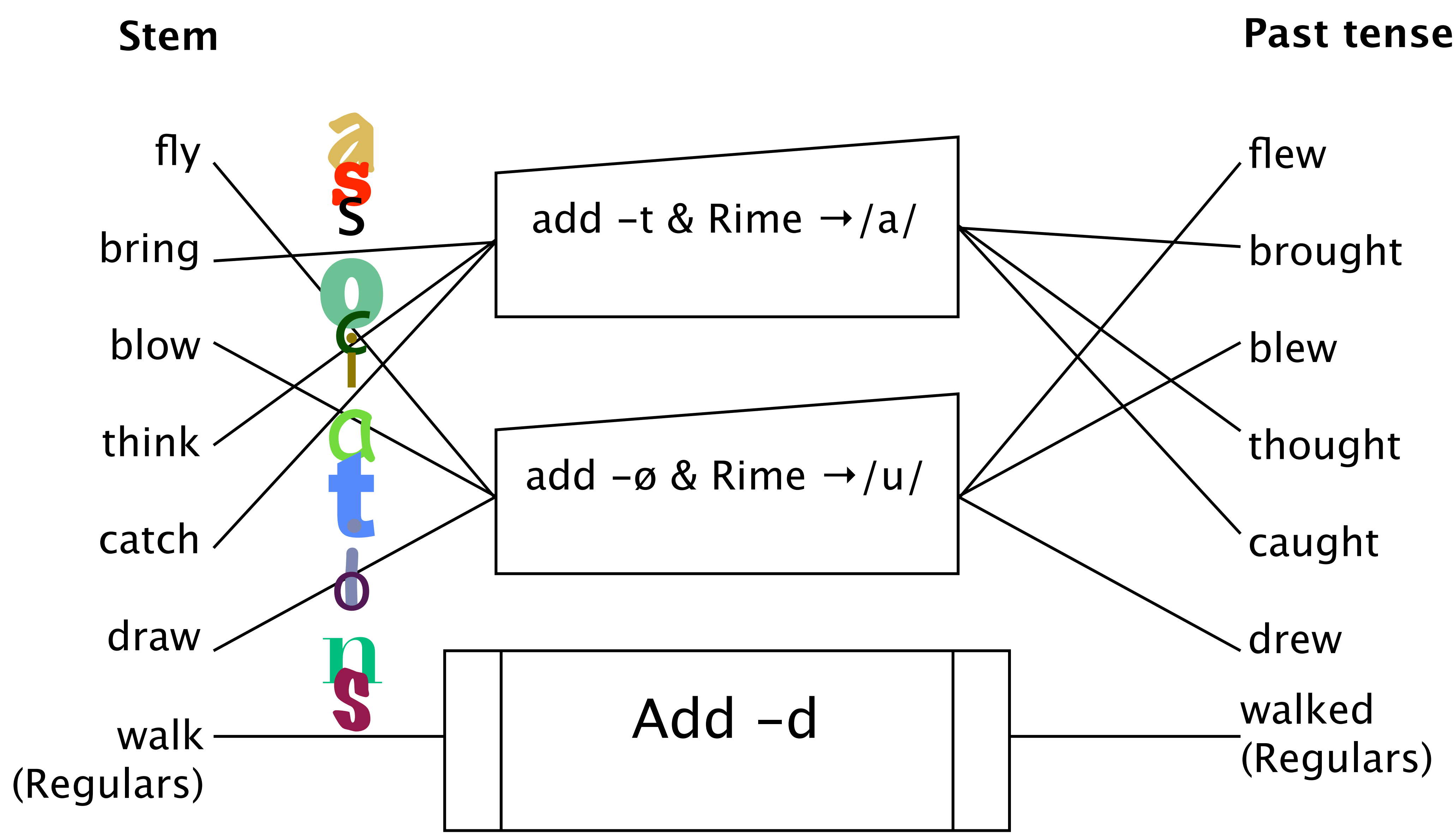
Morphological Change

- a preliminary report
- Where syntax and morphology are different:
  - morphology is (much more) data/induction-driven than syntax: co-existing parameter values, yes, English children using Japanese past tense morpheme, no.
- Where syntax and morphology are the same:
  - historical processes are often attested in language learning

One Long Argument

- Language development: English past tense (Yang 1999, 2002)
- A model of morphological learning
- A model of morphological productivity
    - detour to morphological processing
- A model of morphological change
    - application to Old English weak verbs
- Conclusion and some implications

Rules over Words

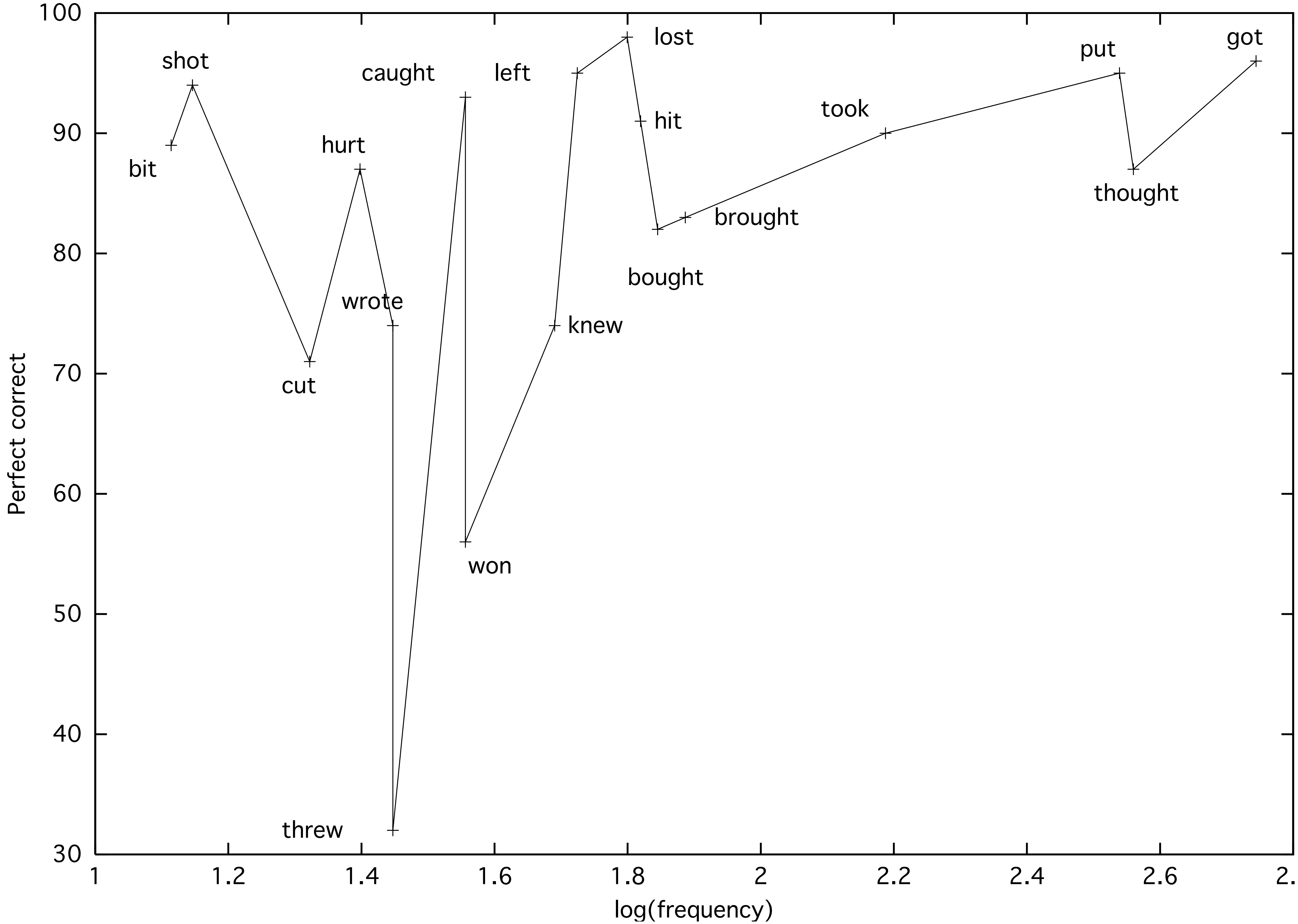


generalization and categorization, however unsystematic

Evidence for Rules

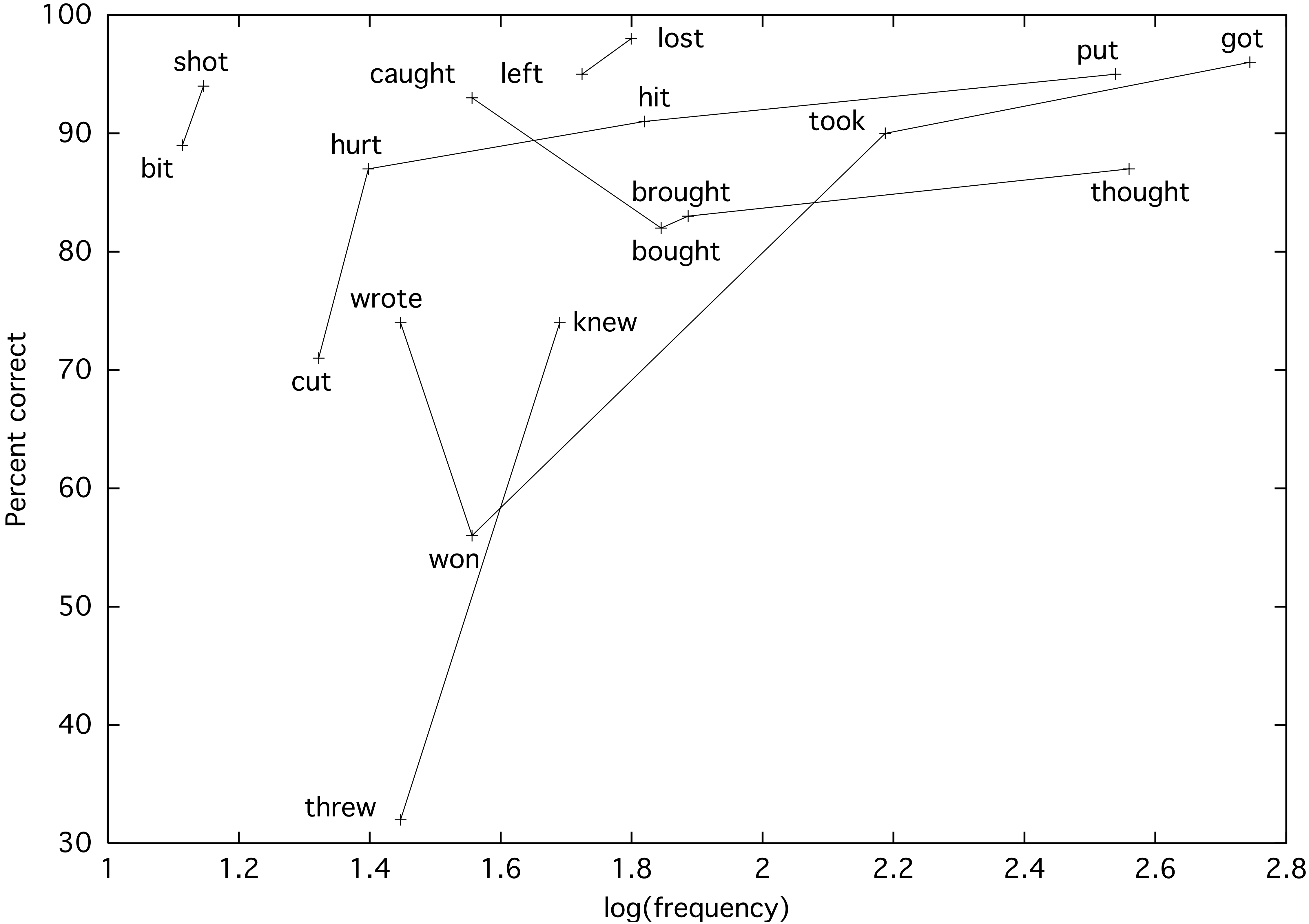
- Within a class, frequency predicts acquisition success
  - almost perfect, while frequency-error correlation without class distinctions is only -0.37 (Marcus et al. 1992)

Words and Rules



frequency-error correlation -.32
(cf. -.37 for all verbs; Marcus et al. 1992)

Rules over Words



Evidence for Rules

- Within a class, frequency predicts acquisition success
  - almost perfect, while frequency-error correlation without class distinctions is only -0.37 (Marcus et al. 1992)
- Control for verb frequency, bigger class--sum of token frequencies of all verbs--makes higher success rate
  - in some cases, “catch-caught” are better than “go-went”
- Morpho-phonologically productive rules make near perfect learning despite token frequency
  - vowel shortening verbs are almost error-free: shoot-shot, bite-bit, lose-lost, leave-left, say-said

Change ~ Learning

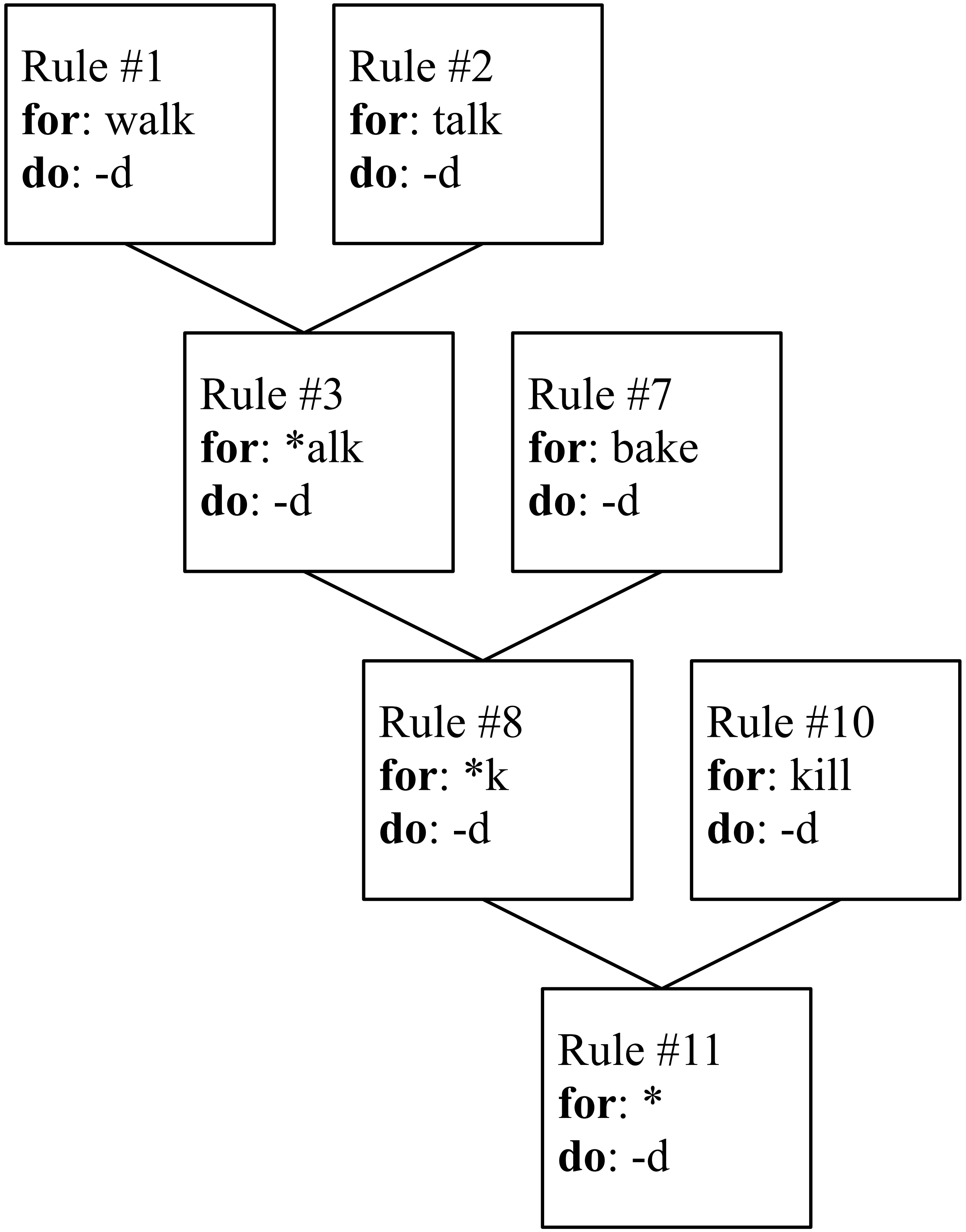
- Analogical **leveling**: words go from a more specific rule to a more general one (e.g., strive-strove)
  - over-**regularization**: hold-holded (quite robust, about 10% of all past tense uses)
- Analogical **extension**: words go from a more general rule to a more specific one (e.g., werede ("weared")-wore)
  - over-**irregularization**: bring-brang
  - exceedingly rare: 40 out of children's 20,000 past tense uses (Xu & Pinker 1995), "bring-brang" is the only error with multiple instances
  - but more on this in German noun plural acquisition

Two Kinds of Rules

- Productive rules: can extend to novel tokens: target of extension/ leveling
- Unproductive rules: **morpholexical** rules (Anderson 1974)
  - e.g., /-t & Rime->a/, as it applies to a fixed set of words (*bring*, *think*, *catch*, *seek*, *teach*, *buy*), and that's it. The hold on words is brute force (repeated exposure in learning)
- Morphological "movement" is controlled by
  - Elsewhere Condition (go to the most specific rule): categorical
  - Frequency effect: stochastic

* This view is consistent with the interpretation of analogy by some linguists (Kiparsky 1974, Anderson 1988).
* We need a theory of morphological productivity

Background: rule learning



Mooney & Califf (1995)
Sussman & Yip (1997)
Molnar (2001), cf.
Mitchell (1978)

A->B/C____D

replace conflicting features with *s (don't care)

Only model that learns minority defaults (e.g., German plurals)

Feature space over which induction operates is completely "up to you"

Imagine ...

- The child has only two words: sing-sang, ring-rang
- Perfectly good rule: **If** /ing/ **Then** /ang/
- Exceptions accumulate as vocabulary grows
  - bring, sting, swing, wing (maybe not in Texas...)
- **If** /ing/ **Then** /ang/ doesn't work so well
- "add -d" would fare better as it doesn't have too many exceptions:
  - how many is "too many"
  - children are exceptionally good at this: morphological errors are almost all omission rather than substitution (Xu & Pinker 1995, Phillips 1995, Guasti 2002)

Terminologies

- **Default:** backup option, when all fails. Maximally general rule (*'s across the board)
- **Productive**: generalizable, "**IF** verb ends in vowel **THEN** add -n"
    - doesn't have to be 100% reliable (can have exceptions)
    - defaults are productive, productive rules are not necessarily defaults
- When is a rule **productive**? When is it **morpholexical** (Anderson 1972)?

Online morphology

- An **evaluation measure** based on the complexity of linguistic performance
  - there are other ways of doing this: ideally, one needs some empirical motivation for a particular approach
  - performance metric feeds back into the grammar
- Processing of morphologically complex words (Caramazza 1997, Levelt 1999)

1. **word search** (e.g., “walk”)
2. **rule search** (e.g., “add -d”)
3. **rule application** (e.g., “walk”+”d”=”walked”)

Rule Search

- Take Elsewhere Condition seriously
- Rule for sibilants in noun plural formation (OK, it's phonology ...)
  - **If** W=fish **then** use the rule "no change"
  - **If** W=perch **then** use the rule "no change"
  - **If** W=tooth **then** use the rule "oo->ee"
  - ....
  - Epenthesis

Bin Search

- Forster (1976): words are organized into orthographic bins, and are listed in decreasing order by frequency
  - non-words take longer to recognize as actual words must be rejected first
  - reaction time is better predicted by **rank** within bins rather than frequency (Murray & Forster 2004)
  - straightforwardly captures frequency effects in lexical processing

Rule-based Bin Search

R: **If** X **Then Y**
α: # of words that **R** applies to
β: # of exceptions that **R** does not apply to **but could**
N= α+ β
(R="add -d to verbs to make past tense", **N**=12,000, **β** =150)

- **If** $W=W_1$ **Then** ...
- **Else If** $W=W_2$ **Then** ...
- ...
- **Else If** $W=W_\beta$ **Then** ...
- **apply R** (**α** items)
- exception clauses ranked by frequency

Optimal Bin

- Ranking seems computationally taxing
- **Randomized algorithms**: near-optimal methods that are computationally trivial
  - **Move Front** (Knuth 1966), **Move Up** (Rivest 1976)
  - fair to assume the human lexicons are near optimal wrt frequency
- Frequency effect = order in which the list is accessed
  - could use tree-like structures (Sleator & Tarjan 1985) for Cohort effects (Marslen-Wilson 1987)

Predictions

- Exceptions exhibit frequency effect
  - This is strongly true (Prasada et al. 1990, Seidenberg 1992, Jaeger et al 1996, Clahsen 1999)
- α's are **slower** than β' during rule search
  - exceptions are expensive: the obedient guys have to wait
  - This is subtle: note that we do not predict an overall reaction delay as reported in Prasada et al. (1990) and Seidenberg (1992), among others

Inflection = Search + Application

- Inflection involves finding the rule and applying the rule, both take time
- Frequency of rules: sum of token frequency
  - Sereno & Jongman (1997): plural nouns (add -s) are faster than past verbs (add -d)--more evidence for “-s” than “-d”
    - also possible: there are few plural irregulars, so the regular nouns are arrived faster
  - Taft (2004): frequency effect can be factored into stem freq., affix freq and **composition cost.**
  - Yang (2002): input frequency matched, different classes of irregular verbs have very different error rates in acquisition (“know-knew” vs. “teach-taught”, “went” vs. “shot”!)

Rule Search Predictions

- Must control both word frequency (for lexical lookup) as well as rule frequency (for rule application)
- English past tense: “add -d” makes up >40% of tokens, far more than any individual irregular class
  - hence regulars are faster than irregulars
- Better test: German past participles (-t and -n are more evenly matched in token frequency)
  - Clahsen et al. (2004): the production latency for regulars is longer than irregulars when word frequencies are matched
  - only for high frequency verbs: that’s OK for us, as the children is making decisions on productivity on a small vocabulary consisting of mostly high frequency words

Shortcuts

- search needn't be exhaustive: German is a case in point
- English nominals:
  - IF **grow, warm, deep, ...** then "add th"
  - IF ending in -**able** then call **-ity** subroutine (this skips hundreds of potential entries for search and rejection)
  - Add -**ness**.
- Actual productivity studies must take into account the specific morpho-phono-syntactico-semantic properties that define morphological classes

To be or not to be

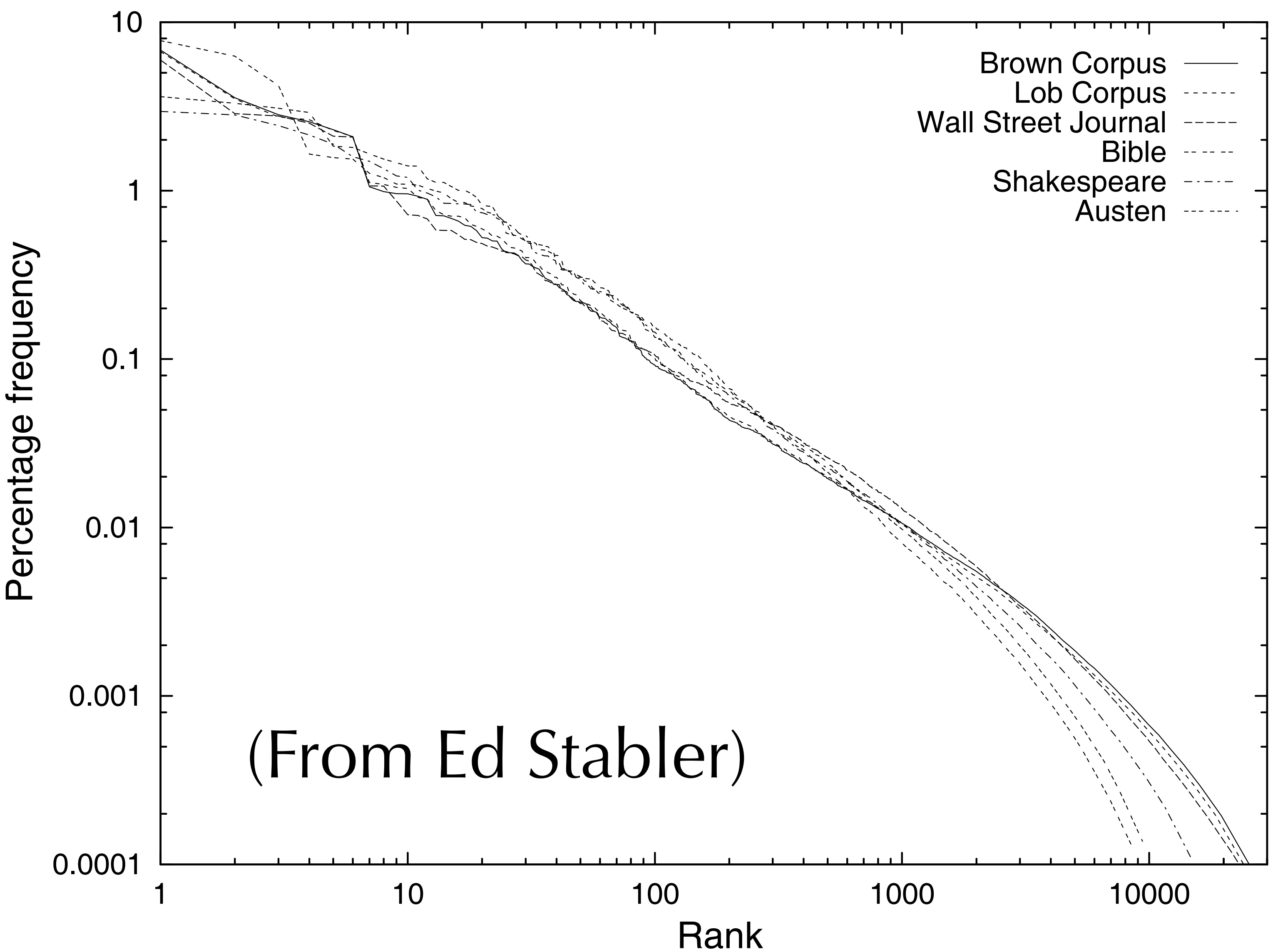
- **If R is productive:**
  - the **α** good words cannot be accessed until all **β** exceptions are accessed and rejected; let **T(α, β)** be the excepted time complexity of retrieving **(α+β)** words with β exceptions
- **If R is morpholexical (i.e. unproductive):**
  - all **N=(α+β)** words are listed as exceptions--even for the **α** good guys that do follow **R**; let time complexity is **T(N,N)**
- <u>Conjecture</u>: The learner will postulate **R** to be productive if **T(α, β) < T(N, N)**

T(N, N)

Assume a Zipfian distribution of word frequencies

Rank * Frequency = Constant

$$T(N,N) = \sum_{i=1}^{N} f_i \times i$$



(From Ed Stabler)

$$f_i = \frac{1}{iH_N}$$

$$H_N = \frac{1}{1} + \frac{1}{2} + \ldots + \frac{1}{N}$$

$$T(N,N) = \frac{N}{H_N}$$

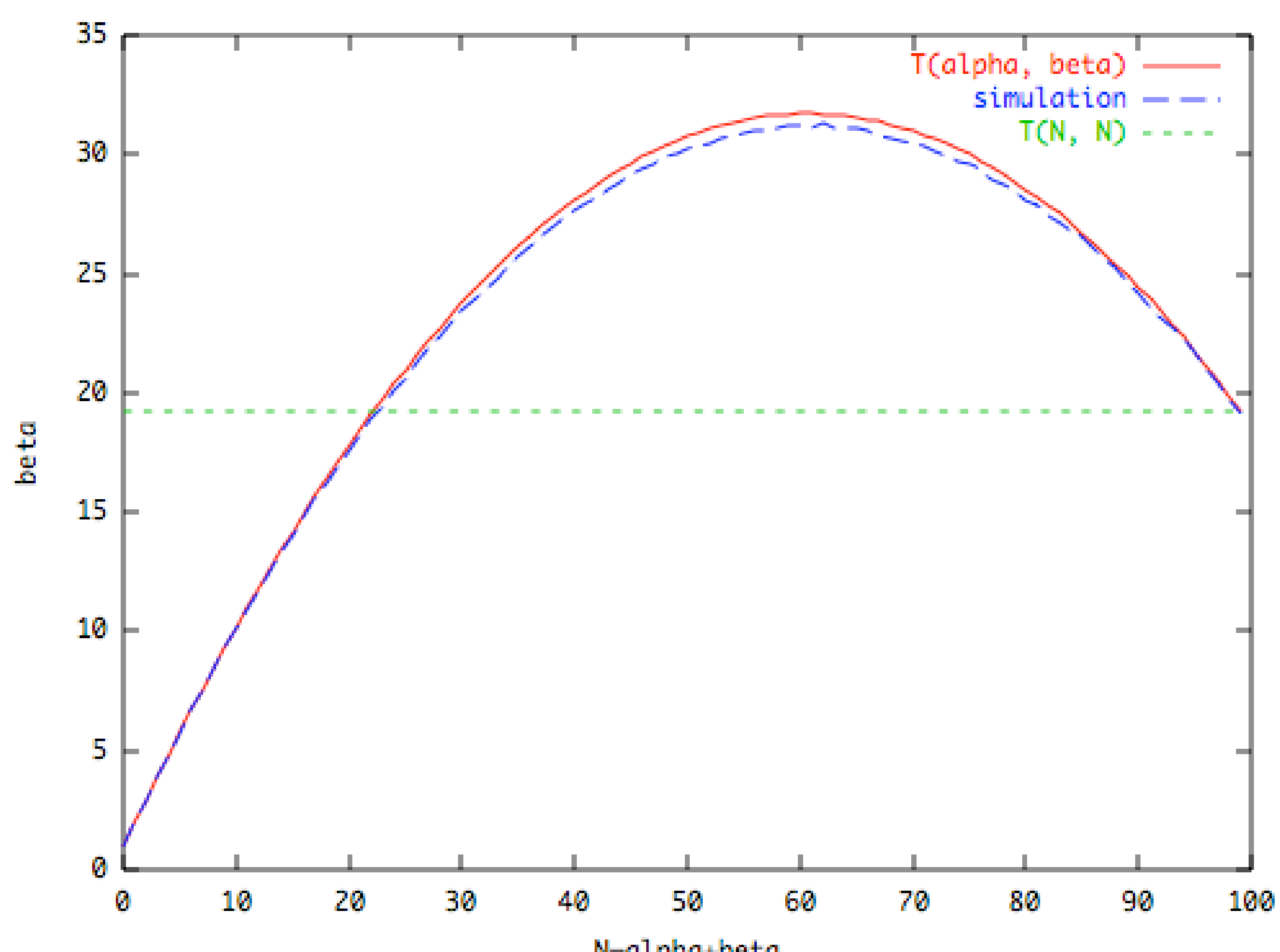
T(α, β)

$$T(\alpha, \beta) = \mu T(\beta, \beta) + (1-\mu)\beta, \text{ where } \mu = \frac{\beta}{\alpha+\beta}$$

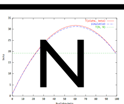
solving for **β** in T(**α**, **β**) = T(**N**, **N**)

$$\beta \approx \frac{N}{\ln N}$$

Joint work with Sam Gutmann (Northeastern)



Tolerance Principle

 N	β
10	4
20	6
50	12
100	22
200	37
500	80
1000	144
2000	263
5000	587
10000	1085

- Productive rules can tolerate few exceptions (sub-linear)
  - English has 150 irregular verbs, so there needs more than 1000 verbs in all: the “add -d” rule is safe.
- Makes predictions about morphological productivity (room for individual variation)
- Makes predictions about morphological learning in relation to word frequencies
- Makes predictions about morphological change

Frequency and Learning

- U-shape learning curve
- most of English past tense probability mass (**token**) is irregular (60%; Grabowski & Mindt 1995), and these will be the earliest verbs: at this point, -ed rule will have a lot of exceptions and will be unproductive
- as more verbs are learned (**types**), -ed rule gets more support and will become productive
- U-shape learning may **not** be universal: it only shows up in exceptions are exceptionally high in token frequency and depends on the particular composition of a child's vocabulary (in past tense, Adam clearly shows it, but other children are not so clear)
- No evidence for U-shape learning in plurals (Brown 1973, Falco & Yang 2005): irregular nouns are fewer and lower in frequency

Lopsided Advantage

- Emergence of productivity: the first instance of overregularization (e.g., “hold-holded”)
- At this point, the child must have considerably more regulars than irregulars according to TP
- Marcus et al. (1992)
  - Adam: 136/218=62% regular, Sarah: 124/193=64% regular
  - somewhat lower than what TP predicts, possibly due to under-estimation of regulars, which are lower in frequency

Two Empirical Cases

- Briefly, acquisition of German plurals: fewer exceptions than meet the eye
- Old English weak verbs: the leveling of productive rules

German noun plurals

- Five classes: Kind-Kinder (children), Wind-Winde (wind), Ochs-Ochsen (oxen), Daumen-Daumen (thumbs), Auto-Autos (cars)
- Marcus et al. (1995), Clahsen (1999), Pinker (1999), etc.: only -s, numerical minority (7%) is the default
  - All other four classes are memorized by rote
- "A quick glance at German shows that the four irregular classes of plurals show no systematic similarity whatever. The horrors of German are real: one must sort each irregular noun into its proper class, as in the traditional rule-based view" (Yang 2000)
  - **Wrong! (93%: too many exceptions for -s to be productive)**

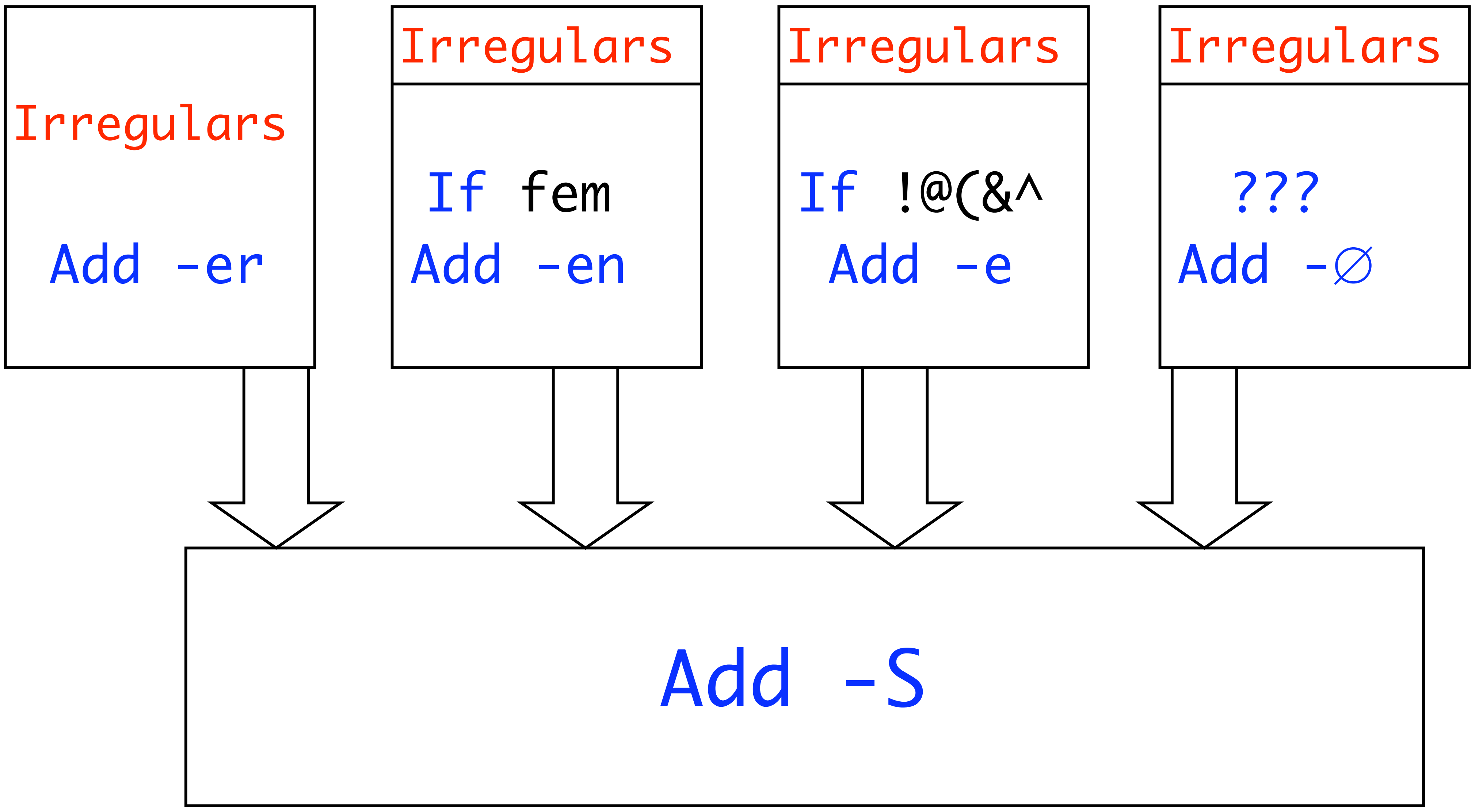
Feminine nouns

- Many feminine nouns add -(e)n
- Some exceptions (Umlaut): Axt (ax), Bank (bench), Braut (bride), Brust (breast), Faust (fist), ... Wand (wall), Wurst (sausage)
- A few, all monosyllabic ones that add -e: Tann (wood), Trupp (troop), ...
- **β**≈80, **N**=500: around 500 feminine nouns taking -(e)n can ensure a **productive** rule for a subclass of nouns
- There are at least 9500 feminine nouns in German (CELEX), 3600 excluding compounds: **<u>there is a productive rule</u>**
    - **IF** feminine **Then** add -(e)n, without some exceptions)
- This is what a group of German morphologists have argued (against the Pinker-Clahsen line): Dressler (1999), Wiese (1999), Wunderlich (1999).

Evidence for feminine rule

- Klampfer et al. (2002): German children use wrong irregular (**not s**) pattern **10%** of time
  - ≫0.2% in English children (Xu & Pinker 1995), and comparable to English over-regularization
- Elicitation (low frequency nouns) error rate 18.5%, 14.6% of which are -(e)n (Clahsen 1999)
- Penke & Krause (2002): aphasics and unimpaired subjects
  - both aphasics and unimpaired subjects show no frequency effect for -(n) feminine nouns
  - lack of frequency effects generally taken as evidence for **regular** rules (Clahsen 1999)

Stacked Bins



very few overuse involve –er (not much data on –∅)

Penke & Krause (2002): add –s class uniformly slower than other classes: bins are search serially as well

horrors of German still real: have to memorize genders!

Productivity and Change

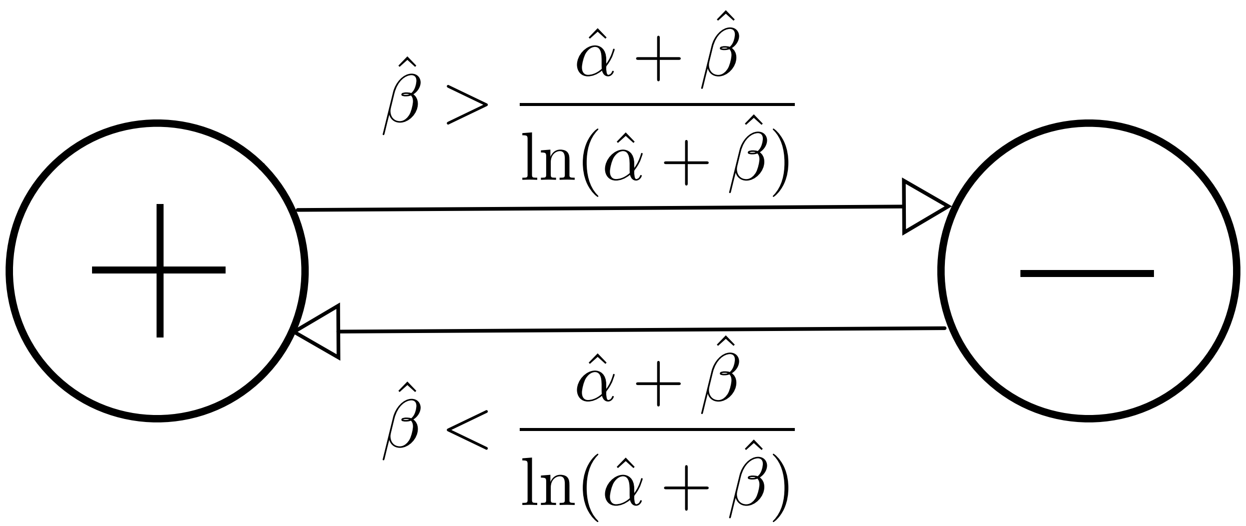
- if a rule is productive (**+**), then it won't lose members but could attract other words matching its structural description (**extension: German -(e)n rule**)
- if a rule is unproductive (**-**), then it could lose members to more include rules that match the structural descriptions of these words (**leveling: English overregularization**)
- productivity is determined by α and β
- productivity may change as a result of learning: in the next generation, $\boldsymbol{\alpha'}$ and $\boldsymbol{\beta'}$ may well be different

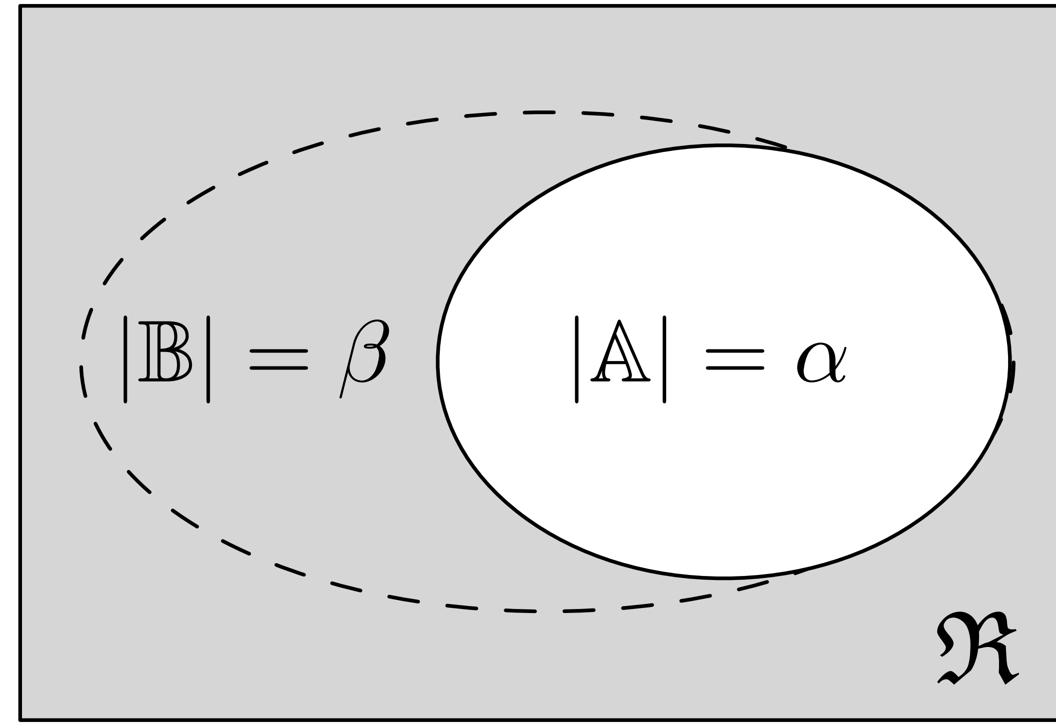
α and β in perturbation

social/linguistic factors	$\alpha' \sim \alpha$	$\beta' \sim \beta$	Example
variation	$>$	$>$	conflicting membership doublets: dive-dived-dove
innovation	$<$	$>$	"Kosovians can move back in" (W.)
phonological reduction	$>$	$<$	Many cases (OE later)
under-learning	$<$	$<$	smaller vocabulary effects (Kroch 1978)
language/dialect contact	?	?	it depends ...

perturbation as a uniform framework to capture the effects on morphological change from diverse sources

Jekyll and Hyde





a. $A_+ \rightarrow A_+$ (A stays in the productive state): $\mathbb{B}$ is subject to extension to A, though $\mathbb{A}$ is not subject to leveling to $\mathfrak{R}$ as it's protected by the Elsewhere Condition.

b. $A_- \rightarrow A_-$: $\mathbb{A}$ is subject to leveling to $\mathfrak{R}$, $\mathbb{B}$ is not subject to extension as there is no productive rule for it to go to.

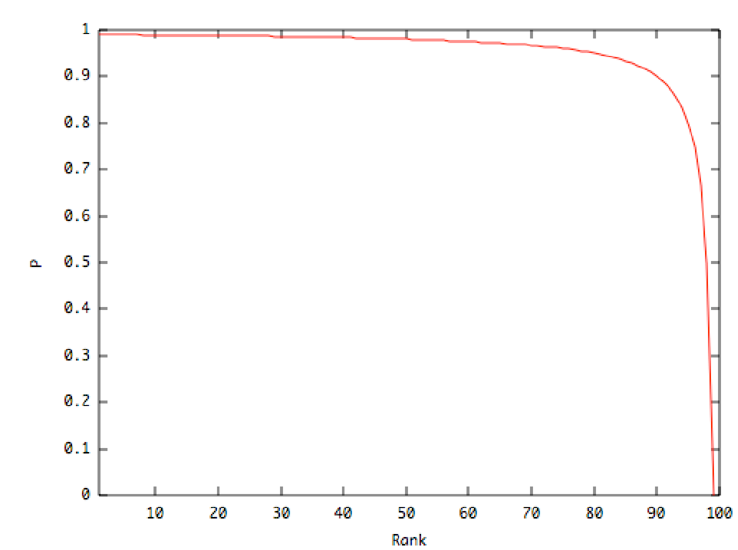
c. $A_+ \rightarrow A_-$: $\mathbb{A}$ is subject to leveling as the productive rule disappears (i.e., $\mathbb{A}$ is now the exceptions to the rule $\mathfrak{R}$).

d. $A_- \rightarrow A_+$: $\mathbb{A}$: $\mathbb{B}$ is subject to extension to A (new sheriff in town), and $\mathbb{A}$ is no longer subject to leveling to $\mathfrak{R}$.

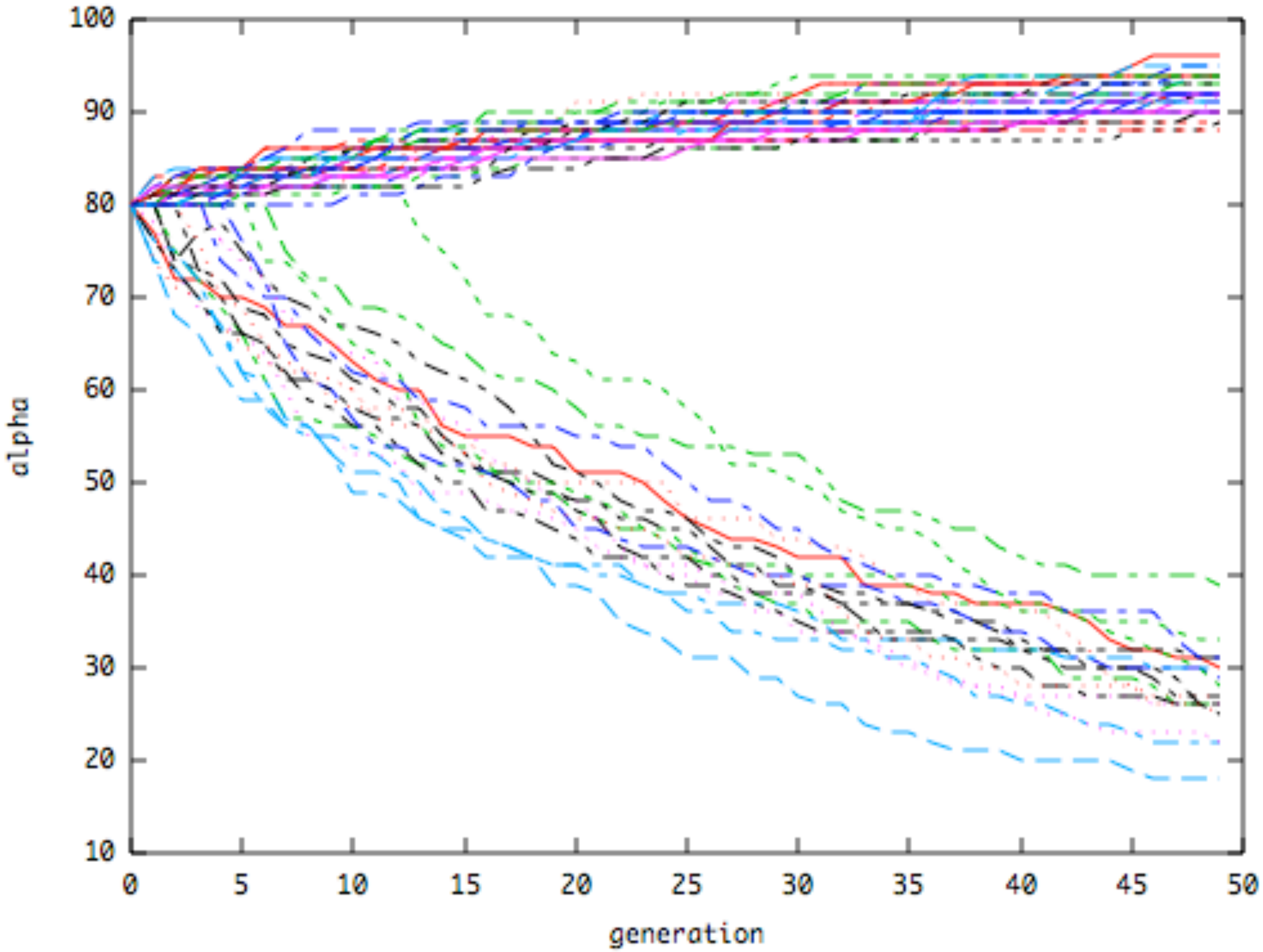
A Formal Model of Change

- initial α and β in the founder population
- for each generation of learners
  - obtain α′ and β′
  - compute the productivity of A based on (α′+β′)/ln(α′+β′)
  - carry out frequency-dependent extension/leveling for words with the procedure described in previous page*
  - iterate

If the nature of the perturbation is clearly understood, it's possible to derive formal results of productivity as a Markov process (random walk)



A simulation before reality



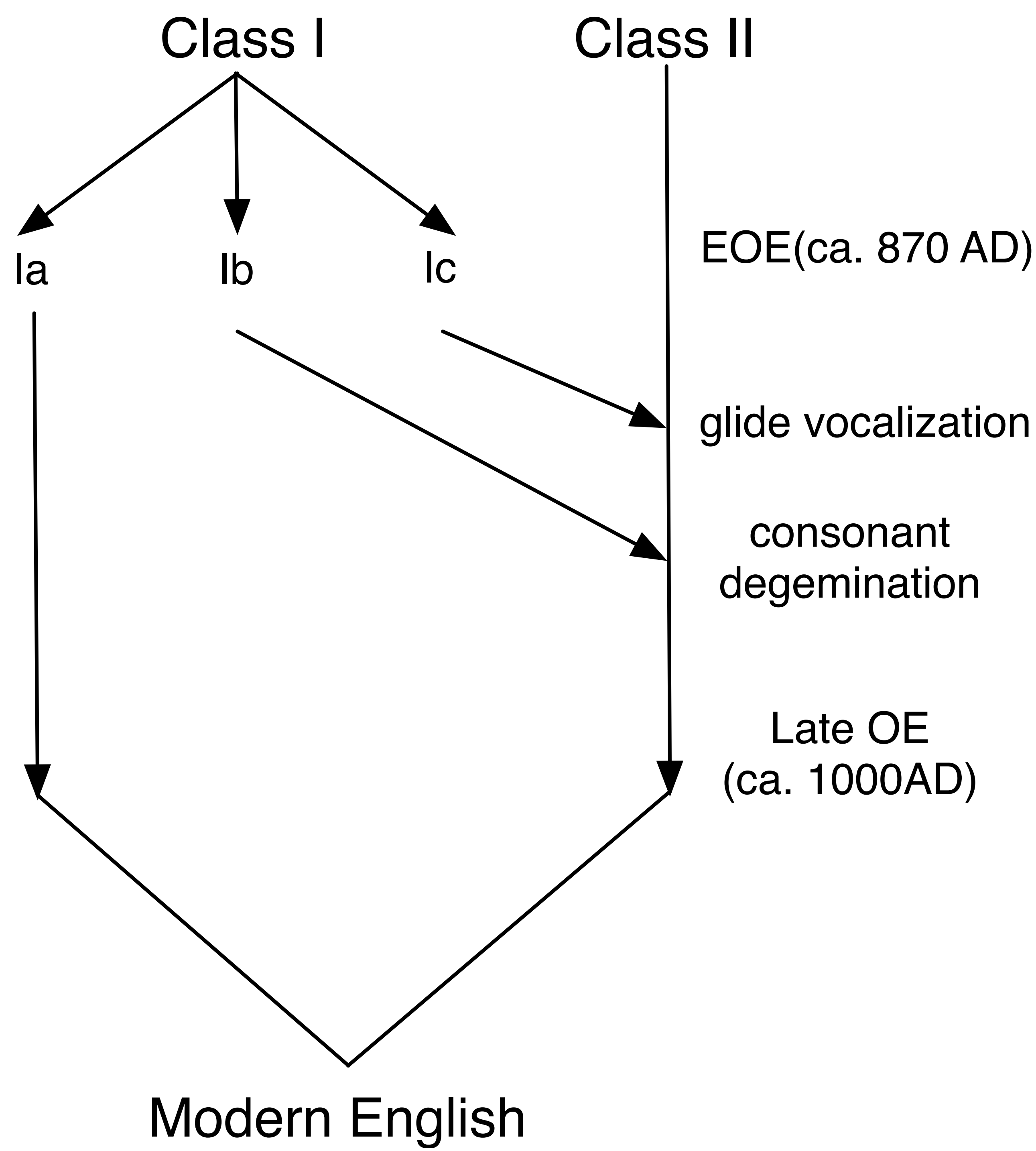
(E)OE Weak Verbs

- Class III: various odd balls
- Class II: present suffix **i**, rather heterogeneous and thus more "open" (recall the rule learning model)
- Class I: present suffix **j**
  - Ia: heavy stems (long vowel or final consonant cluster), **j** deletes
  - Ib: light stem, stem-final consonant germinates, **j** also deletes
  - Ic: light stem ending in ***r***, no **j** deletion, no germination
- Fairly well-defined structural (phonological) descriptions of the Class I rules

(E)OE Weak Verbs

	Class Example	Ia dēman (judge)	Ib fremman (do)	Ic nerjan (save)	II lufiian (love)
	1st.sing	dēm+e	fremm+e	nerj+e	lufij+e
Present	**2nd.sing**	dēm+st	frem+est	ner+st	luf+ast
	Plural	dēm+ath	fremm+ath	nerj+ath	lufij+ath
	1st.sing	dēm+de	frem+e+de	ner+e+de	luf+o+de
Past	**Plural**	dēm+don	frem+e+don	ner+e+don	luf+o+don
	P. Participle	dēm+e+d	frem+e+d	ner+e+d	luf+o+d

Diachronics of Weak Verbs



- glide vocalization: **j** in Ic and II (1st.sing) became **i**
  - Ic merged with II, and picked up **o** in past tense
- degemination: the distinction between Ib and **II′** lost
- Ia and II″ remained in Late OE: further phonological change (vowel reduction, elimination of final **i** in II) lost the distinction between the two classes

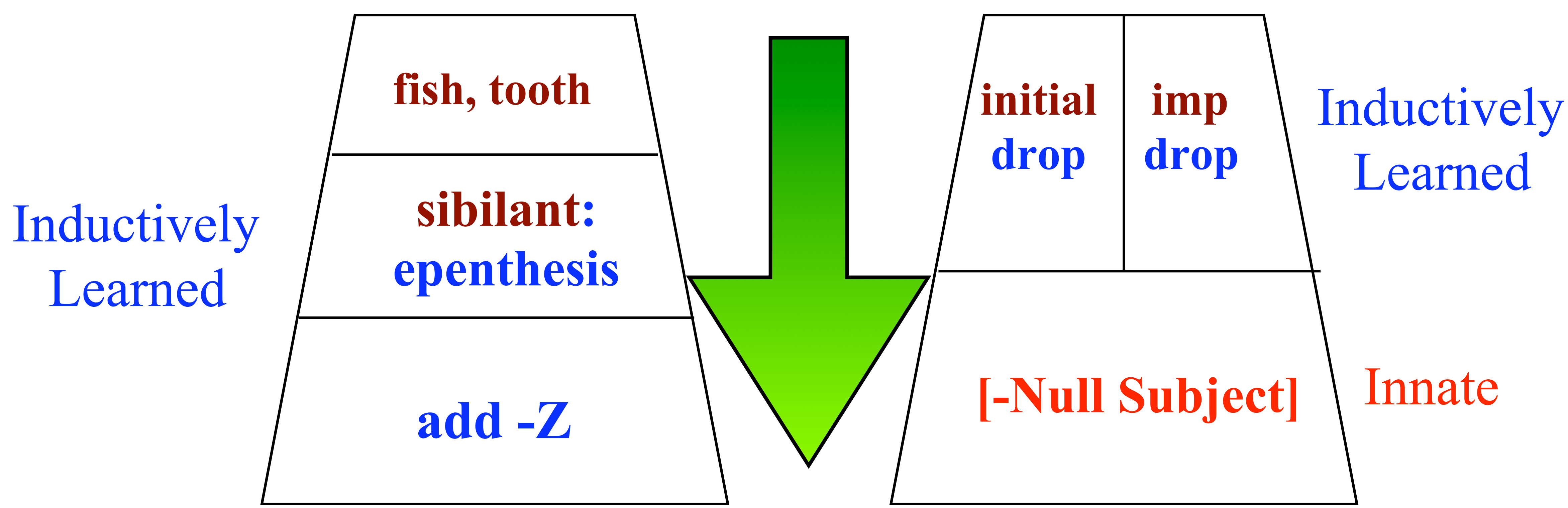
How Phonology Killed Morphology

Class	#
Ia	128
Ib	64
Ic	32
II	256*

Hare & Elman (1995)

- **Ic->II**: β=32, α=256, threshold=50)
  - leveling of Ic **must** happen
- **Ib->II′** (β=64, α=288, threshold=60)
  - leveling is highly **probable**
- **Ia->II″**: β=128, needed about N=α+β=900 verbs altogether to level)
  - Need better verb counts from the right period, but this is easily met at least in modern English
- NB: leveling was not complete, some historical weak verbs didn't level and became strong/irregular (*send-sent*, *bend-bent* in Ia), and there were also lots of variations (Lass & Anderson 1975)
- Not a complete theory of change, but rather what might happen after some (unpredictable) change has taken place

Elsewhere Everywhere

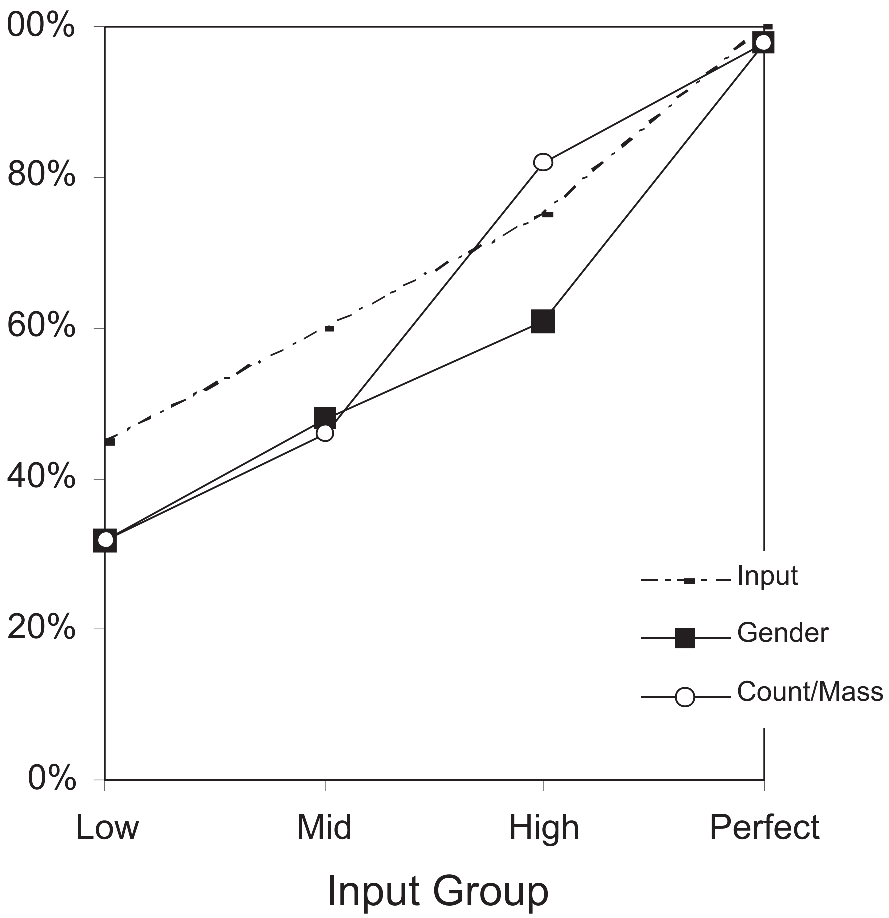


- application of syntactic rules govern by Elsewhere Condition (Poser 1988, Andrews 1990, Hankamer & Mikkelsen 2005)
- Only difference between phonology and syntax:
  - phonological default (if it exists) may be **inductively learned**
  - syntactic default is **innate**: parameter value of the native language

Productivity in Lexical Change

- Lexical learning plays an important rule
- Within a homogeneous class of words, the learner may see mixed patterns: e.g., some verbs trigger subject inversion (**α**), and some do not (**β**)
- Tolerance Principle can tell the learner which pattern is productive: e.g., compare **β** with **N/lnN**
- As α and β evolve, one may observe rather abrupt changes in the syntactic behavior of a lexical class
- Phonological change: resistance to vowel shift in a class of words that are characterized semantically (Quebec French, Yaeger-Dror 1996)

L2 learning



- Artificial language learning (Kam & Newport 2005): regularization under inconsistent data.
  - adults probability match up to a point, kids forge ahead to make rules, but see the figure on the left
  - Adults handle α and β somewhat differently which may affect α and β in later generations

Conclusion

- exceptions--cumulative or abrupt--can change regularity
- it is useful to build basic models that explore the dynamics of learning and change: one can always make them more complicated and closer to reality later
- formal models of synchronics may shed light on the study of diachronics
- formal models of synchronics can certainly shed light on the study of synchronics