

# Evaluation Measure in Language Acquisition

Charles Yang  
University of Pennsylvania

LING50  
MIT

# From LSLT to Aspects: A Psychological Turn

# From LSLT to Aspects: A Psychological Turn

- LSLT
  - “any simplification along these lines is immediately reflected in the length of the grammar” (§26: MDL)
  - “defined the best analysis as the one that minimizes information per word in the generated language of grammatical discourses” (§35.4: entropy)
  - 23.771 Mathematical Backgrounds for Communication (Hall/Partee)

# From LSLT to Aspects: A Psychological Turn

- LSLT
  - “any simplification along these lines is immediately reflected in the length of the grammar” (§26: MDL)
  - “defined the best analysis as the one that minimizes information per word in the generated language of grammatical discourses” (§35.4: entropy)
  - 23.771 Mathematical Backgrounds for Communication (Hall/Partee)
- Aspects
  - “... theories require supplementation by an evaluation measure if language acquisition is to be accounted for ... such a measure is not given a priori, in some manner. Rather, any proposal concerning such a measure is an empirical hypothesis about the nature of language” (p37)

# Three Psychological Conditions

# Three Psychological Conditions

- Formal sufficiency: Does the evaluation measure choose the “correct” grammar?
  - Distributional methods: Harris (1951), Fowler (1952), Holt (1953)

# Three Psychological Conditions

- Formal sufficiency: Does the evaluation measure choose the “correct” grammar?
  - Distributional methods: Harris (1951), Fowler (1952), Holt (1953)
- Ecological validity: Does the evaluation measure operate under reasonable assumptions about the learning data and mechanisms?

# Three Psychological Conditions

- Formal sufficiency: Does the evaluation measure choose the “correct” grammar?
  - Distributional methods: Harris (1951), Fowler (1952), Holt (1953)
- Ecological validity: Does the evaluation measure operate under reasonable assumptions about the learning data and mechanisms?
- Developmental compatibility: Does the evaluation measure employed by the learner produce similar developmental patterns in language acquisition?



# MDL: An Evaluation Measure

$$\text{DL}(D, G) = |G| + \log \frac{1}{p(D|G)}$$

# MDL: An Evaluation Measure

$$\text{DL}(D, G) = |G| + \log \frac{1}{p(D|G)}$$

- MDL is similar with (or equivalent to) many other approaches such as Bayesian inference

# MDL: An Evaluation Measure

$$\text{DL}(D, G) = |G| + \log \frac{1}{p(D|G)}$$

- MDL is similar with (or equivalent to) many other approaches such as Bayesian inference
- A method for hypothesis **selection** rather than hypothesis **proposing**
  - Three conditions for choosing alternative methods, e.g., reinforcement learning, Fourier transform

# MDL: An Evaluation Measure

$$DL(D, G) = |G| + \log \frac{1}{p(D|G)}$$

- MDL is similar with (or equivalent to) many other approaches such as Bayesian inference
- A method for hypothesis **selection** rather than hypothesis **proposing**
  - Three conditions for choosing alternative methods, e.g., reinforcement learning, Fourier transform
- The composition of data

# MDL: An Evaluation Measure

$$DL(D, G) = |G| + \log \frac{1}{p(D|G)}$$

- MDL is similar with (or equivalent to) many other approaches such as Bayesian inference
- A method for hypothesis **selection** rather than hypothesis **proposing**
  - Three conditions for choosing alternative methods, e.g., reinforcement learning, Fourier transform
- The composition of data
- Subset Principle (Berwick 1985): the first Evaluation Measure to influence empirical work in language acquisition

# Evaluation Measuring Parameters

# Evaluation Measuring Parameters

- Aspects: “We want the hypotheses compatible with fixed data to be “scattered” in value, so that choice among them can be made relatively easily” (p61)

# Evaluation Measuring Parameters

- Aspects: “We want the hypotheses compatible with fixed data to be “scattered” in value, so that choice among them can be made relatively easily” (p61)
- Parameters can be viewed as low dimensional description of syntactic variation, or MDL



# Evaluation Measuring Parameters

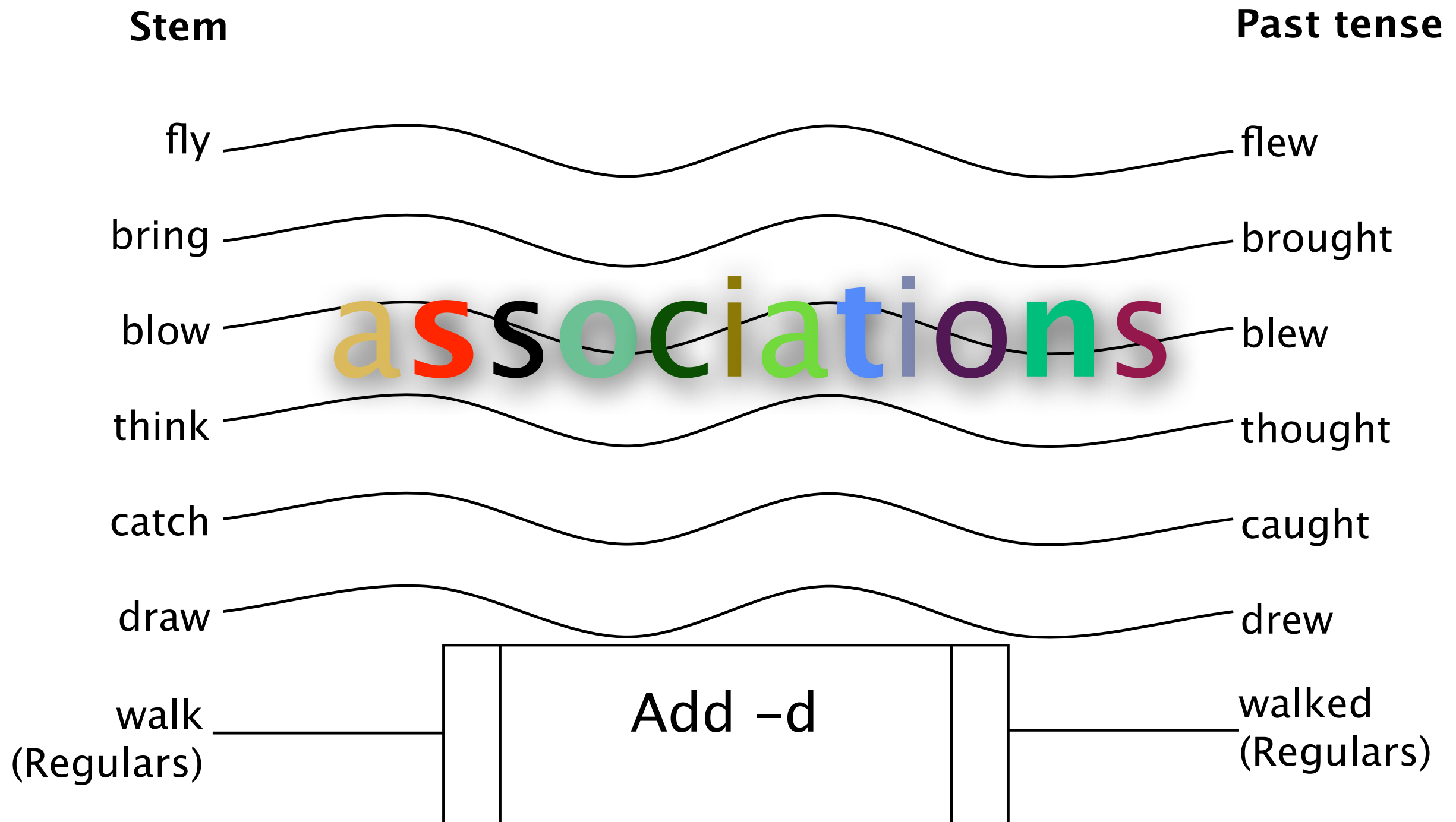
- Aspects: “We want the hypotheses compatible with fixed data to be “scattered” in value, so that choice among them can be made relatively easily” (p61)
- Parameters can be viewed as low dimensional description of syntactic variation, or MDL
- CUNY CoLAG Parameter Domain (Sakas & J.D.Fodor *in press*):  
13 parameters, 3072 grammars, 48086 distinct degree-0 sentences
  - Most parameters are favorable for the learner and can be set independently (thus “scattered” well)

# Evaluation Measuring Parameters

- Aspects: “We want the hypotheses compatible with fixed data to be “scattered” in value, so that choice among them can be made relatively easily” (p61)
- Parameters can be viewed as low dimensional description of syntactic variation, or MDL
- CUNY CoLAG Parameter Domain (Sakas & J.D.Fodor *in press*):  
13 parameters, 3072 grammars, 48086 distinct degree-0 sentences
  - Most parameters are favorable for the learner and can be set independently (thus “scattered” well)
- Evidence for parameters in language acquisition

# MDL in Action

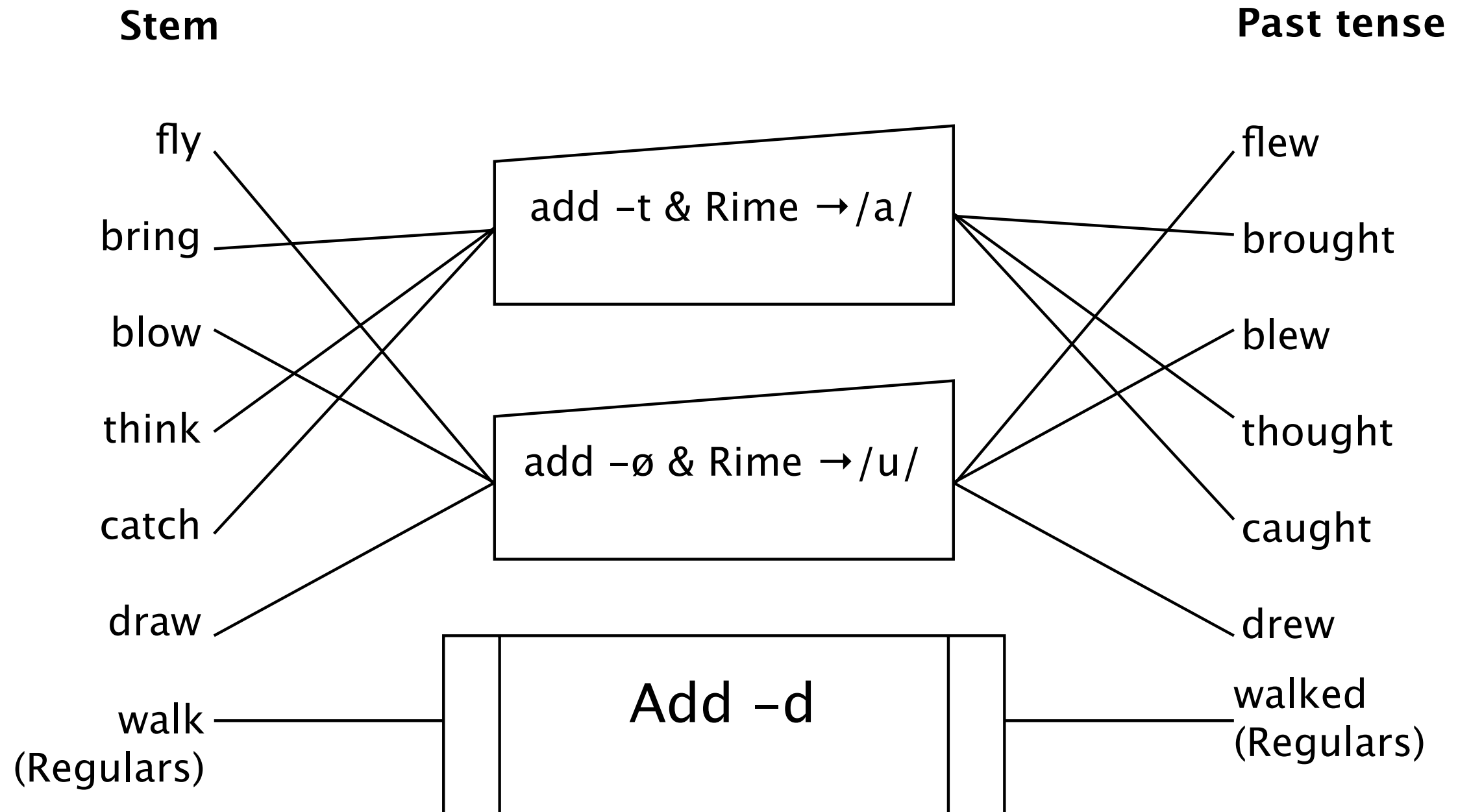
# MDL in Action



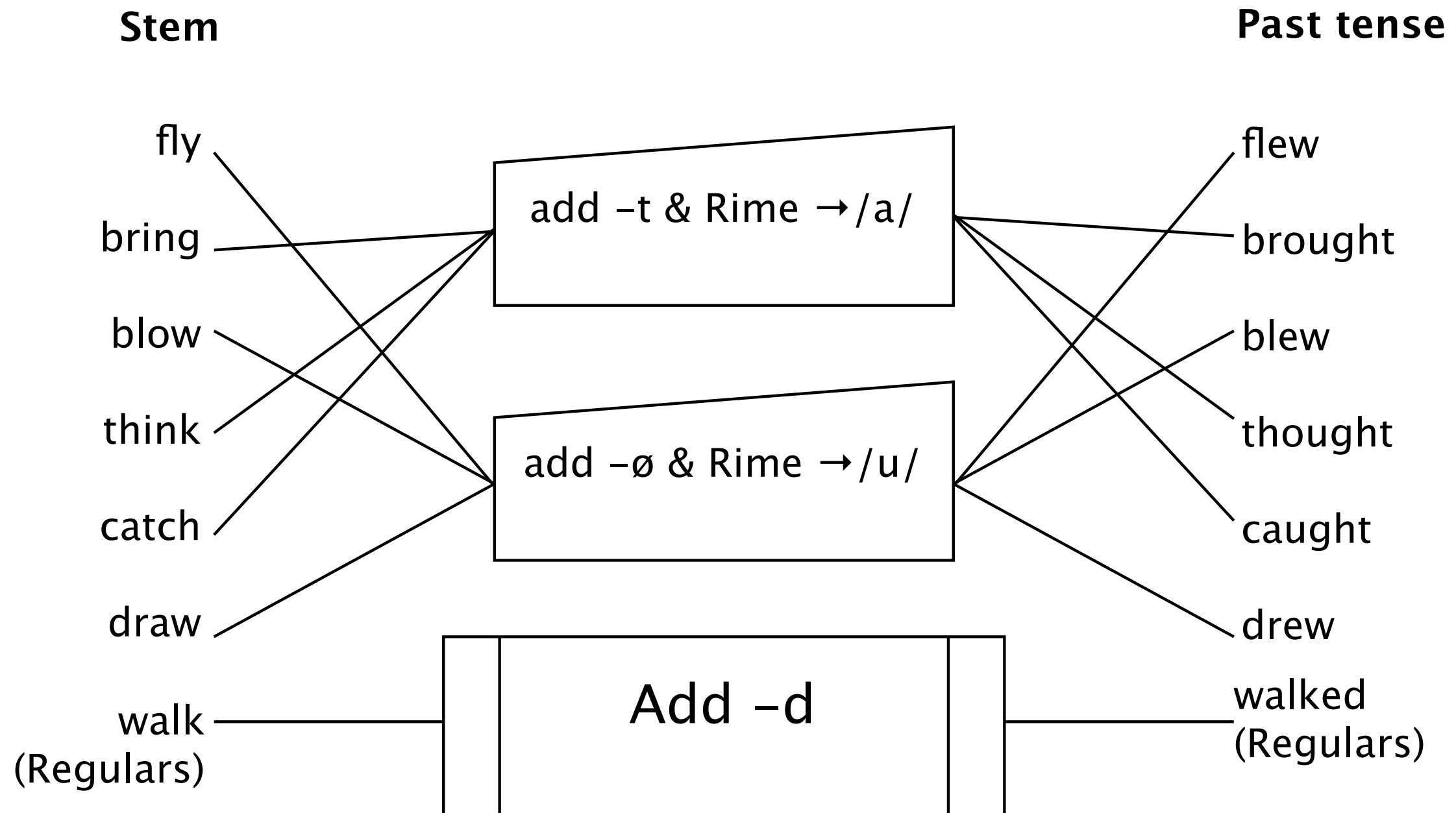
# MDL in Action

# MDL in Action

# MDL in Action



# MDL in Action



“... the acceptance of these Laws (Grimm’s and Verner’s) as historical fact is based wholly on considerations of simplicity”  
Halle (1961: On the role of simplicity in linguistic descriptions)




# Evidence from Children

- Largest past tense analysis to date (Gorman & Yang 2011)
- Free-rider effect: verbs belonging to larger rules learned better

# Evidence from Children

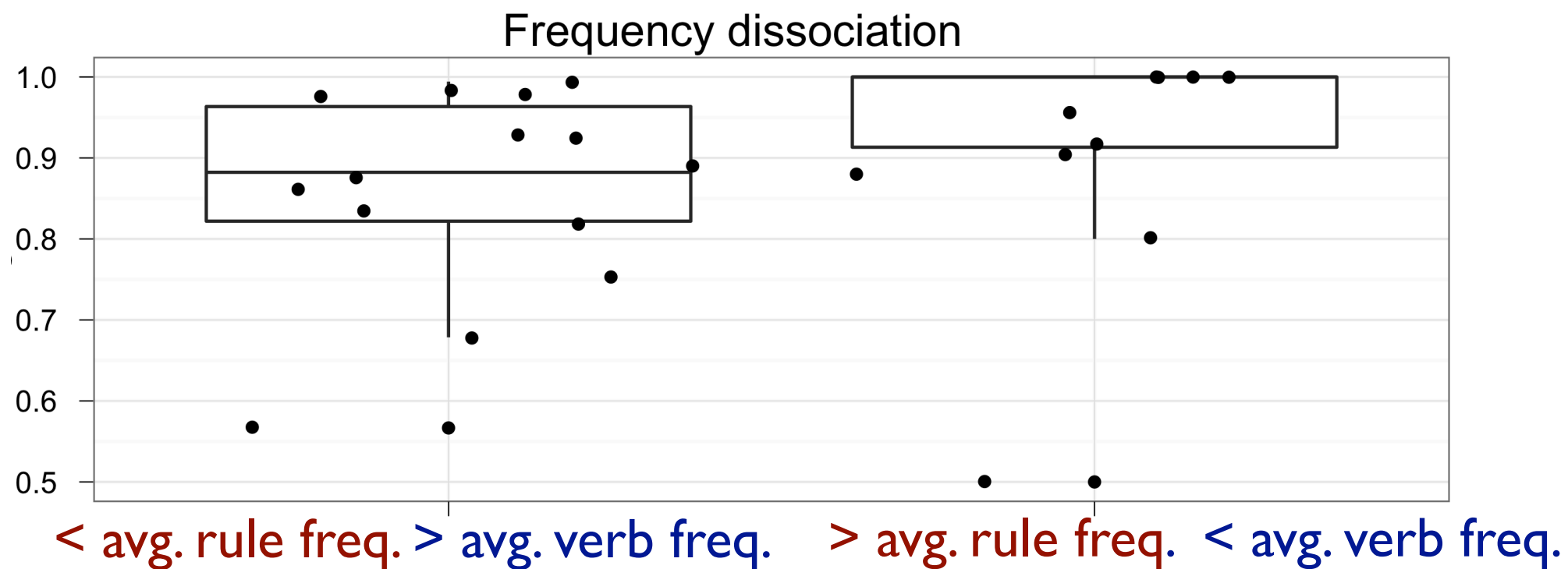
- Largest past tense analysis to date (Gorman & Yang 2011)
- Free-rider effect: verbs belonging to larger rules learned better

	Spearman $\rho$	Kendall $\tau$	G-K $\gamma$
 <b>Abstract rules</b>	<b>0.276</b>	<b>0.191</b>	<b>0.202</b>
Surface rules	0.267	0.180	0.190
<b>Words only</b>	<b>0.128</b>	<b>0.133</b>	<b>0.140</b>

# Evidence from Children

- Largest past tense analysis to date (Gorman & Yang 2011)
- Free-rider effect: verbs belonging to larger rules learned better

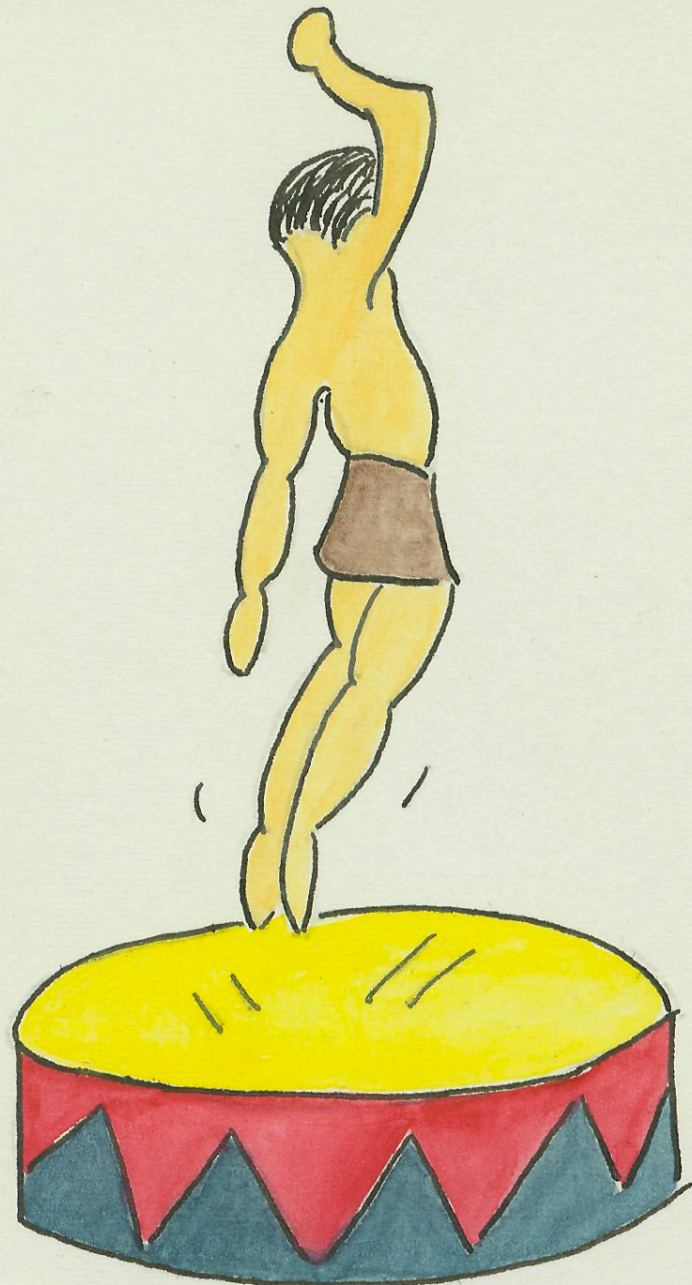
	Spearman $\rho$	Kendall $\tau$	G-K $\gamma$
👉 Abstract rules	<b>0.276</b>	<b>0.191</b>	<b>0.202</b>
Surface rules	0.267	0.180	0.190
<b>Words only</b>	<b>0.128</b>	<b>0.133</b>	<b>0.140</b>



two-tailed Mann-Whitney  $W=156.5$ ,  $p=0.019$

Yesterday he \_\_\_\_\_

Yesterday he \_\_\_\_\_



This is a man who knows how to GLING.  
He is GLINGING. He did the same thing  
yesterday. What did he do yesterday?  
Yesterday he \_\_\_\_\_.

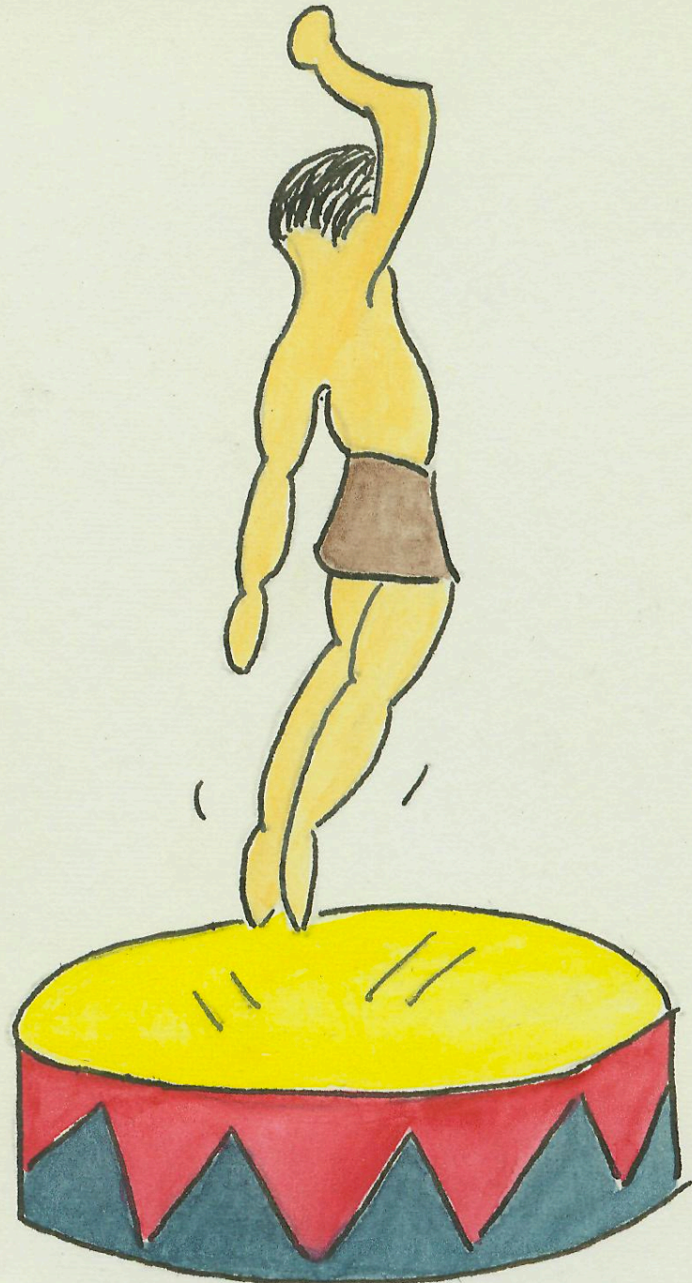


# Yesterday he \_\_\_\_\_



- The forgotten **Wug test** (Berko 1958)
- Only **one** out of 86 children produced *bing-bang, gling-glang*

# Yesterday he \_\_\_\_\_

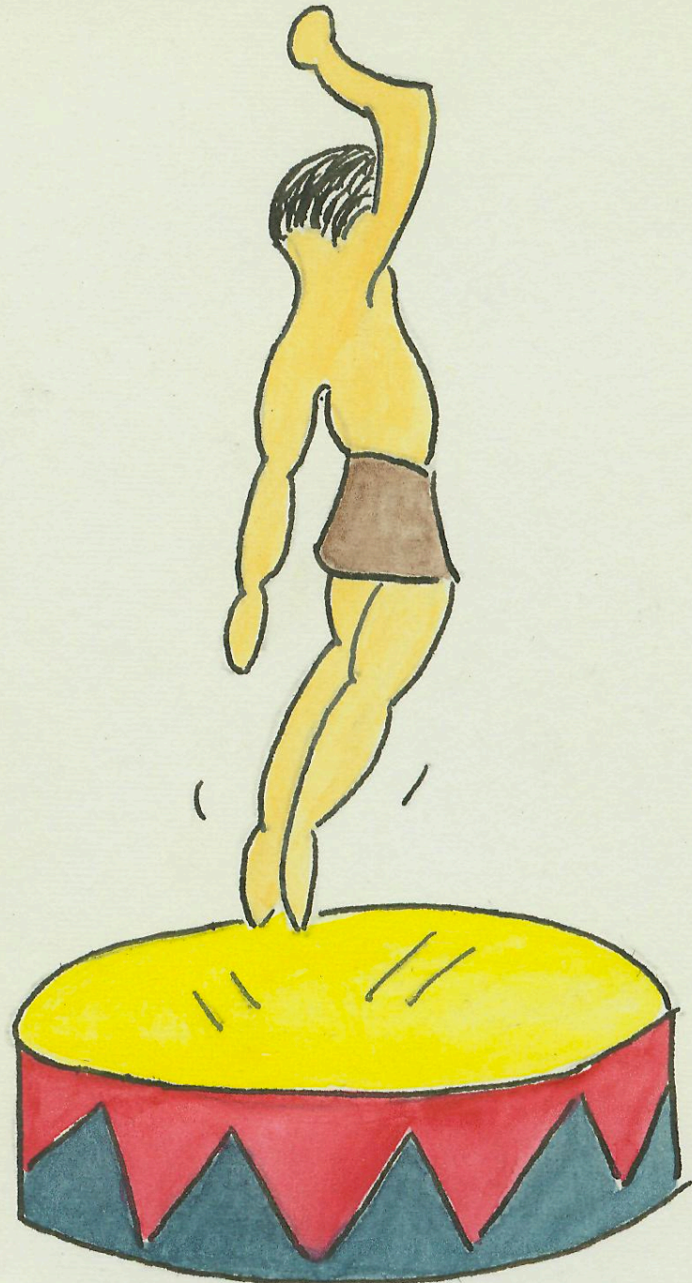


This is a man who knows how to GLING.  
He is GLINGING. He did the same thing  
yesterday. What did he do yesterday?  
Yesterday he \_\_\_\_\_.

- The forgotten **Wug test** (Berko 1958)
  - Only **one** out of 86 children produced *bing-bang, gling-glang*
- Children over-regularize: **8-10%**



# Yesterday he \_\_\_\_\_

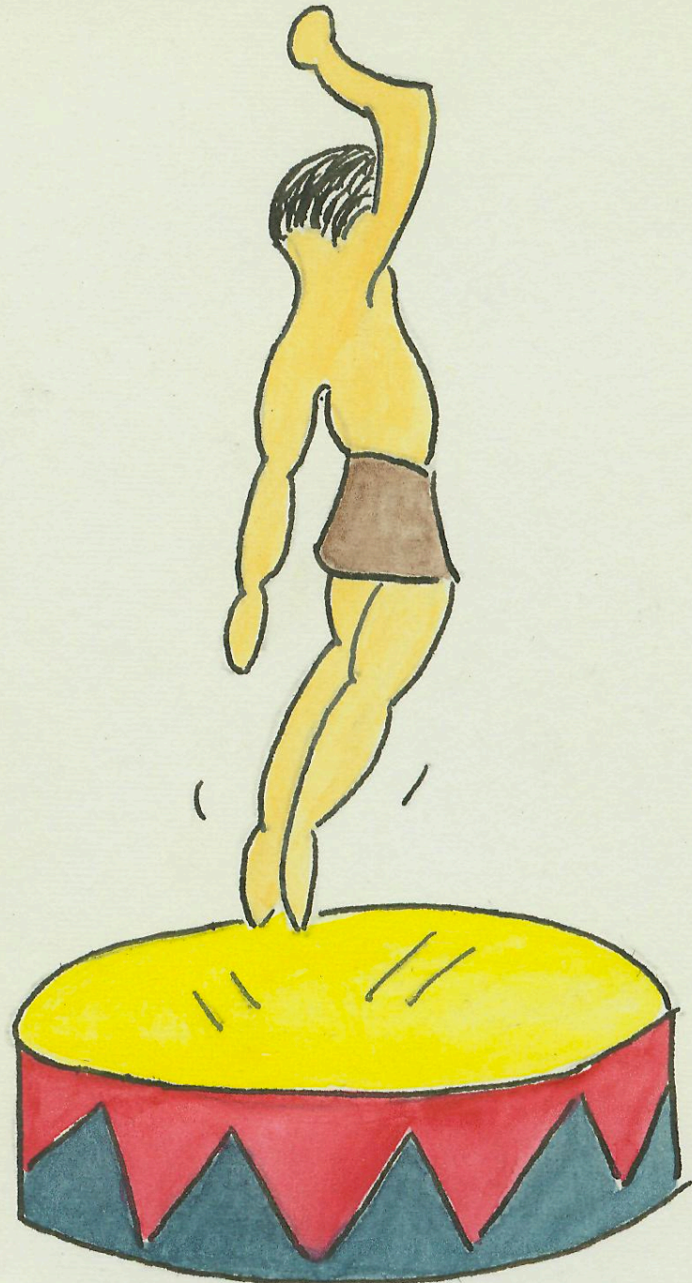


This is a man who knows how to GLING.  
He is GLINGING. He did the same thing  
yesterday. What did he do yesterday?  
Yesterday he \_\_\_\_\_.

- The forgotten **Wug test** (Berko 1958)
  - Only **one** out of 86 children produced *bing-bang, gling-glang*
- Children over-regularize: **8-10%**
- Children over-irregularize: **<0.2%**



# Yesterday he \_\_\_\_\_



This is a man who knows how to GLING.  
He is GLINGING. He did the same thing  
yesterday. What did he do yesterday?  
Yesterday he \_\_\_\_\_.

- The forgotten **Wug test** (Berko 1958)
  - Only **one** out of 86 children produced *bing-bang, gling-glang*
- Children over-regularize: **8-10%**
- Children over-irregularize: **<0.2%**
- Children's Evaluation Measure produces a binary outcome: productive or lexical
  - probability spreading insufficient

# Exceptions in Evaluation Measure

# Exceptions in Evaluation Measure

“core”, “basic word order”, “default case”, “unmarked form”

VS.

“periphery”, “lexical listing”, “exceptional marking”, “diacritics”

# Exceptions in Evaluation Measure

“core”, “basic word order”, “default case”, “unmarked form”

VS.

“periphery”, “lexical listing”, “exceptional marking”, “diacritics”

- SPE: “Clearly, we must design our linguistic theory in such a way that the existence of exceptions does not prevent the systematic formulation of those regularities that remain ... Finally, an overriding consideration is that the evaluation measure must be designed in such a way that the wider and more varied the class of exceptions to a rule, the less highly valued is the grammar” (p172)

# Exceptions in Evaluation Measure

“core”, “basic word order”, “default case”, “unmarked form”

VS.

“periphery”, “lexical listing”, “exceptional marking”, “diacritics”

- SPE: “Clearly, we must design our linguistic theory in such a way that the existence of exceptions does not prevent the systematic formulation of those regularities that remain ... Finally, an overriding consideration is that the evaluation measure must be designed in such a way that the wider and more varied the class of exceptions to a rule, the less highly valued is the grammar” (p172)
- But majority doesn’t rule: 90% of English words in speech are stress initial (Cutler & Carter 1987); Legate & Yang poster

# Measuring Rules

# Measuring Rules

# Measuring Rules

- Optimization again: instead of space, it's time



# Measuring Rules

- Optimization again: instead of space, it's time
- Exceptions = Real time processing slowdown
  - “kicked the bucket” faster than “lifted the bucket” by **51ms** (Swinney & Cutler 1979)
  - Production: German irregular past participle (-n) faster than regular (-t) by **38ms** (Clahsen & Fleischhauer 2011)
  - Lexical decision: English irregular verbs faster than regulars by **19ms** (English Lexicon Project; Lignos 2011)

# Measuring Rules

- Optimization again: instead of space, it's time
- Exceptions = Real time processing slowdown
  - “kicked the bucket” faster than “lifted the bucket” by **51ms** (Swinney & Cutler 1979)
  - Production: German irregular past participle (-n) faster than regular (-t) by **38ms** (Clahsen & Fleischhauer 2011)
  - Lexical decision: English irregular verbs faster than regulars by **19ms** (English Lexicon Project; Lignos 2011)
- Exceptions delay rule computation

# Measuring Rules

- Exception 1
- Exception 2
- Exception 3
- ...
- Rule

- Optimization again: instead of space, it's time
- Exceptions = Real time processing slowdown
  - “kicked the bucket” faster than “lifted the bucket” by **51ms** (Swinney & Cutler 1979)
  - Production: German irregular past participle (-n) faster than regular (-t) by **38ms** (Clahsen & Fleischhauer 2011)
  - Lexical decision: English irregular verbs faster than regulars by **19ms** (English Lexicon Project; Lignos 2011)
- Exceptions delay rule computation

# Tolerance Principle

# Tolerance Principle

To be productive, the maximum exceptions to a rule/process applicable to **N** items is

$$\frac{N}{\ln N}$$

# Tolerance Principle

To be productive, the maximum exceptions to a rule/process applicable to  $N$  items is

$$\frac{N}{\ln N}$$

- If English has **150** irregular verbs, we need **900** regulars to have a productive -ed rule:  **$1050/\ln(1050) = 150$**

# Tolerance Principle

To be productive, the maximum exceptions to a rule/process applicable to  $N$  items is

$$\frac{N}{\ln N}$$

- If English has **150** irregular verbs, we need **900** regulars to have a productive -ed rule:  **$1050/\ln(1050) = 150$**
- Children start over-regularization when they reach the tipping point

# Tolerance Principle

To be productive, the maximum exceptions to a rule/process applicable to  $N$  items is

$$\frac{N}{\ln N}$$

- If English has **150** irregular verbs, we need **900** regulars to have a productive -ed rule:  **$1050/\ln(1050) = 150$**
- Children start over-regularization when they reach the tipping point
- $N$  (e.g., vocabulary size) and the number of exceptions may vary from speaker to speaker, accounting for certain individual patterns in language acquisition and sociolinguistic variation



# A Birth-er Problem

# A Birth-er Problem

- The suffix **-er** is productive and segmented in real time even for **broth-er** (Rastle, Davis & New 2004) resulting in slowdown (Lignos 2011)

# A Birth-er Problem

- The suffix **-er** is productive and segmented in real time even for **broth-er** (Rastle, Davis & New 2004) resulting in slowdown (Lignos 2011)
- While some **-er**'s are real (**hunt-hunter**), some are not (**corn-corner, cent-center, sock-soccer**): children need to learn **-er** despite exceptions

# A Birth-er Problem

- The suffix **-er** is productive and segmented in real time even for **broth-er** (Rastle, Davis & New 2004) resulting in slowdown (Lignos 2011)
- While some **-er**'s are real (**hunt-hunter**), some are not (**corn-corner, cent-center, sock-soccer**): children need to learn **-er** despite exceptions
- English Lexicon Project (Balota et al. 2007)
  - **hunt-hunter** type: 94, **cent-center** type: 18
  - The suffix **-er** is productive:  $18 < 112/\ln(112)=24$

# A Birth-er Problem

- The suffix **-er** is productive and segmented in real time even for **broth-er** (Rastle, Davis & New 2004) resulting in slowdown (Lignos 2011)
- While some **-er**'s are real (**hunt-hunter**), some are not (**corn-corner, cent-center, sock-soccer**): children need to learn **-er** despite exceptions
- English Lexicon Project (Balota et al. 2007)
  - **hunt-hunter** type: 94, **cent-center** type: 18
  - The suffix **-er** is productive:  $18 < 112/\ln(112)=24$
- The suffix **-th** fails to reach productivity: **warmth, width, depth** etc. overwhelmed by **tooth, booth, filth, forth, ...**

stride-strode-???

$$\frac{N}{\ln N}$$

stride-strode-???

$$\frac{N}{\ln N}$$

- Mere majority is not sufficient; filibuster proof majority required

# stride-strode-???

$$\frac{N}{\ln N}$$

- Mere majority is not sufficient; filibuster proof majority required
- [-lexical insertion] (Halle 1972, esp. fn1)
  - gaps only arise in unproductive corners of morphology



# stride-strode-???

$$\frac{N}{\ln N}$$

- Mere majority is not sufficient; filibuster proof majority required
- [**-lexical insertion**] (Halle 1972, esp. fn1)
  - gaps only arise in unproductive corners of morphology
- 102 out of 161 irregular verbs (**36%**) show preterite and past participle syncretism
  - Tolerance Principle only allows  $1/\ln(161)=20\%$  exceptions
  - \*forwent, \*sightsaw, \*stridden

# Evaluation Metrics

# Evaluation Metrics

- What not to do: Computer chess

# Evaluation Metrics

- What not to do: Computer chess
- Resource bounded optimization

# Evaluation Metrics

- What not to do: Computer chess
- Resource bounded optimization
- Convergence of methods and disciplines

# Evaluation Metrics

- What not to do: Computer chess
- Resource bounded optimization
- Convergence of methods and disciplines
- Simple theories are usually right ones

# Thank you, to my teachers

- Bob Berwick
- Noam Chomsky
- Morris Halle