Quantifying language learnability: The simplicity hypothesis

Anne Hsu
supervised by Nick Chater
University College London

Outline

- Background on learning language without negative evidence
- Simplicity principle and Minimum Description Length hypothesis
- MDL applied to assessing language learnability
- Comparison with experiments

Outline

- Background on learning language without negative evidence
- Simplicity principle and Minimum Description Length hypothesis
- MDL applied to assessing language learnability
- Comparison with experiments

No negative evidence in 1st language acquisition

- Children don't receive/pay attention to negative feedback from parents (Brown and Hanon 1970, Hirsh-Pasek, Treiman, and Schneiderman, 1984; Demetras, Post, and Snow, 1986; Penner, 1987; Bohannon & Stanowicz, 1988; Marcus 1999).

Problem of language acquisition:

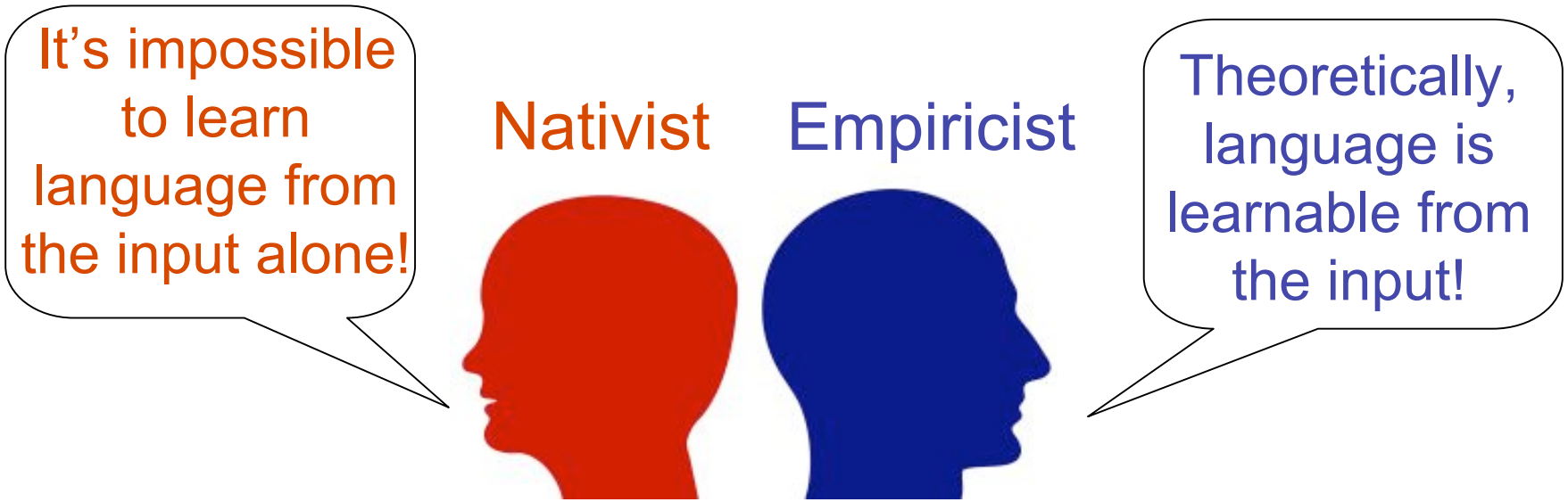
Many sentences we say we've never heard before: Language requires "generalization"

Yet linguistic rules abound with exceptions:

John asked Mary a question
**John shouted Mary a question*
John gave Mary sheets
**John donated Mary sheets*
Betty splashed the floor with suds
**Betty spilled the floor with suds*
Betty wrapped the pole with ribbons
**Betty coiled the pole with ribbons*
Betty painted flowers onto the wall

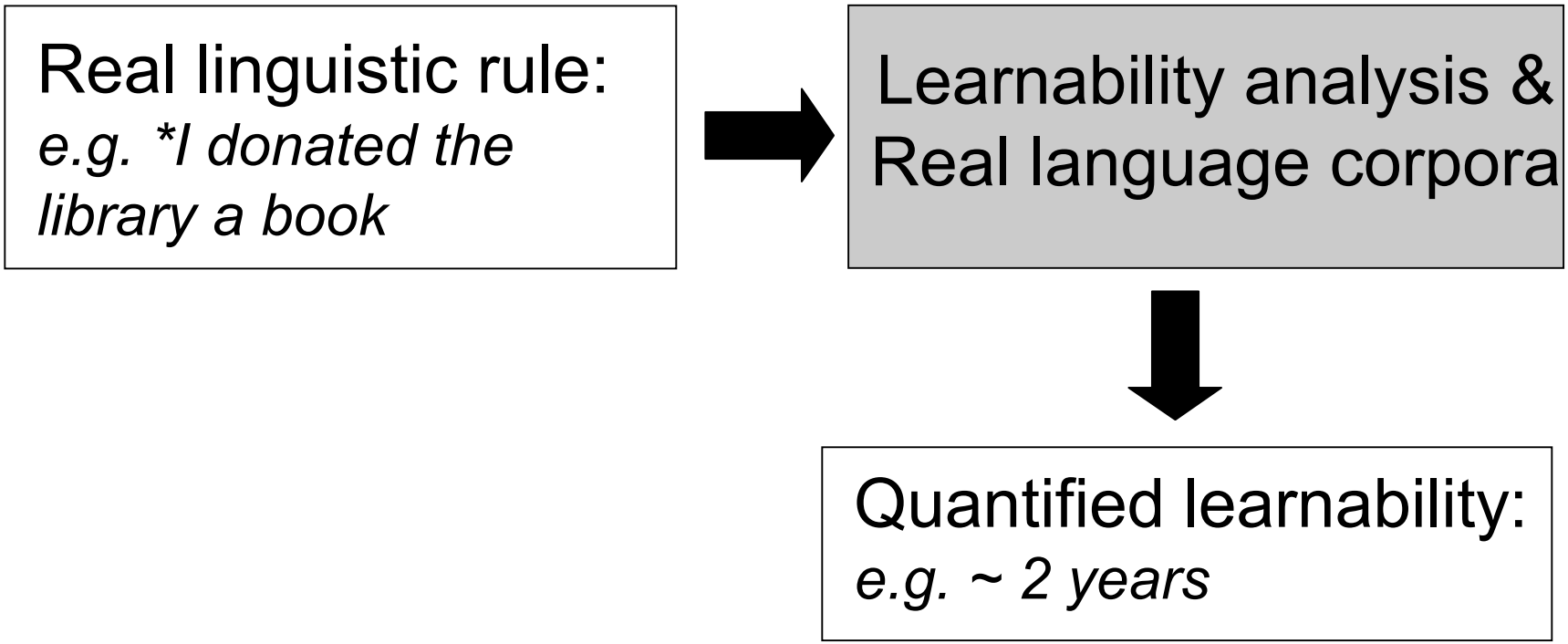
How do we learn what's grammatical and what's not?

Two extremes of the language acquisition debate:



Exactly how learnable are specific language constructions?

We need a method for assessing learnability of specific constructions



Outline

- Background on learning language without negative evidence
- Simplicity principle and Minimum Description Length hypothesis
- MDL applied to assessing language learnability
- Comparison with experiments

Simplicity implemented through coding theory: 2 part Minimum Description Length (MDL)

- Goal: to find regularity in the data.
- Regularity means `ability to compress'.

Contains
1) hypothesis: probability model of the data
2) representation of data given the hypothesis.
   Code length=-log(p(data))

Data: 001001001001…
Hypothesis : endless repetition of 001 p(001)=1
Code length of 001 given hypothesis : -log(1)=0 bits

Data: 0001010000…
Hypothesis: Bernoulli with p(0)=0.8
Code length of 0 and 1 given hypothesis : -log(.8)=0.3 and -log(.2)=2.3 bits

Two-part version of MDL

Goal: minimize the sum L(H) + L(D|H),

L(H) is the length, of the hypothesis; (grammar description)

L(D|H) is length of data representation under the hypothesis encoded sentences under the grammar

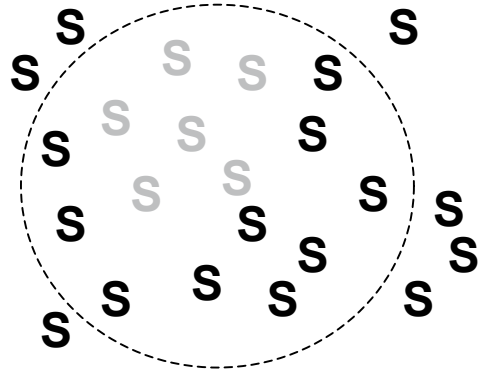
= *L(hypothesis)*

= *L(data)*$_1$

S = grammatical sentences

S = ungrammatical sentences

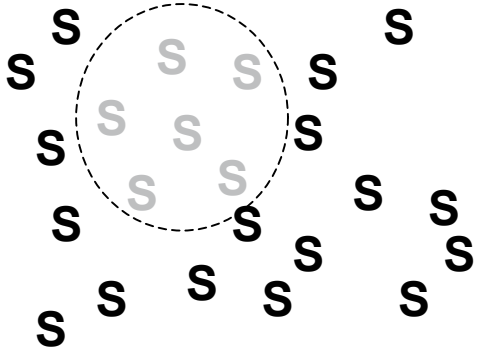
Simpler over-general grammar



L(hypothesis) is short:

L(data)$_1$ is long:

More complex, specific grammar



L(hypothesis) is long:

L(data)$_1$ is short:

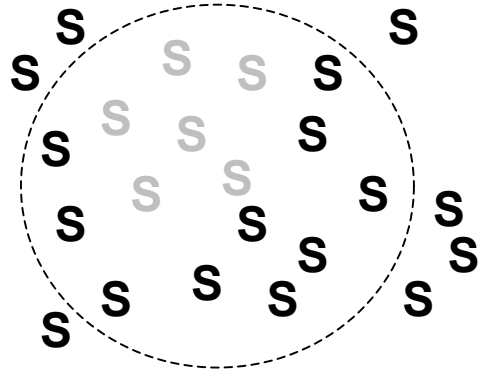
= *L(hypothesis)*

= *L(data)*$_1$

s = grammatical sentences

S = ungrammatical sentences

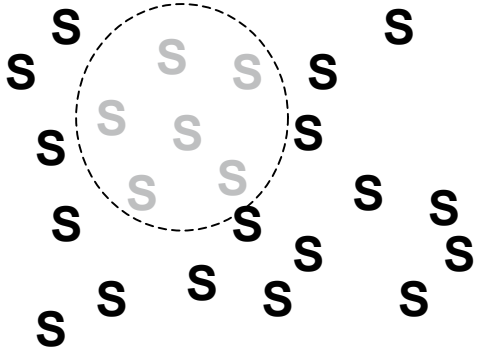
Simpler over-general grammar



L(hypothesis) is short:

L(data)$_1$ is long:

More complex, specific grammar



L(hypothesis) is long:

L(data)$_1$ is short:

Less data:

Simple:

Complex:

More data:

Simple:

Complex:

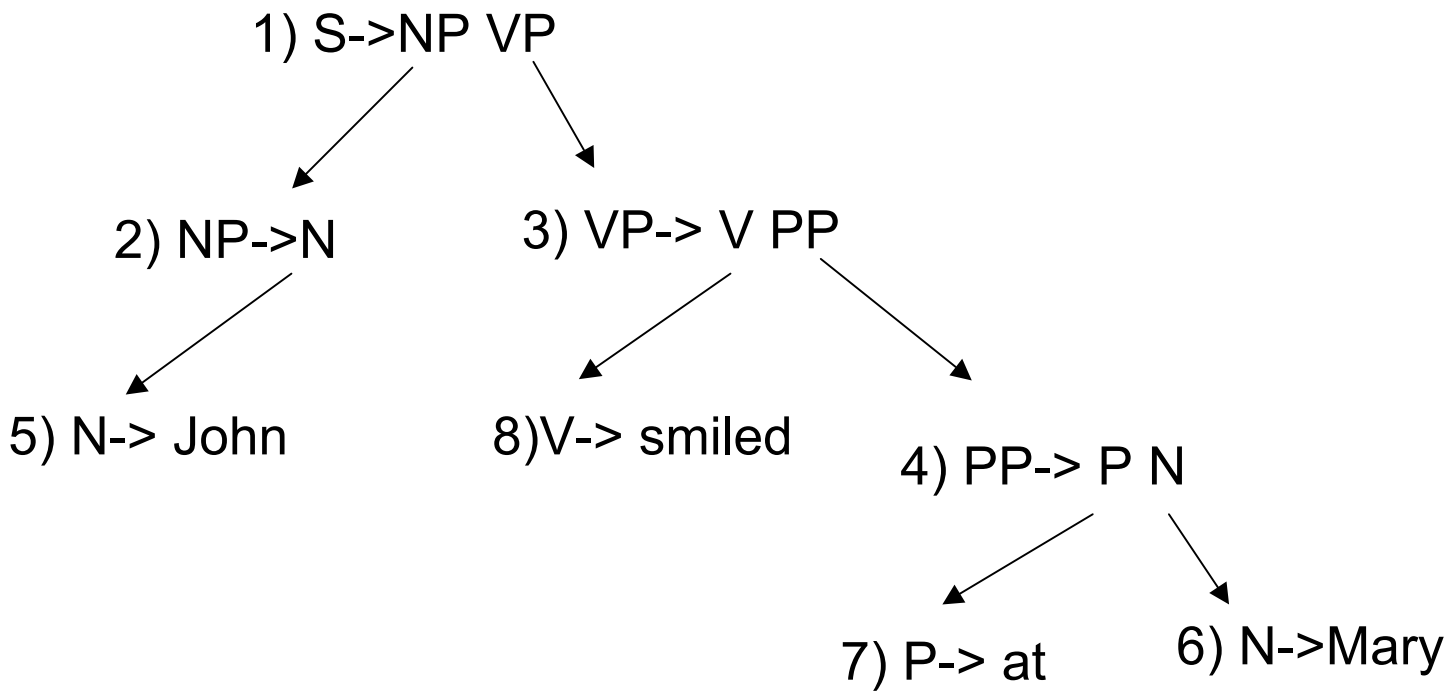
With less data, over all encoding length is shorter with simpler grammar. With more data, it is shorter with more complex grammar.

Outline

- Background on learning language without negative evidence in language
- Simplicity principle and Minimum Description Length hypothesis
- MDL applied to assessing language learnability
- Comparison with experiments

Language encoded using phrase structure grammar

1) S->NP VP #
2) NP->N #
3) VP-> V PP #
4) PP-> P N #
5) N-> John #
6) N->Mary #
7) P-> at #
8) V-> smiled #



1, 2, 3, 4, 5, 8, 7, 6 = John smiled at Mary

The data

Ethel thinks John ran
John thinks Ethel ran
Mary ran
Ethel hit Mary
Mary thinks John hit Ethel
John screamed
Noam hopes John screamed
Mary hopes Ethel hit John
Noam kicked Mary

John hit Mary
Mary hit Ethel
Ethel ran
John ran
Mary ran
Ethel hit John
Noam hit John
Ethel screamed
Mary kicked Ethel
John hopes Ethel thinks Mary hit Ethel

The initial grammar

$S \rightarrow X\ S$ $\quad$ $S \rightarrow X$

$X \rightarrow$ John $\quad$ $X \rightarrow$ Ethel

$X \rightarrow$ Mary $\quad$ $X \rightarrow$ Noam

$X \rightarrow$ ran $\quad$ $X \rightarrow$ screamed

$X \rightarrow$ hit $\quad$ $X \rightarrow$ kicked

$X \rightarrow$ thinks $\quad$ $X \rightarrow$ hopes

The learned grammar

$V_s \rightarrow$ thinks
$V_s \rightarrow$ hopes
NP $\rightarrow$ John
NP $\rightarrow$ Ethel
NP $\rightarrow$ Mary
NP $\rightarrow$ Noam

S $\rightarrow$ NP VP
VP $\rightarrow$ ran
VP $\rightarrow$ screamed
VP $\rightarrow V_t$ NP
VP $\rightarrow V_s$ S
$V_t \rightarrow$ hit
$V_t \rightarrow$ kicked

Dowman

- In order to assess learnability we need to apply MDL analysis to **natural language corpora**.

- MDL has been used in natural language to show learnability of particular linguistic constructions:

-anaphoric one (Foraker, Regier, Khetarpal, Perfors, & Tenenbaum, 2009)

-hierarchical phrase structure (Perfors, Regier, & Tenenbaum, 2006)

We present a general method for assessing learnability of any given linguistic construction given two assumptions:

1) Choice of grammar representation and rule description

2) Choice of input corpus

Testing MDL in real language

Instead of conducting full learning over all possible grammars, we will compare specific models and evaluate the relative gains in compression obtained from coding specific exceptions:

Cost= L(new grammar) - L(original grammar)

Gain= ΔL(exception|H)*frequency(exception)

Construction is learnable when Gain = Cost.

KEY: Only need to specify the part that differs between the two grammars.

Sample old and new grammars: *dative alternation*

e.g. I gave the money to her / I donated the money to her
I gave her the money / * I donated her the money

Original grammar:

```
[case definition give/donate]
    [direct-dative]  V->V' NP NP #
    [prepositional-dative] V->V' NP PP #
    [dative give/donate ]  give donate #
[end]

[case] [dative give/donate]
     [direct-dative] #     0.9
    [prepositional-dative]  #    0.1
[end]
```

New grammar:

```
[case definition donate/give]
[direct-dative]  V->V' NP NP #
    [prepositional-dative] V->V' NP PP
    [both datives verb1] verb1
    [prepositional-dative-only verb2] verb2
[end]

[case] [both-datives give]
    [direct-dative] #     0.9
    [prepositional-dative]  #    0.1
 [end]
[case] [prepositional-dative-only donate]
    [prepositional-dative]  #   1.0
[end]
```

Sample old and new grammars: *dative alternation*

e.g. I gave the money to her / * I donated the money to her
I gave her the money / * I donated her the money

Original grammar:

```
[case definition give/donate]
    [direct-dative]  V->V' NP NP #
    [prepositional-dative] V->V' NP PP #
    [dative verb1/verb2 ] verb1 verb2 #
[end]

[case] [dative give/donate]
     [direct-dative] #     0.9
    [prepositional-dative]  #    0.1
[end]
```

New grammar:

```
[case definition donate/give]
[direct-dative]  V->V' NP NP #
    [prepositional-dative] V->V' NP PP
    [both datives verb1] verb1
    [prepositional-dative-only verb2] verb2
[end]

[case] [both-datives give]
    [direct-dative] #     0.9
    [prepositional-dative]  #    0.1
 [end]
[case] [prepositional-dative-only donate]
    [prepositional-dative]  #   1.0
[end]
```

Grammar cost difference: 53 bits

MDL for model selection in real data:

Original grammar
Donate and Give

P(prep)=0.1-> 2.3 bits
P(direct)=0.9-> 0.1 bits

.

After 20 encounters of donate,
Data cost = 56 bits

Grammar cost= original cost

New grammar
Give:
P(prep)=0.1-> 2.3 bits
P(direct)=0.9-> 0.1 bits

Donate:
P(prep)=1 -> 0 bits

After 20 encounters of donate,
Data cost = 0 bits

Grammar cost=original cost + 53 bits

Learnability: Savings on Data cost >= grammar cost

An ideal learner should acquire dative restriction on *donate* after seeing *donate* 20 times.

Assessing learnability of specific constructions using MDL:

1) Specify original vs. new specific-rule grammar and evaluate grammar cost difference
2) Evaluate data cost savings between original and new grammars
3) Evaluate how many occurrences of a construction is needed for the new grammar to be worth "learning" (data savings >= grammar cost)
4) Use corpus to evaluate how often a construction occurs in real language

$$O_{1yr} = \frac{T_{year}}{T_{corpus}} \times O_{corpus}$$

5) Years needed to learn ~ # occurrences needed/ # occurrences per year

PREDICTIONS:

1) More learnable constructions are learned more quickly/easily

2) More learnable constructions will have grammatical and ungrammatical forms perceived as more extremely grammatical/ungrammatical.

Outline

- Background on learning language without negative evidence in language
- Simplicity principle and Minimum Description Length hypothesis
- MDL applied to assessing language learnability
- Comparison with experiments

Comparison with Data 1:

Theakston (2004) asked 5 and 8 year old children to assess grammaticality of ungrammatical sentences.

I told the idea to her. / I told her the idea.
I (whispered, shouted) the idea to her. / *I (whispered, shouted) her the idea.

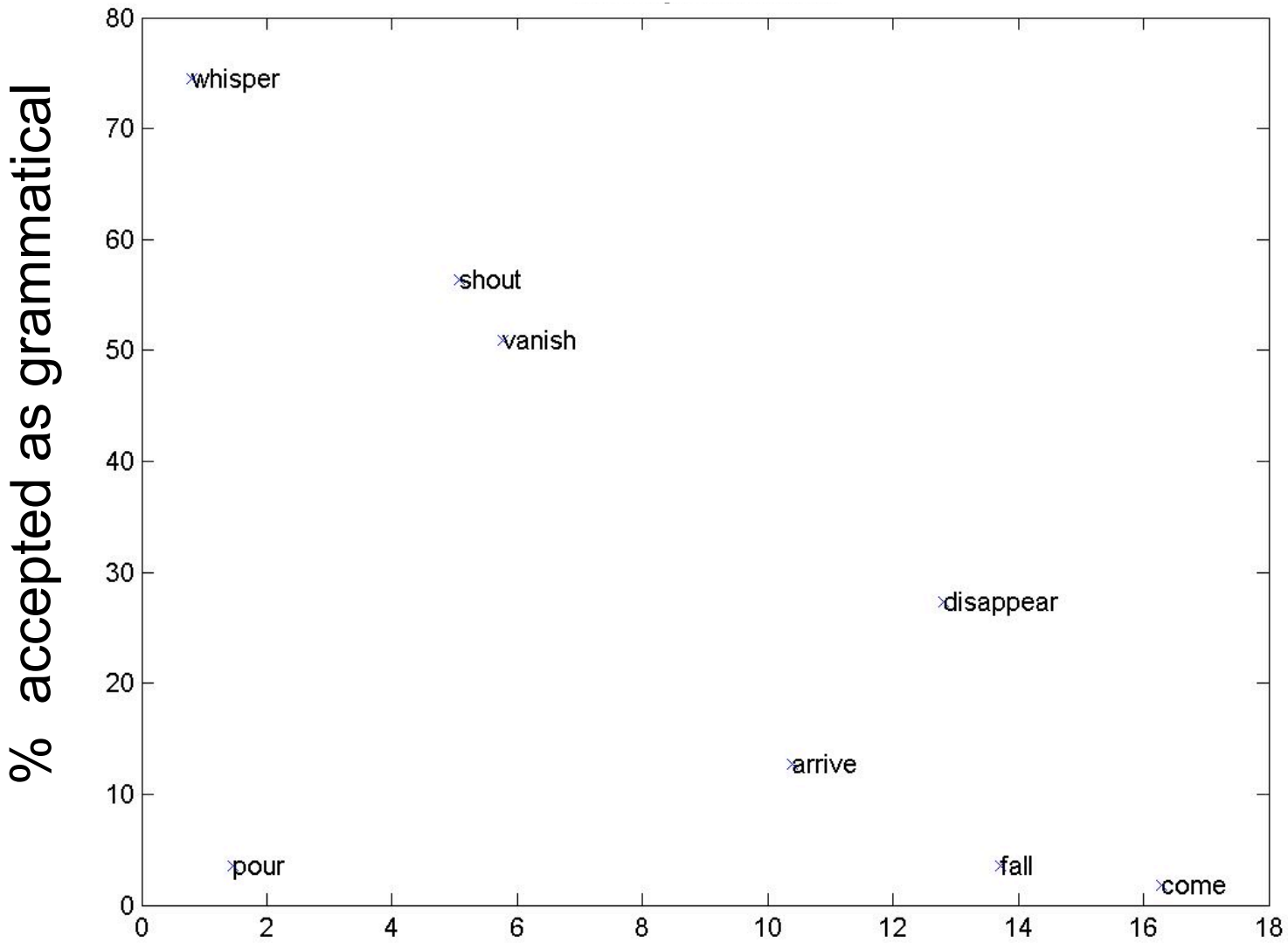
I loaded pebbles into the tank. / I loaded the tank with pebbles.
I poured pebbles into the tank. / *I poured the tank with pebbles.

John hid. / John hid the rabbit.
John (disappeared, vanished). /*John (disappeared,vanished) the rabbit.

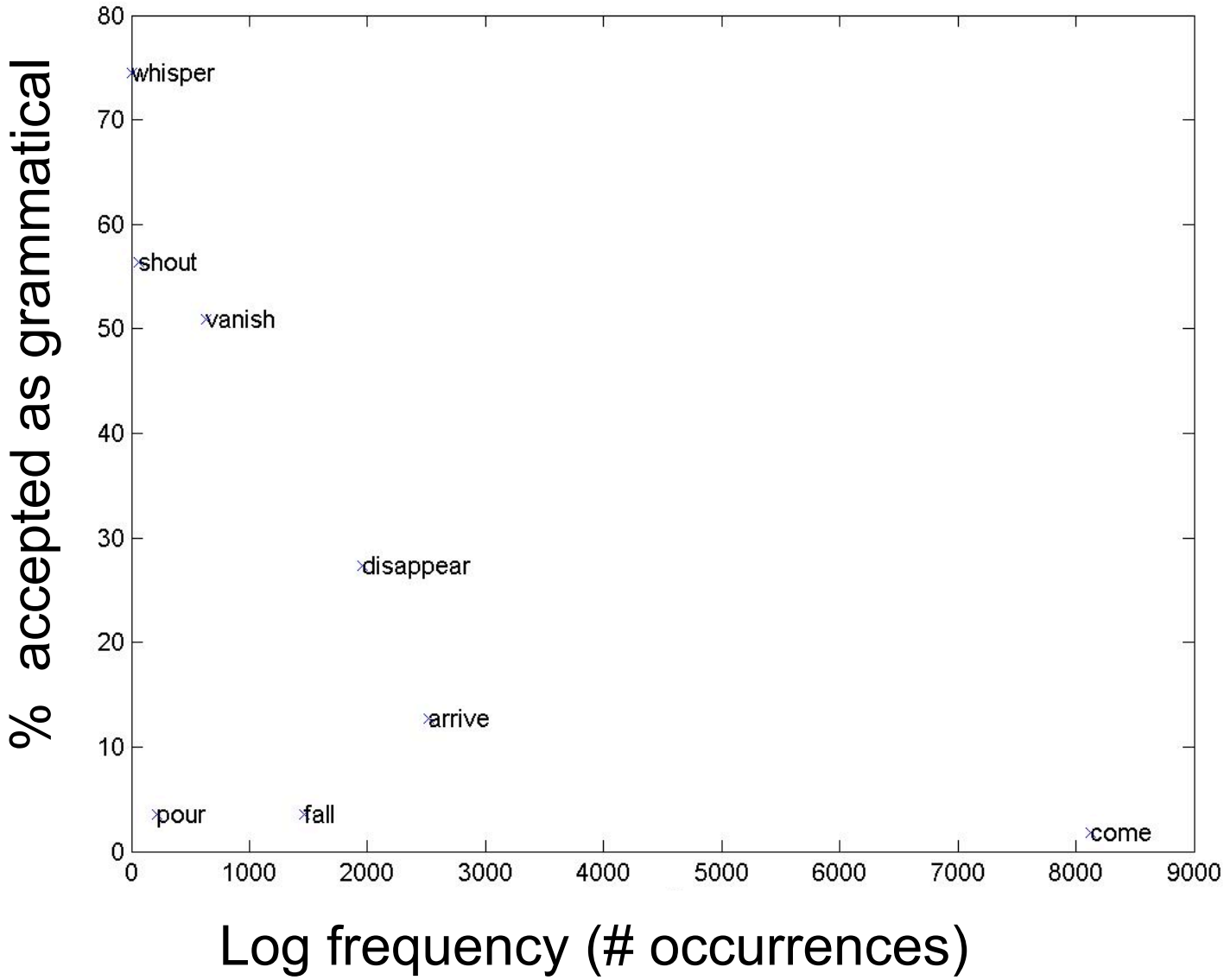
The plane landed. / He landed the plane.
The plane (came, arrived). / * He (came, arrived) the plane.

It dropped. / Somebody dropped it.
It fell. / *Somebody fell it.

MDL analysis with Theakston data (British National Corpus)



Frequency counts with Theakston data (British National Corpus)

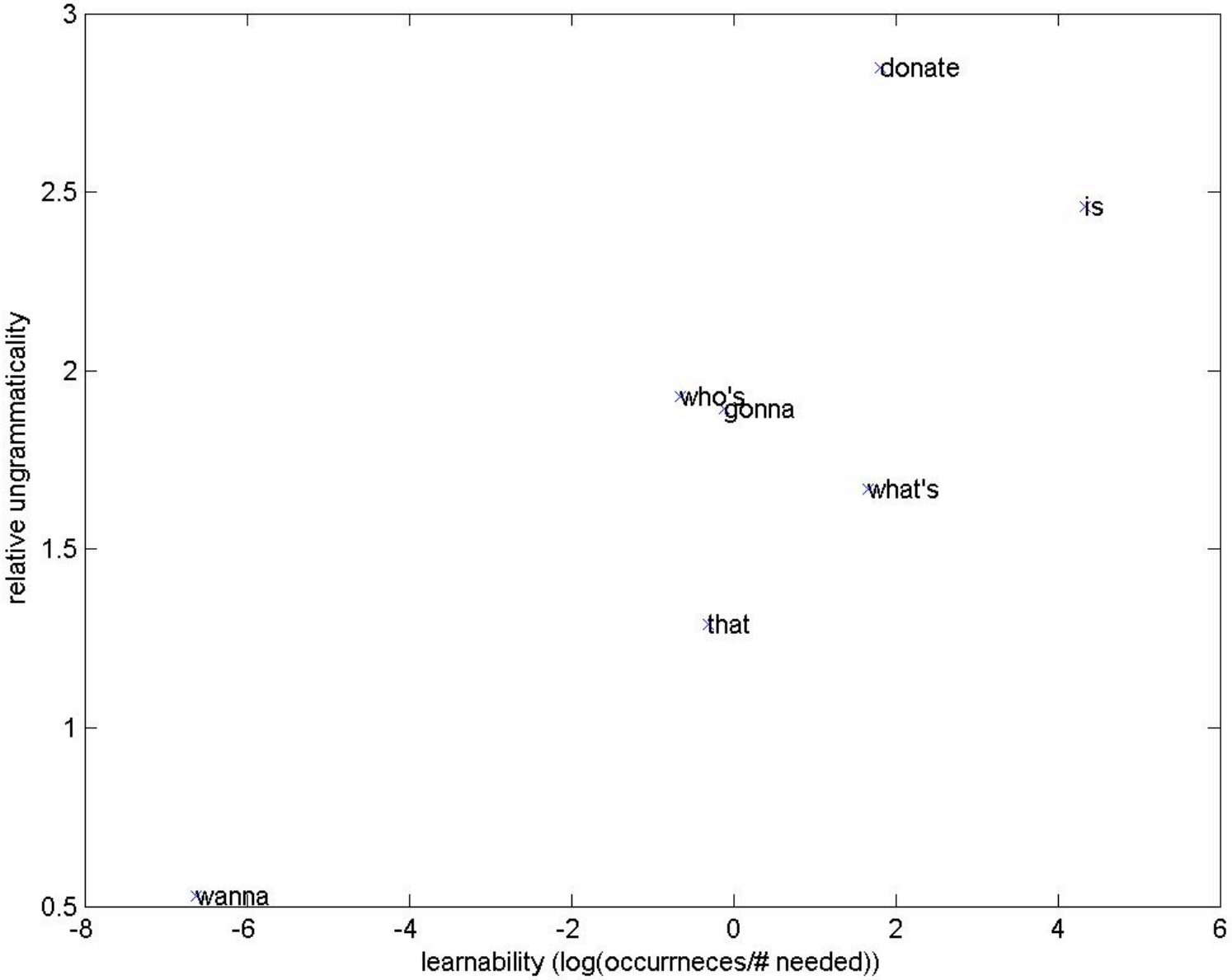


Comparison with Data 2:

Internet experiment: ages 7-70 (mean 31). Rate grammaticality 1-5

Who do you think mom called? / Who do you think that mom called?
Who do you think called mom? / *Who do you think that called mom?
Which team do you want to beat? / Which team do you wanna beat?
Which team do you want to win? / *Which team do you wanna win?
I'm going to help her. / I'm gonna help her.
I'm going to the store. / *I'm gonna the store.
Jane is taller than John. / Jane's taller than John.
Jimmy is shorter than she is. / *Jimmy is shorter than she's.
What is there? / What's there?
What is it?. / *What's it?
Who is here? / Who's here?
Who is it? / *Who's it?
I gave a book to the library. / I gave the library a book.
I donated a book to the library. / *I donated the library a book.

Relative grammar judgment vs. learnability



Summary:

We can use MDL to assess learnability of real language constructions

MDL assessment of learnability seems to be supported by data so far.

Related work

Effects of sampling assumptions on learning from implicit negative evidence

Work done in Collaboration with Nick Chater, University College London

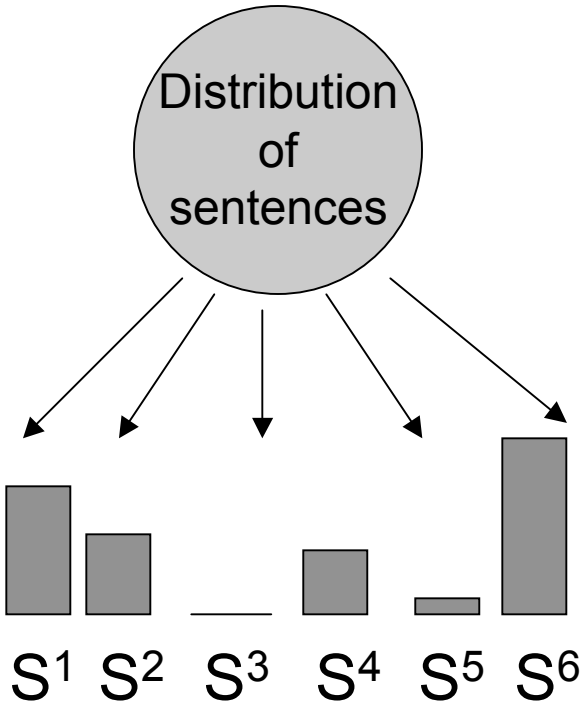
Thank you to Tom Griffiths, UC Berkeley for helpful input and discussion

Language Study 2:

Learning comparison

Probabilities

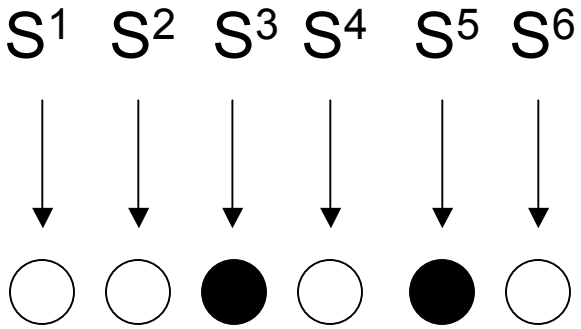
P(S|Grammatical)



vs.

Rules

P(Grammatical|S)



Language Study 2:

The artificial language:

S1) *Verb Subject Object*

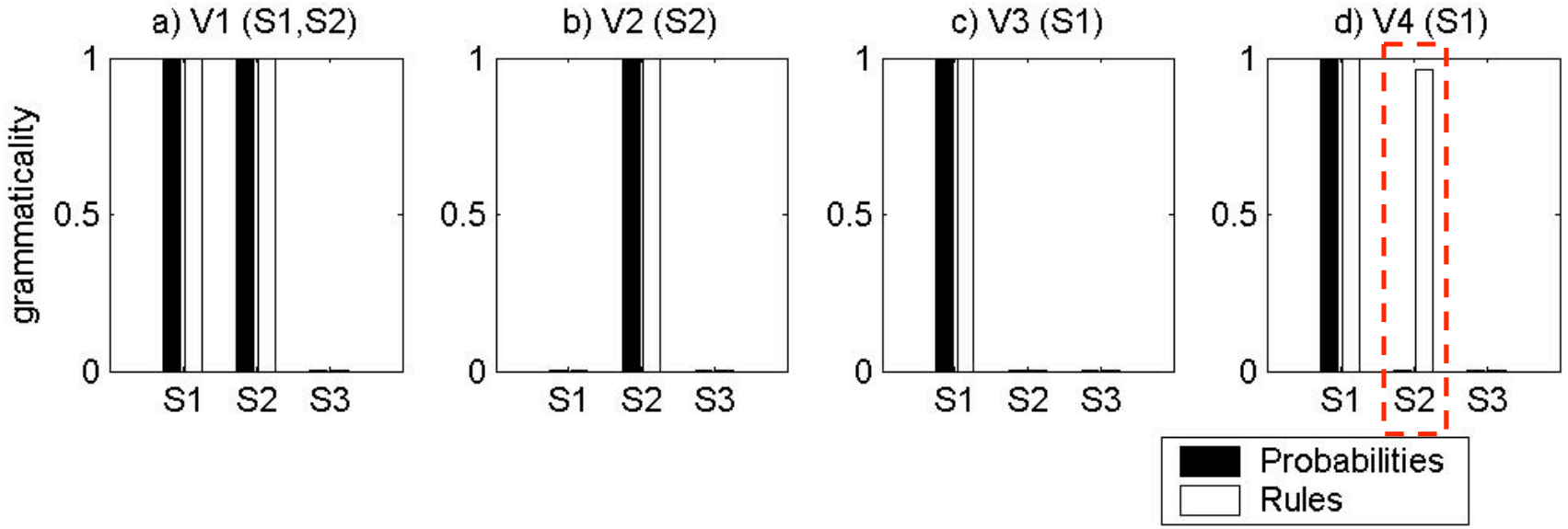
S2) *Subject Verb Object*

S3) *Subject Object Verb*

	S1	S2	S3
V1	+ (9)	+ (9)	- (6)
V2	- (3)	+(18)	- (3)
V3	+ (18)	- (3)	- (3)
V4*	+ (18)	- (0)	- (6)

Language Study 2:

Model Predictions



Language Study 2:

Language Study 2:

Language Study 2:

Language Study 2:

Language Study 2:

blergen tombat flern

Language Study 2:

Language Study 2:

Language Study 2:

Language Study 2:

Language Study 2:

Language Study 2:

nagid blergen semz

Language Study 2:

Condition 1: Probabilities learning

Always grammatically correct adult

Always grammatically incorrect child

Language Study 2:

Condition 1: Probabilities learning

scene 4/84

blergen norg nagid

blergen

nagid

tombat

Click here for next scene

Language Study 2:

Condition 1: Probabilities learning

Language Study 2:

Condition 2: Rules learning

Language Study 2:

Condition 2: Rules learning

Language Study 2:

Condition 2: Rules learning

scene 4/84

blergen semz tombat

Ungrammatical

Was that sentence grammatical?

Grammatical Ungrammatical

Sorry you were wrong.

blergen

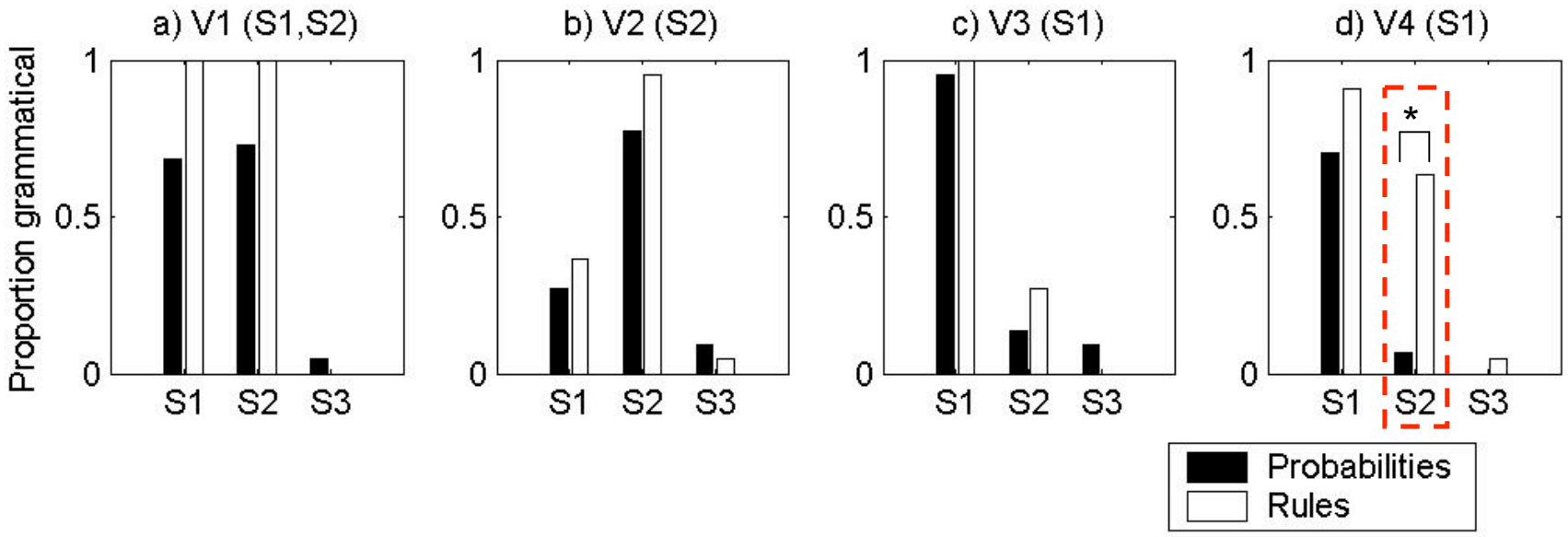
nagid

tombat

Click here for next scene

Language Study 2:

Experimental results



* $\chi 2(1) = 7.28$, $p = 0.007$