Plato's problem
LING 611 Spring 2022

Brian Dillon
Shota Momma (Negishi)

University of Massachusetts, Amherst
Department of Linguistics

3/30/2022

Poverty of the stimulus

A speaker of a language has observed a certain limited set of utterances in his language. On the basis of this finite linguistic experience he can produce an indefinite number of new utterances which are immediately acceptable to other members of his speech community. He can also distinguish a certain set of 'grammatical' utterances, among utterances that he has never heard and might never produce. He thus projects his past linguistic experience to include certain new strings while excluding others.

(Chomsky 1955/1975: p. 61 of 1975 version)

Pullum and Scholz (2002) identify five steps to a poverty argument:

i) that speakers acquire some aspect of grammatical representation;

ii) that the data the child is exposed to is consistent with multiple representations;

iii) that there is data that could be defined that would distinguish the true representation from the alternatives;

iv) that that data does not exist in the primary linguistic data;

v) conclusion: the aspect of the grammatical representation acquired in (i) is not determined by experience but by properties internal to the learner.

Classic case - auxiliary inversion

<u>Aux inversion rule:</u>

The bird that can swim can fly. ->

Can the bird that can swim fly?
* Can the bird that swim can fly?

The girl can see the boy who can swim. ->

Can the girl see the boy who can swim?
* Can the girl can see the boy who swim?

Move the auxiliary of ***the main clause***

Classic case - auxiliary inversion

Test sentences (Subject relatives)

a. The dog that is sleeping is on the blue bench.
b. The ball that the girl is sitting on is big.
c. The boy who is watching Mickey Mouse is happy.
d. The boy who is unhappy is watching Mickey Mouse.
e. The boy who is being kissed by his mother is happy.
f. The boy who was holding the plate is crying.

Mean age: 4;7

	TOTAL	TYPE I	TYPE II	TYPE III
Group I	50 (62%)	30 (60%)	10 (20%)	0
Group II	17 (20%)	9 (53%)	5 (29%)	0
Total	67 (40%)	39 (58%)	15 (22%)	0

TABLE 3. Types of errors by group.

Type I error: *Is the boy that is watching Mickey Mouse is happy?
Type II error: *Is the boy that is watching Mickey Mouse, is he happy?
Type II error: *Is the boy that watching Mickey Mouse is happy?

Crain & Nakayama, 1987

Reconstruction

(1) a. Norbert remembered that Ellen painted a picture of herself
b. * Norbert remembered that Ellen painted a picture of himself
c. Norbert remembered that Ellen was very proud of herself
d. * Norbert remembered that Ellen was very proud of himself

(2) a. Norbert remembered which picture of herself Ellen painted
b. Norbert remembered which picture of himself Ellen painted
c. Norbert remembered how proud of herself Ellen was
d. * Norbert remembered how proud of himself Ellen was

No single example of a *wh*-phrase containing a reflexive pronoun, a non-reflexive pronoun or a name in 10000 wh-sentences in CHILDES

Leddon & Lidz (2006), example
Taken from *faculty of language* blog

Principle C

Truth-value judgement task

First, the Ninja Turtle ate pizza while dancing. This makes the interpretation in which the pronoun (*he*) and the referring expression (*the Ninja Turtle*) are coreferentially true. Second, there was an additional salient character who did not eat pizza while the Ninja Turtle danced. This aspect of the story makes the interpretation in which the pronoun refers to a character not named in the test sentence false. Thus, if children allow coreference in these sentences, they should accept them as true, but if children disallow coreference, then they should reject them as false.

a. While he was dancing, the Ninja Turtle ate pizza.

b. He ate pizza while the Ninja Turtle was dancing.

Children as young as 3 years old accepted sentences like (a), but overwhelmingly rejected sentences like (b)

Korean verb (non-)raising

Scope freezing

Nwukwunka-ka manhun salam-ul piphanhay-ss-ta.
someone-NOM many person-ACC criticize-PST-DECL

some > many: “A particular person criticized many.”

many > some: “*For many people, some person or other criticized him.”

Object raising

Toli-ka **maykcwu-lul cal** masi-n-ta.
Toli-NOM beer-ACC well drink-PRES-DECL

“Toli drinks beer well.”

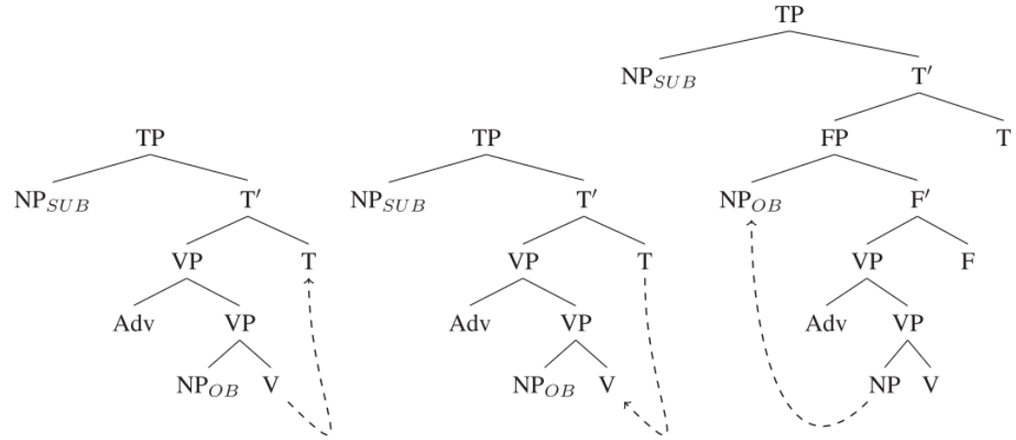


Fig. 2. Korean verb raising (*Left*), tense lowering (*Middle*), and object raising (*Right*).

Short and long negation

Toli-ka maykcwu-lul cal masi-ci **ani ha-n-ta.**
Toli-NOM beer-ACC well drink-CONN NEG do-PRES-DECL

“Toli doesn’t drink beer well.”

Toli-ka maykcwu-lul cal **an masi-n-ta.**
Toli-NOM beer-ACC well NEG drink-PRES-DECL

“Toli doesn’t drink beer well.”

Han et al,. (2016)

Korean verb (non-)raising

Truth-value judgement task

Senario 1:
Subject QP: Two out of three horses (i.e., not all horses) jumped over the fence
Object QP: Two out of three cookies (i.e., not all cookies) were eaten.

Senario 2:
Subject QP: None of the horses jumped over the fence.
Object QP: None of the cookies were eaten.

- **Motun mal-i** wulthali-lul **an** nem-ess-ta.
  every horse-NOM fence-ACC NEG jump.over-PST-DECL

"Every horse didn't jump over the fence."

- Khwuki monste-ka **motun khwuki-lul an** mek-ess-ta.
  cookie monster-NOM every cookie-ACC NEG eat-PST-DECL

"Cookie monster didn't eat every cookie."

Han et al,. (2016)

Korean verb (non-)raising

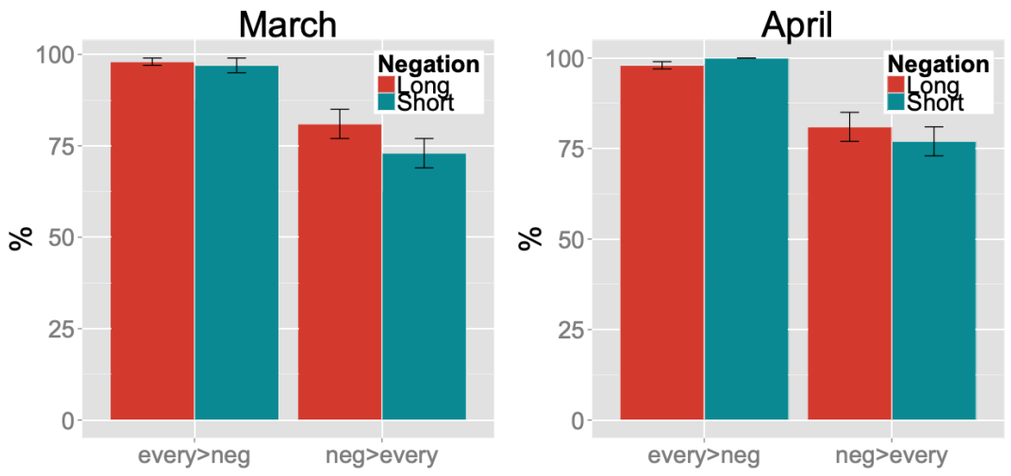


Fig. 3. Mean percentages of acceptances: two test sessions. Error bars indicate one SE from the mean.

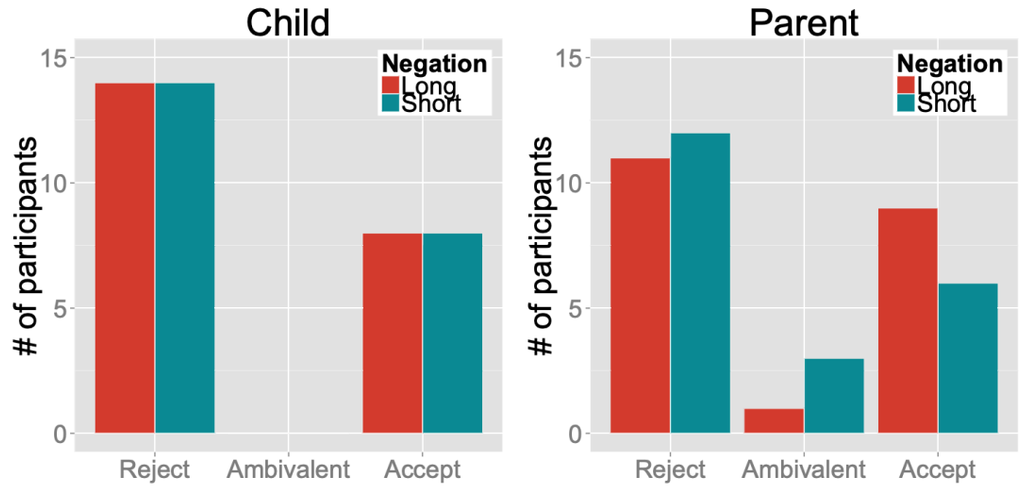


Fig. 7. Number of participants accepting neg > every: children and parents.

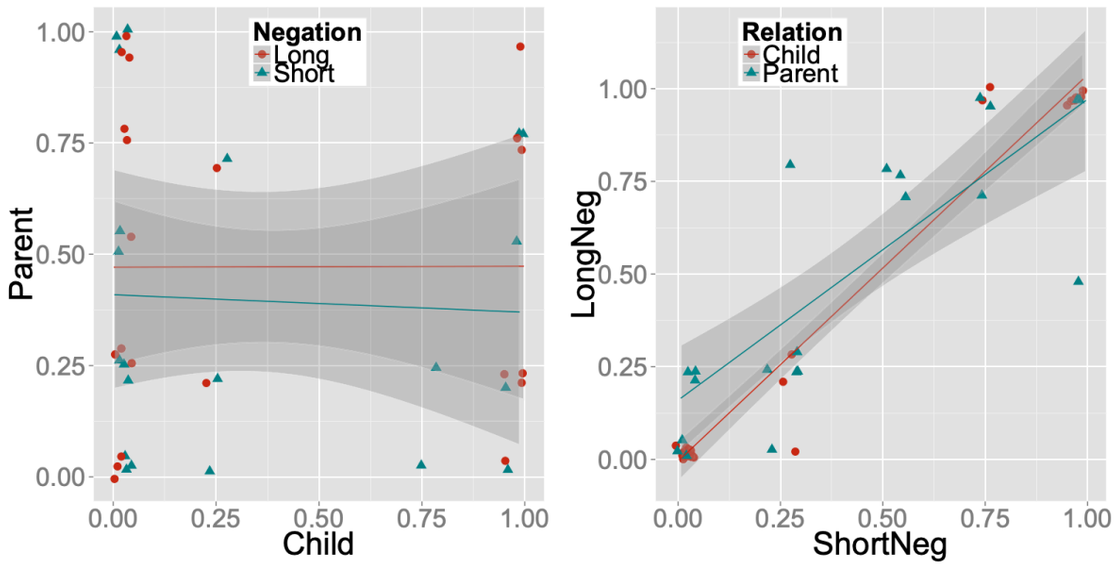


Fig. 8. Correlation between parents and their children's proportions of "yes" responses (*Left*) and correlation between mean acceptance rates of short negation and long negation (*Right*).

Han et al,. (2016)

Islands

Whether-island
*Who did you wonder whether they saw *t*?

Subject-island
* Who did the story about *t* impress John?

Adjunct island
*What did you go home because you needed to do *t*?

Complex NP island
* What did you hear the claim that John met *t*?

- Domain-specific grammatical constraints?
- Non-grammatical constraints (e.g., working memory limitation)

Islands

A factorial design for measuring island effects: STRUCTURE × GAP-POSITION

a.	*Who* __ thinks [that John bought a car]?	NON-ISLAND | MATRIX
b.	*What* do you think [that John bought __]?	NON-ISLAND | EMBEDDED
c.	*Who* __ wonders [whether John bought a car]?	ISLAND | MATRIX
d.	*What* do you wonder [whether John bought __]?	ISLAND | EMBEDDED

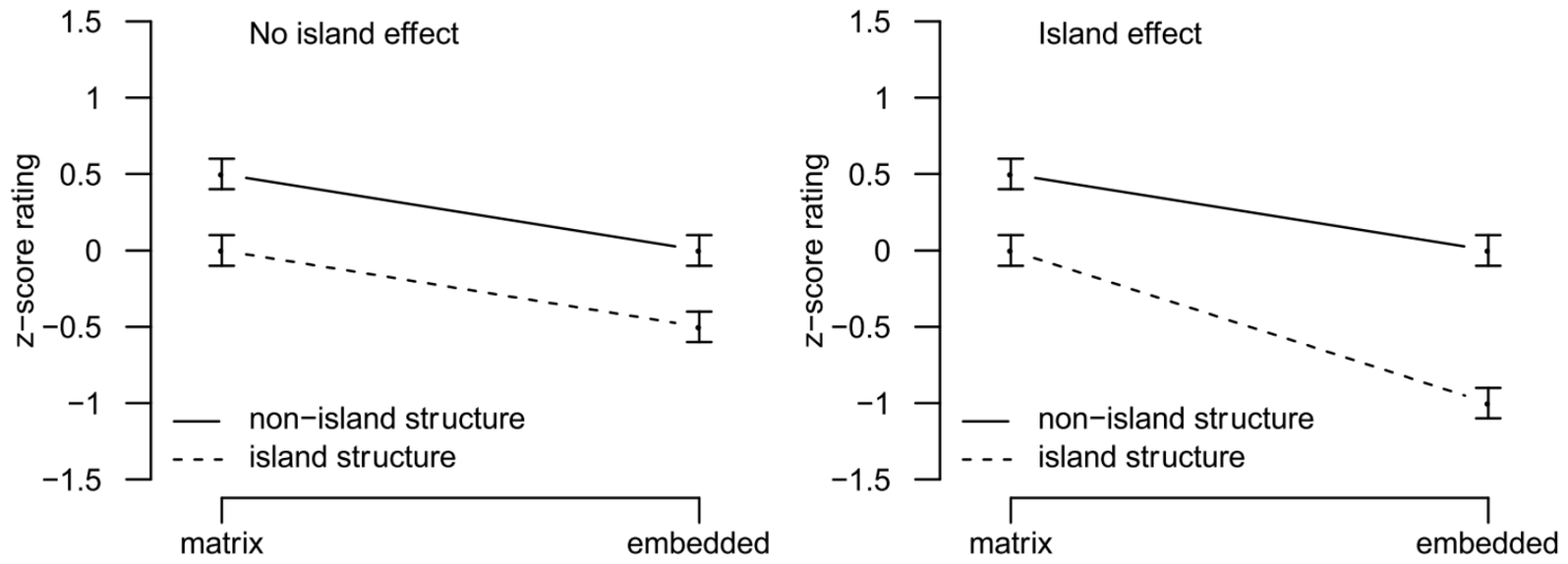


Fig. 1 The *left panel* demonstrates the pattern predicted when no island effect is present. The *right panel* demonstrates the pattern predicted when an island effect is present

Functional explanations of islands

Prediction: working memory capacity should negatively correlate with the magnitude of island effects across individuals.

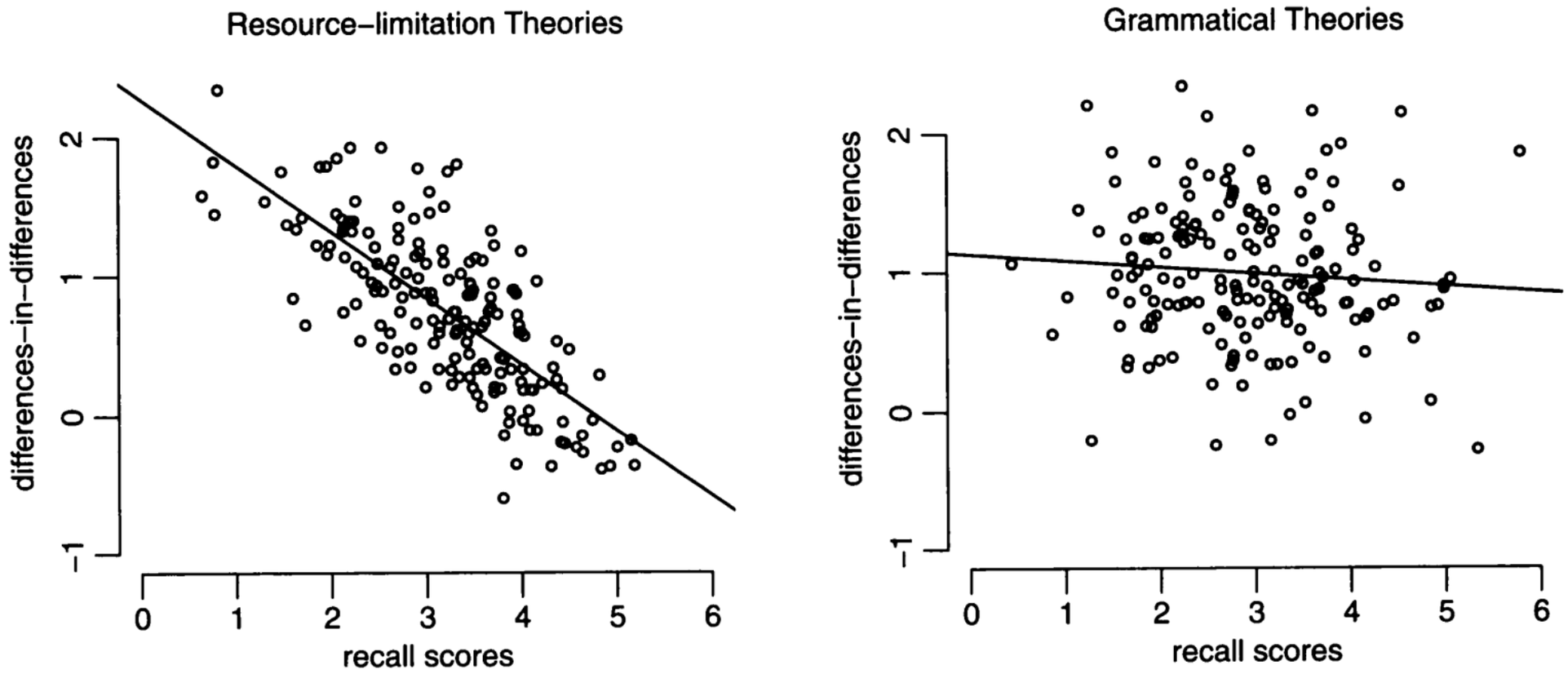
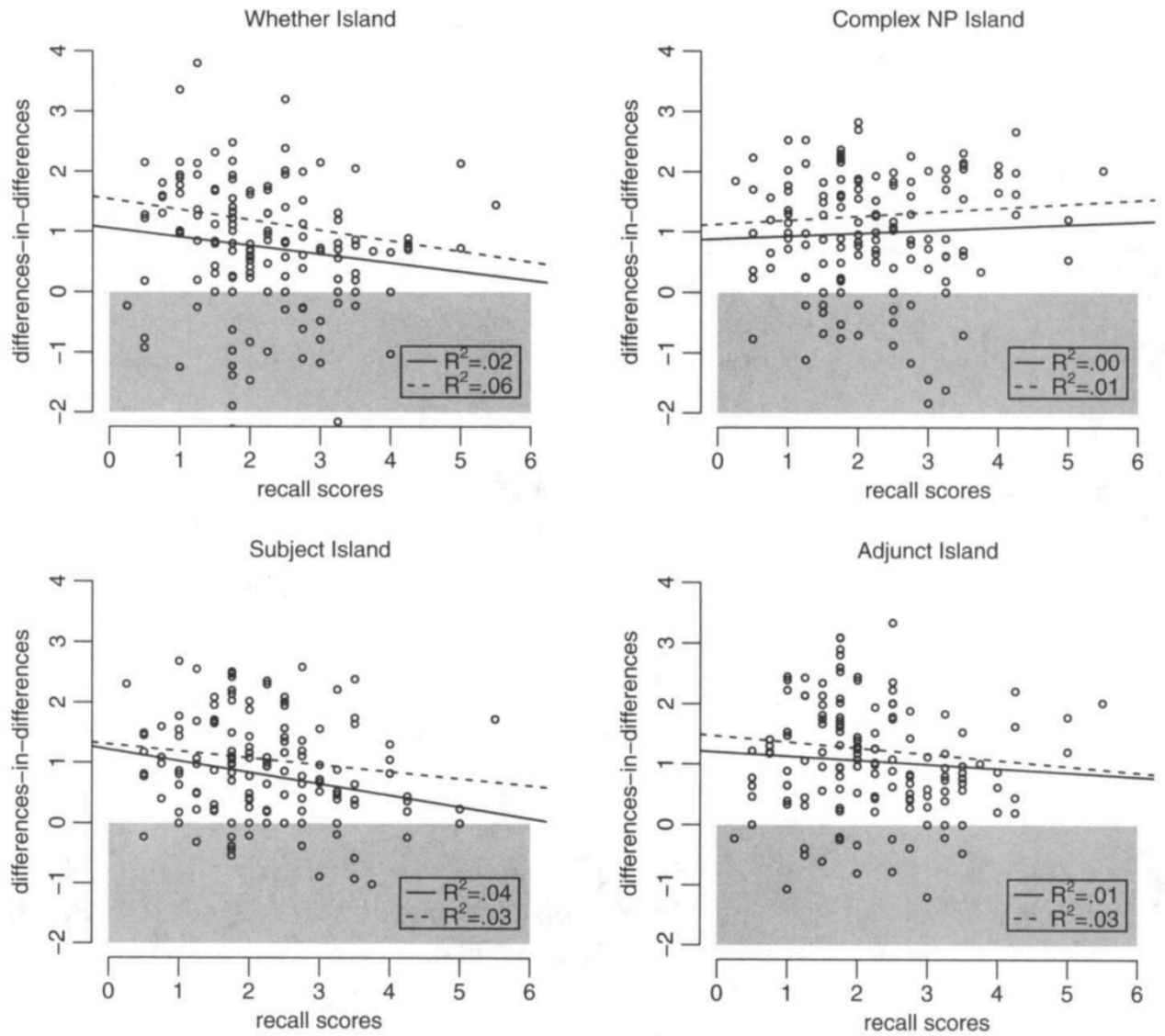


FIGURE 2. Predictions of the resource-limitation (left panel) and grammatical theories (right panel).

Functional explanations of islands

It didn't. But… "absence of evidence ≠ evidence for absence"?



Functional explanations of islands

But… parasitic gap? Why making sentences even more complex improve acceptability?

*What did the attempt to repair ultimately damage the car?”
What did the attempt to repair ultimately damage?

RNN and innateness

Can Recurrent Neural Network (RNN) learn the constraints on anaphoric interpretations?

Network architecture

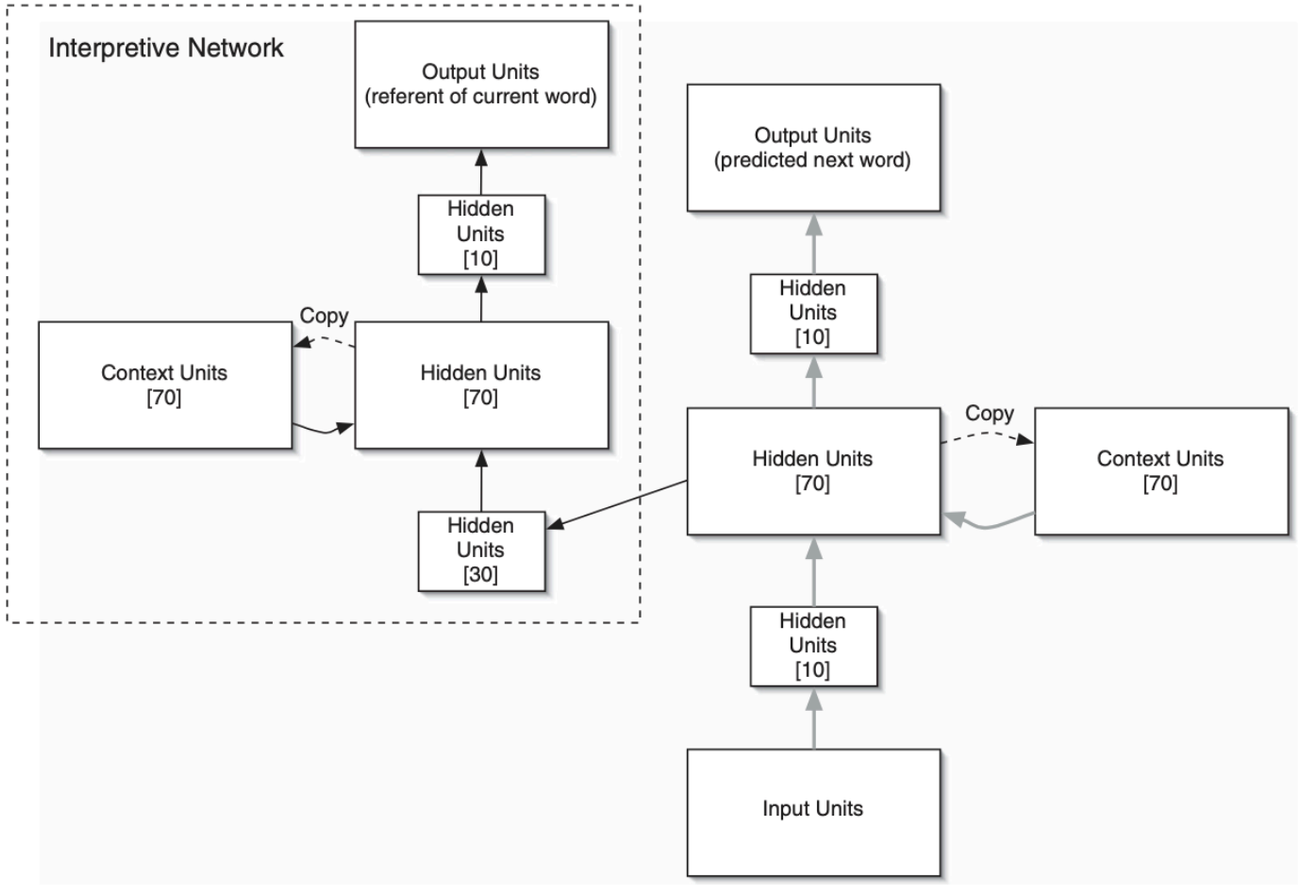


FIGURE 2 Phase 2 network (word prediction + interpretive network).

The word prediction network (outside the dotted box) is trained first. After its weights are frozen, the interpretive network is trained (two-staged training).

Interpretive network was added to interpret pronouns. Its goal is to activate, e.g., ‘John’ when ‘him/himself’ refers to John.

Target language

S	→	NP VP
NP	→	Name (Rel)
VP	→	V NP-obj
NP-obj	→	Name (Rel) | Refl | Pronoun | Distinctive-Obj
Rel	→	who VP | who NP V
Name	→	John | Harold | Nate | Mary | Alice | Sue
Refl	→	himself | herself
Pronoun	→	him | her
Distinctive-Obj	→	junipers | hotdogs | nachos | mangos | avocados | salamanders
V	→	sees-M | loves-M | admires-M | kisses-M | visits-M | sees-F | loves-F | admires-F | kisses-F | visits-F

FIGURE 3 Grammar for the training corpus.

\+

SV agreement (in gender)
Antecedent pronoun/anaphora agreement
Subject - distinctive object ‘agreement’

The co-occurance restriction between names & objects (e.g., *junipers* can only occur when *John* is the subject) -> This forces the prediction network to represent names distinctively.

Example sentences generated by this grammar:

(16) a. Subject verb agreement:
John sees-M Mary; Mary likes-F Bill
b. Subject and object relative clauses:
John who sees-M Sue admires-M Bill; Alice likes-F Nate who Mary admires-F
c. Reflexive and pronominal objects:
John sees-M himself; Alice who Harold likes-M admires-F her

Evaluation

‘Correct’ if the interpretive network shows the highest activation value for the right referent.

Good news: overall ‘good’ performance

Accuracy for reflexive interpretation: 89.7 ± .6%

Accuracy for pronoun interpretation: 71.9 ± 1.5%

Bad news: linearity effect (not human-like)

(17) a. Alice who Mary loves admires herself
herself: $p(Alice) = .99$
b. Alice who loves Mary admires herself
herself: $p(Alice) = .80, p(Mary) = .16, p(Sue) = .03$

Evaluation

Bad news 2: Failing to make abstraction (not human-like)

The activation pattern of the hidden layer of the interpretive network should be identical for ‘grammatically equivalent context.’ (*)

(21) a. Simple Matrix: John admires * himself.
b. Object Relative: John who Bill sees admires * himself.
c. Subject Relative: John who sees Bill admires * himself.

Evaluation

Bad news 2: Failing to make abstraction (not human-like)

> The activation pattern of the hidden layer of the interpretive network should be identical for ‘grammatically equivalent context.’ (*)

(21) a. **Simple Matrix: John admires * himself.**
b. **Object Relative: John who Bill sees admires * himself.**
c. **Subject Relative: John who sees Bill admires * himself.**

But the network treated those context in a construction-specific fashion. That is, the network used information that is not relevant.

The hierarchical clustering of activation pattern

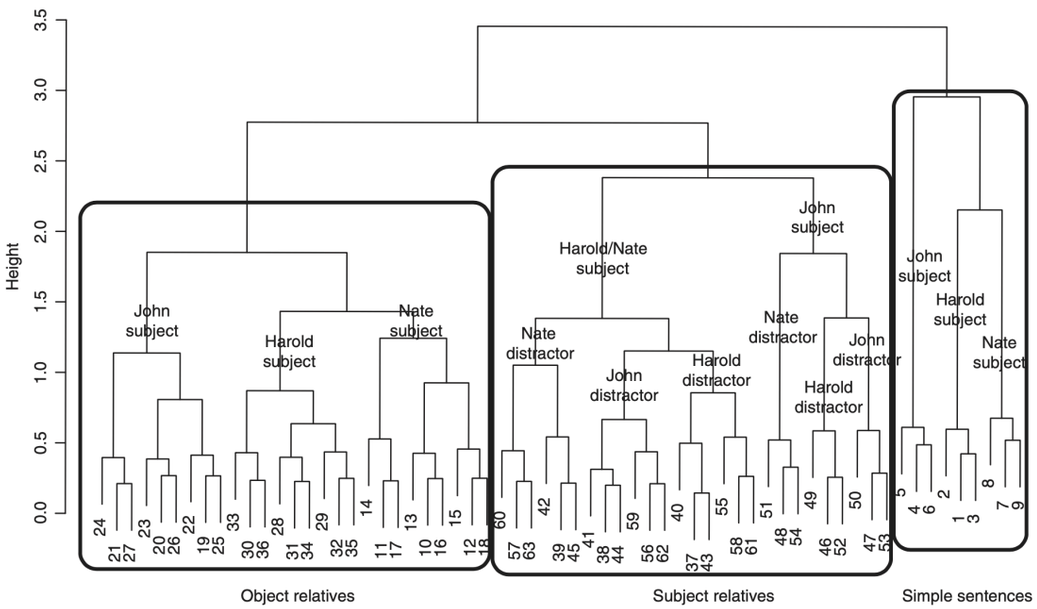


FIGURE 4 Hierarchical Cluster Analysis of network activation immediately preceding the reflexive.

The “lesion” experiment

“for every sentence pair we examined, we were able to find at least one hidden unit whose removal increased the error on the object relative sentence but not on the paired simple matrix sentence”

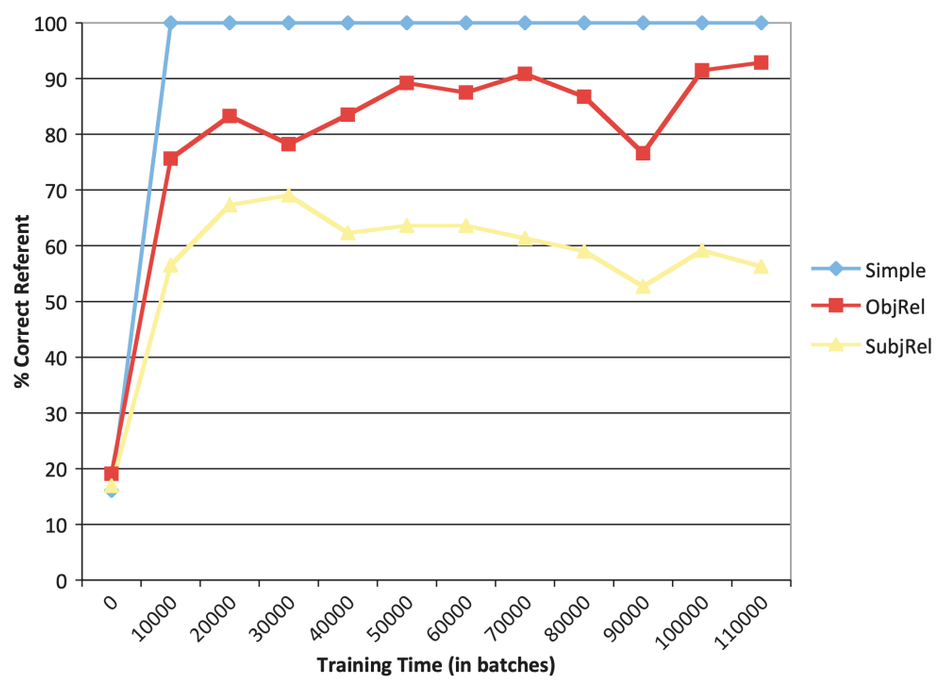
Evaluation

Bad news 2: Failing to make abstraction

If the subject relative clause containing sentences are withheld from the training data (for the interpretive network), the network doesn't do well on sentences containing subject relative clauses (89.7% for sentences that do NOT contain SRC vs. 60.5% for sentences that contain SRC).

Ditto for object relative clauses. (90.3% for sentences that do NOT contain ORC vs. 57.0% for sentences that contain ORC)

… throughout training (= it's not due to 'overfitting')



Single-phase network

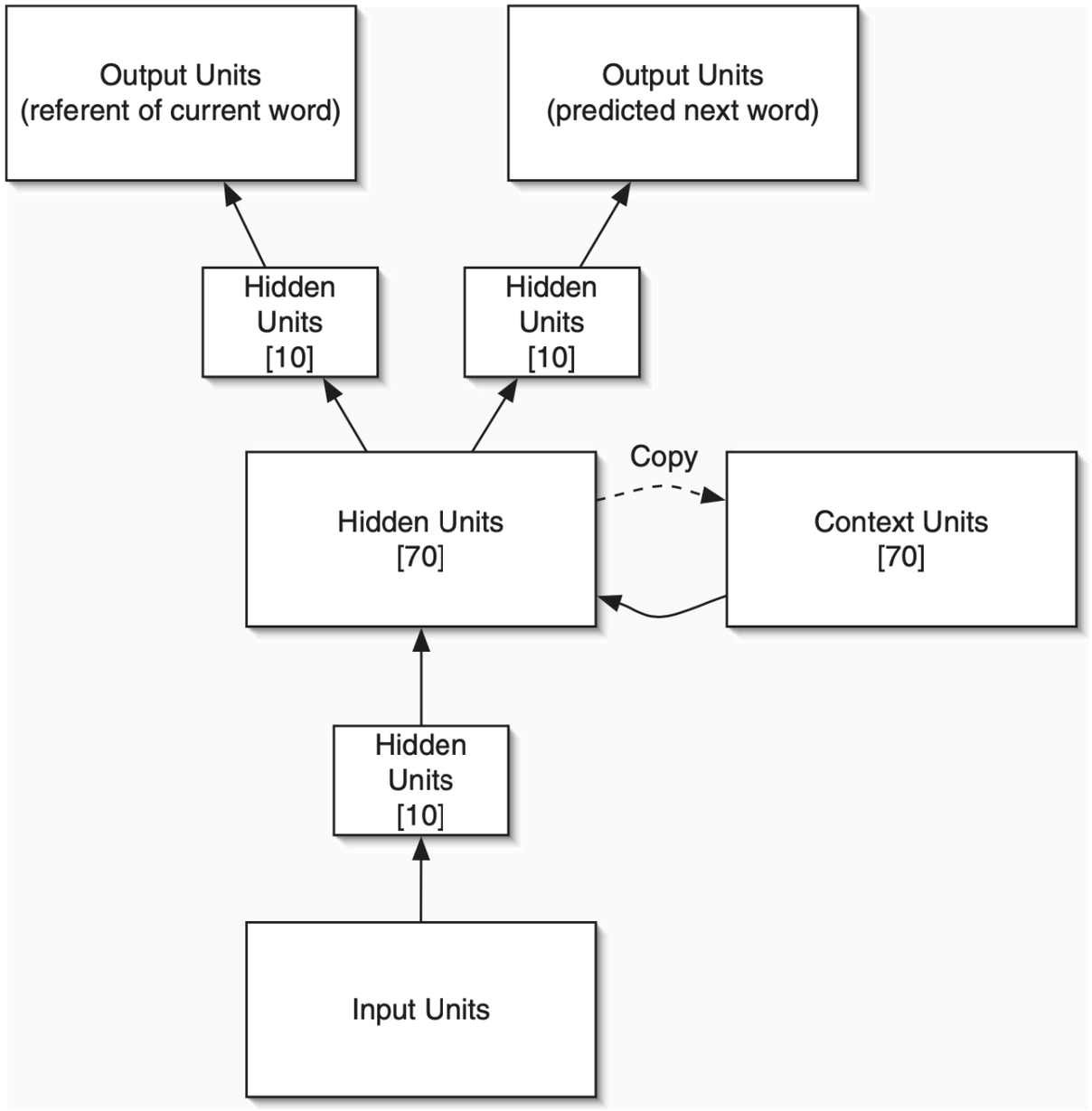


FIGURE 7 Single SRN architecture for anaphoric reference.

Maybe it's unfair to use network trained to predict next words for reference resolution?

This 'single-phase' network directly learns reference resolution (reference resolution does not rely on prediction).

Evaluation

Overall performance is better:

Accuracy for reflexive interpretation: 98.2% (prev. 89.7%)

Accuracy for pronoun interpretation: 81.5% (prev. 71.9%)

But the linearity effect is still present:

TABLE 4
Mean Activation of Linear Distractor as Reflexive Interpretation (Correct Activation = 0)

	ObjRel	*SubjRel*	*F-ratio*	*Prob*
Net A	.009	.035	497.2	<.001
Net B	.017	.042	465.1	<.001
Net C	.017	.084	1063	<.001
Net D	.035	.047	33.34	<.001
Net E	.039	.090	566.2	<.001

TABLE 5
Mean Activation of Subject Referent as Reflexive Interpretation (Correct Activation = 1)

	ObjRel	*SubjRel*	*F-ratio*	*Prob*
Net A	.984	.952	509.8	<.001
Net B	.971	.914	948.9	<.001
Net C	.967	.869	1456	<.001
Net D	.930	.924	3.883	<.05
Net E	.911	.792	1297	<.001

Evaluation

This network did generalize to across structures.

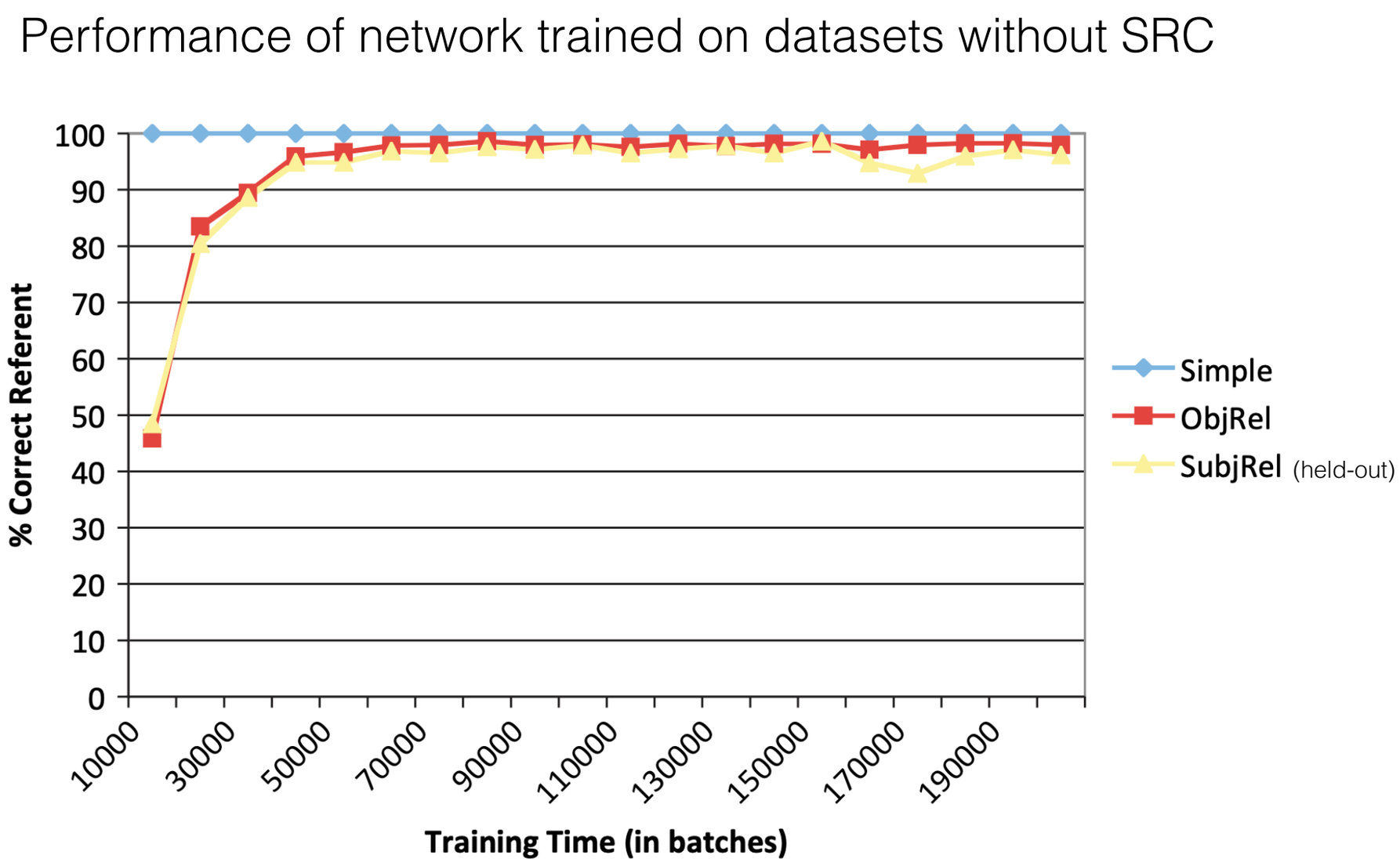


FIGURE 9 Structural generalization in one-phase SRN (color figure available online).

… sort of…

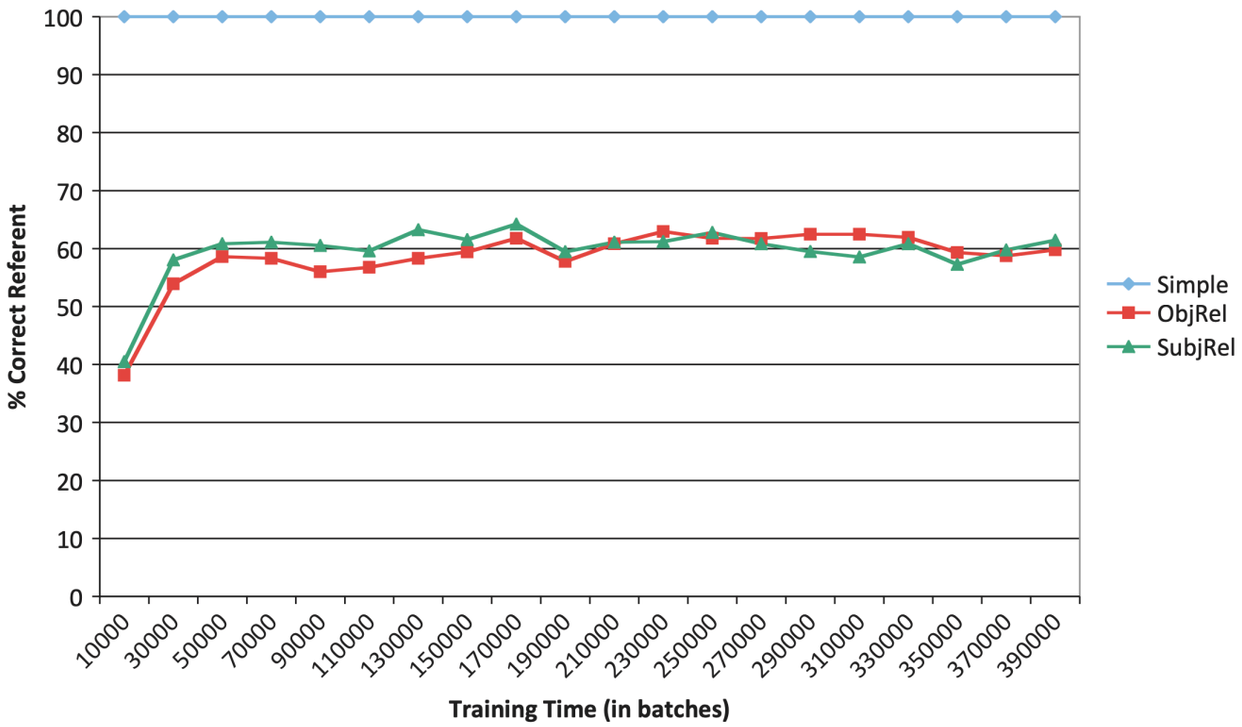


FIGURE 10 Generalization of reflexive interpretation from simple to (held-out) complex sentences in one-phase SRN (color figure available online).

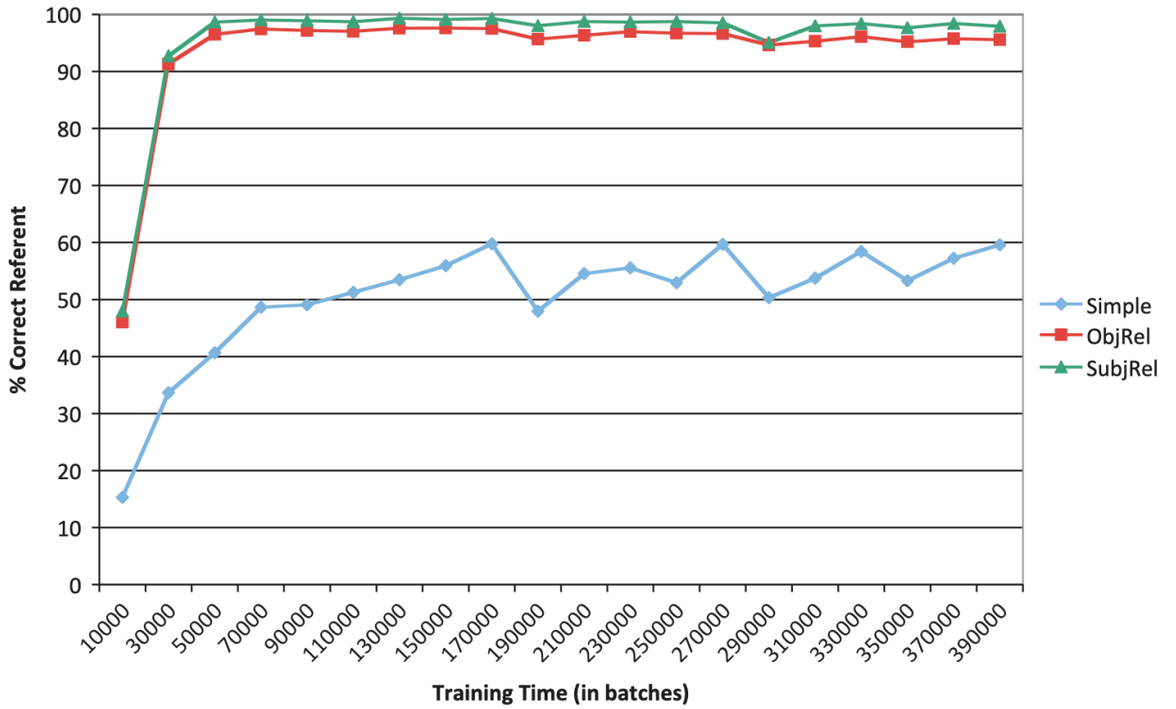
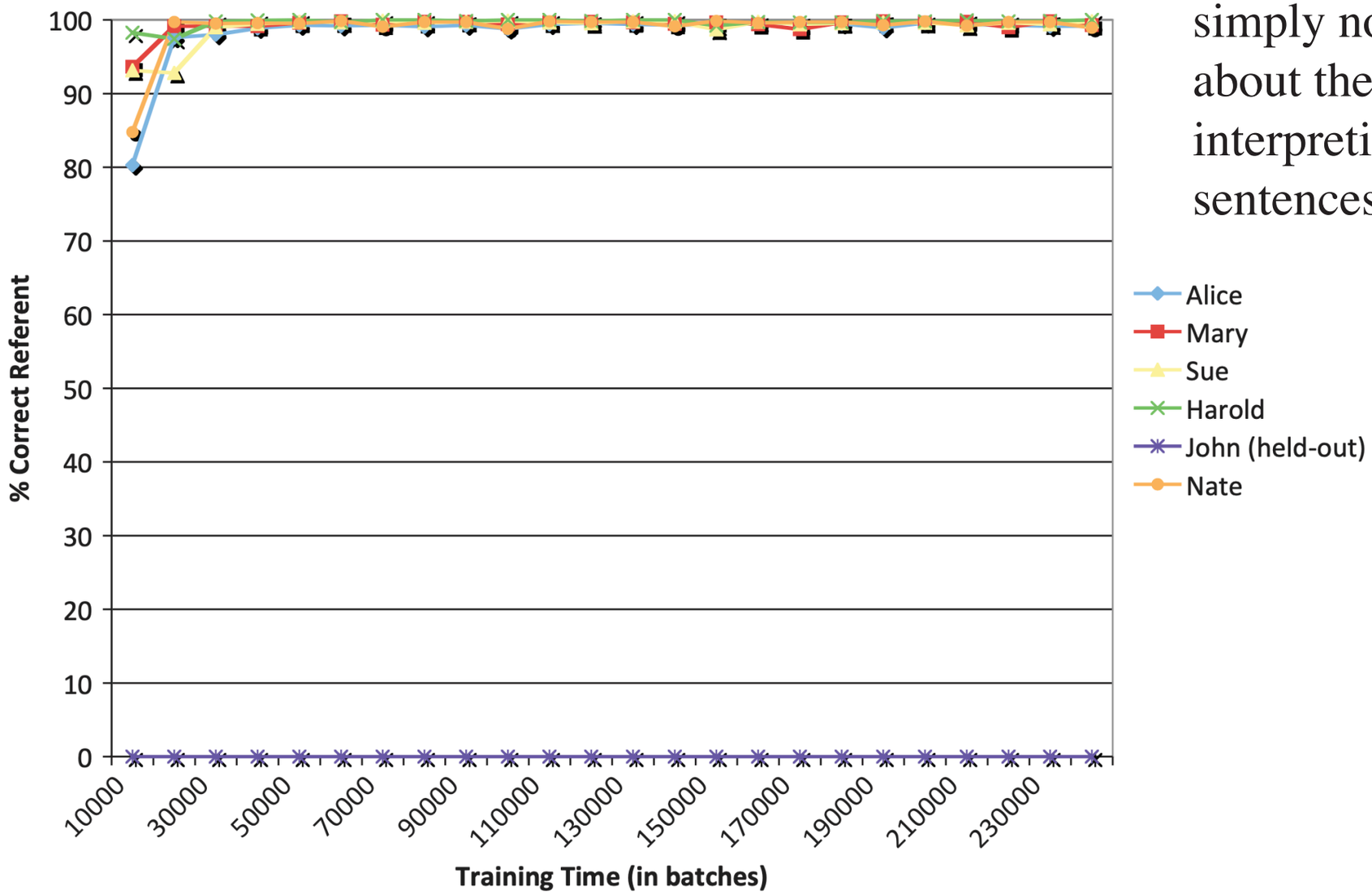


FIGURE 11 Generalization from complex to (held-out) simple sentences in one-phase SRN (color figure available online).

… but not across different subject nouns.



withheld training on the reference for all reflexives whose interpretation was *John*. Note that it is not the case that sentences with *John* as antecedent of a reflexive never occurred during the training of the network. Rather, the networks were simply not given feedback during training about the appropriateness of its interpretive output for reflexives in sentences of this sort.

Summary of the results

The two-phase trained RNN…

showed overall good performance on anaphora resolution

but showed linearity bias (the dropped performance when there is a SRC modifying the subject)

… and failed to generalize across different structures (networks have to be trained on SRC/ORC containing sentences to do well on SRC/ORC containing sentences)

The single-phase trained RNN…

showed even better performance on anaphora resolution

but still showed some linearity bias

did generalize across different structural contexts (networks did fine on SRC containing clauses even when the SRC is withheld in training)

… but did NOT generalize across different subjects (when ‘John ’ as the subject is withheld in the training, it doesn’t do well on sentences with ‘John’ as the subject).

Question

Can simple Recurrent Neural Network (sRNN) learn the constraints on anaphoric interpretations?

- sort of, but it failed to make structural and/or lexical generalizations.