Getting the Right Stuff Wrong: Distributional Equivalence in Modeling Language Acquisition

Sarah Brogden Payne

sarah.payne@stonybrook.edu

paynesa.github.io

Rutgers Subregular Workshop

15th November 2025

Introduction: Returning to the Computational Level

Central Mystery: how do children come to be **competent, fluent speakers** of their native language(s) from such **small, sparse** input?

Introduction: Returning to the Computational Level

Central Mystery: how do children come to be **competent, fluent speakers** of their native language(s) from such **small, sparse** input?

Computational Models: are one tool to explore this question

- Propose **mechanisms** that may underlie acquisition

Introduction: Returning to the Computational Level

Central Mystery: how do children come to be **competent, fluent speakers** of their native language(s) from such **small, sparse** input?

Computational Models: are one tool to explore this question

- Propose **mechanisms** that may underlie acquisition

Another way to view this: **Marr's Levels**

Introduction: Returning to the Computational Level

Central Mystery: how do children come to be **competent, fluent speakers** of their native language(s) from such **small, sparse** input?

Computational Models: are one tool to explore this question

- Propose **mechanisms** that may underlie acquisition

Another way to view this: **Marr's Levels**

- **Computational Level:** what are the goals of the computation? Map from **small, sparse input** to the **learned grammar**

Introduction: Returning to the Computational Level

Central Mystery: how do children come to be **competent, fluent speakers** of their native language(s) from such **small, sparse** input?

Computational Models: are one tool to explore this question

- Propose **mechanisms** that may underlie acquisition

Another way to view this: **Marr's Levels**

- **Computational Level:** what are the goals of the computation? Map from **small, sparse input** to the **learned grammar**
- **Algorithmic Level:** what are the **representations** used for the input and output and what are the **mechanisms** that map between them? **Computational modeling** attempts to address this.

(Marr, 1982)

Marr's Levels: An Example

The Past Tense Debate: Computational Level

Marr's Levels: An Example

The Past Tense Debate: Computational Level

- **Goal:** learn **-ed** and its allomorphs

Marr's Levels: An Example

The Past Tense Debate: Computational Level

- **Goal:** learn **-ed** and its allomorphs
- **Input:** ≤ 500 verbs, not full paradigms

Marr's Levels: An Example

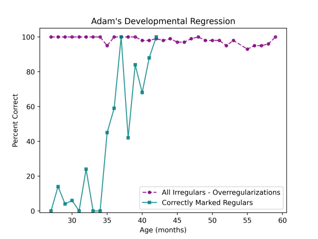
The Past Tense Debate: Computational Level

- **Goal:** learn **-ed** and its allomorphs
- **Input:** ≤ 500 verbs, not full paradigms
- **Output:** learned grammar that
  - Has an **asymmetry** between over-regularization and over-irregularization
  - Shows **developmental regression**

Marr's Levels: An Example

The Past Tense Debate: Computational Level

- **Goal:** learn **-ed** and its allomorphs
- **Input:** ≤ 500 verbs, not full paradigms
- **Output:** learned grammar that
  - Has an **asymmetry** between over-regularization and over-irregularization
  - Shows **developmental regression**

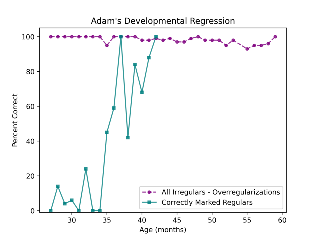


Dale Caldwell, Lt. Gov. Elect!

Marr's Levels: An Example

The Past Tense Debate: Computational Level

- **Goal:** learn **-ed** and its allomorphs
- **Input:** ≤ 500 verbs, not full paradigms
- **Output:** learned grammar that
  - Has an **asymmetry** between over-regularization and over-irregularization
  - Shows **developmental regression**



Dale Caldwell, Lt. Gov. Elect!

Rumelhart and McClelland (1986)

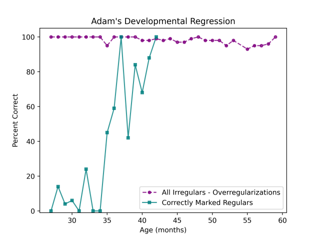
Distributed representations

Regulars and irregulars processed with the **same associative memory mechanism**

Marr's Levels: An Example

The Past Tense Debate: Computational Level

- **Goal:** learn **-ed** and its allomorphs
- **Input:** ≤ 500 verbs, not full paradigms
- **Output:** learned grammar that
  - Has an **asymmetry** between over-regularization and over-irregularization
  - Shows **developmental regression**



Dale Caldwell, Lt. Gov. Elect!

Rumelhart and McClelland (1986)

Distributed representations

Regulars and irregulars processed with the **same associative memory mechanism**

Pinker and Prince (1988)

Symbolic representations

Irregulars processed with **associative memory**, regulars processed with **symbolic rule**

Introduction: Returning to the Computational Level

Rumelhart and McClelland (1986) and **Pinker and Prince (1988)** agreed on the **computational-level** problem they were trying to solve!

- Their disagreements were at the **algorithmic level**

Introduction: Returning to the Computational Level

Rumelhart and McClelland (1986) and **Pinker and Prince (1988)** agreed on the **computational-level** problem they were trying to solve!

- Their disagreements were at the **algorithmic level**

Today, we **no longer agree on the computational level** problem

Introduction: Returning to the Computational Level

Rumelhart and McClelland (1986) and **Pinker and Prince (1988)** agreed on the **computational-level** problem they were trying to solve!

- Their disagreements were at the **algorithmic level**

Today, we **no longer agree on the computational level** problem

- **Input:**
  - Rumelhart and McClelland: **420** verbs
  - Kirov and Cotterell (2018): **3,500** verbs *in their complete paradigm*
  - Dankers et al. (2021): **46,000** German noun plurals
  - Warstadt and Bowman (2022): **100 million** tokens ($\approx$ input to 10 y.o.)
  - Hayes and Wilson (2008) remove **marginal forms** from train

Introduction: Returning to the Computational Level

Rumelhart and McClelland (1986) and **Pinker and Prince (1988)** agreed on the **computational-level** problem they were trying to solve!

- Their disagreements were at the **algorithmic level**

Today, we **no longer agree on the computational level** problem

- **Input:**
  - Rumelhart and McClelland: **420** verbs
  - Kirov and Cotterell (2018): **3,500** verbs *in their complete paradigm*
  - Dankers et al. (2021): **46,000** German noun plurals
  - Warstadt and Bowman (2022): **100 million** tokens ($\approx$ input to 10 y.o.)
  - Hayes and Wilson (2008) remove **marginal forms** from train
- **Developmental Patterns:**
  - Rumelhart and McClelland tried to achieve **developmental regression**
  - Kirov and Cotterell claim **micro U-shaped learning** across epochs

(Marcus et al., 1992; Belth et al., 2021; Payne and Kodner, 2025)

Returning to the Computational Level

As computational models of language acquisition proliferate, it seems high time to return to our **formal, computational-level specification** of the problem of acquisition. We wish to develop:

Introduction: Returning to the Computational Level

Returning to the Computational Level

As computational models of language acquisition proliferate, it seems high time to return to our **formal, computational-level specification** of the problem of acquisition. We wish to develop:

1. A **computational-level description** of language acquisition

Introduction: Returning to the Computational Level

Returning to the Computational Level

As computational models of language acquisition proliferate, it seems high time to return to our **formal, computational-level specification** of the problem of acquisition. We wish to develop:

1. A **computational-level description** of language acquisition
2. A method for **evaluating algorithmic implementations** to determine whether they're plausible models

Outline

Formalizing the Acquisition Problem

Kodner and Payne (2025) provide a formalization of language acquisition as a mapping from **some input sample** to **a sequence of hypothesized output grammars**.

Formalizing the Acquisition Problem

Kodner and Payne (2025) provide a formalization of language acquisition as a mapping from **some input sample** to **a sequence of hypothesized output grammars**.

This can be given as a typed function (Pierce, 2002) below:

$$A :: L(g_t) \rightarrow \langle g_h^1, g_h^2, ..., g_h^n \rangle$$

Formalizing the Acquisition Problem

Kodner and Payne (2025) provide a formalization of language acquisition as a mapping from **some input sample** to **a sequence of hypothesized output grammars**.

This can be given as a typed function (Pierce, 2002) below:

$$A :: L(g_t) \rightarrow \langle g_h^1, g_h^2, ..., g_h^n \rangle$$

Our approach builds on this, with two critical changes.

Outline

Change 1: The Input Sampling Function I

The input to acquisition is not the **entire extension** of the target grammar

$$A :: L(g_t) \to \langle g_h^1, g_h^2, ..., g_h^n \rangle$$

Change 1: The Input Sampling Function I

The input to acquisition is not the **entire extension** of the target grammar

$$A :: L(g_t) \to \langle g_h^1, g_h^2, ..., g_h^n \rangle$$

Instead, it's some **sequence of input utterances**

$$\langle u^1, u^2, ..., u^n \rangle$$

sampled from $L(g_t)$ with particular distributional properties.

Brown (1973); MacWhinney (1996); Chan (2008); Lignos and Yang (2016)

Change 1: The Input Sampling Function I

This **sequence of input utterances**:

Change 1: The Input Sampling Function I

This **sequence of input utterances**:

- Is made up of **language** (sequences of symbols) only

(Landau et al., 1985; Bedny et al., 2019; Madasu and Lal, 2023)

Change 1: The Input Sampling Function I

This **sequence of input utterances**:

- Is made up of **language** (sequences of symbols) only

  (Landau et al., 1985; Bedny et al., 2019; Madasu and Lal, 2023)

- Is made up of **positive examples** only
  - **Direct** negative evidence (corrections) are **ignored**
  - **Indirect** negative evidence is a non-starter given **gaps & sparisty**

  (Brown, 1970; Braine et al., 1971; Marcus et al., 1992; Marcus, 1993)

Change 1: The Input Sampling Function *I*

This **sequence of input utterances**:

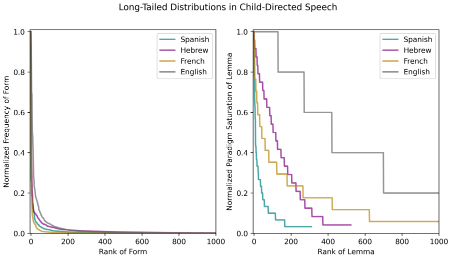
- Is made up of **language** (sequences of symbols) only

  (Landau et al., 1985; Bedny et al., 2019; Madasu and Lal, 2023)

- Is made up of **positive examples** only
  - **Direct** negative evidence (corrections) are **ignored**
  - **Indirect** negative evidence is a non-starter given **gaps & sparisty**

  (Brown, 1970; Braine et al., 1971; Marcus et al., 1992; Marcus, 1993)

- Follows a **sparse, skewed** frequency distribution



(Brown, 1973; MacWhinney, 1996; Chan, 2008; Lignos and Yang, 2016)

Change 1: The Input Sampling Function I

We define I to be the function which samples **a sequence of n utterances** from $L(g_t)$ which have these properties:

$$I :: L(g_t) \to n \to \langle u^1, u^2, ..., u^n \rangle$$

Change 1: The Input Sampling Function I

We define I to be the function which samples **a sequence of n utterances** from $L(g_t)$ which have these properties:

$$I :: L(g_t) \to n \to \langle u^1, u^2, ..., u^n \rangle$$

Now we have:

$$A :: L(g_t) \to \langle g_h^1, g_h^2, ..., g_h^n \rangle$$
$$\Downarrow$$
$$A :: \langle u^1, u^2, ..., u^n \rangle \to \langle g_h^1, g_h^2, ..., g_h^n \rangle$$

Where the sequence $\langle u^1, u^2, ..., u^n \rangle$ is sampled by I!

Outline

Change 2: The Lexicon

Not everything in the child's **input** is in their **intake** and thus added to their lexicon. (Corder, 1967; Pearl, 2007; Gagliardi, 2012)

Change 2: The Lexicon

Not everything in the child's **input** is in their **intake** and and thus added to their lexicon. (Corder, 1967; Pearl, 2007; Gagliardi, 2012)

- The acquisition algorithm builds not just the **algorithm** but the **lexicon** as well

Change 2: The Lexicon

Not everything in the child's **input** is in their **intake** and thus added to their lexicon. (Corder, 1967; Pearl, 2007; Gagliardi, 2012)

- The acquisition algorithm builds not just the **algorithm** but the **lexicon** as well

The acquisition function should output not just **a sequence of hypothesized grammars**, but a sequence of **paired grammars and lexicons**:

Change 2: The Lexicon

Not everything in the child's **input** is in their **intake** and thus added to their lexicon. (Corder, 1967; Pearl, 2007; Gagliardi, 2012)

- The acquisition algorithm builds not just the **algorithm** but the **lexicon** as well

The acquisition function should output not just **a sequence of hypothesized grammars**, but a sequence of **paired grammars and lexicons**:

$$A :: \langle u^1, u^2, ..., u^n \rangle \rightarrow \langle g_h^1, g_h^2, ..., g_h^n \rangle$$

$$\Downarrow$$

$$A :: \langle u^1, u^2, ..., u^n \rangle \rightarrow \langle (g_h^1, L^1), (g_h^2, L^2), ..., (g_h^n, L^n) \rangle$$

Outline

1. Computational-Level Description
   - The Input Sampling Function I
   - The Role of the Lexicon L
   - Finalizing our Formalization
   - Subcomponents and their Acquisition Functions A_C
   - The Lexical Acquisition Function A_L
2. Evaluation
   - Evaluating A_L and I
   - Implementing E
   - Distributional Equivalence
3. Conclusion

Finalizing our Formalization

The **final form of our acquisition function**:

$$A :: \langle u^1, u^2, ..., u^n \rangle \to \langle (g_h^1, L^1), (g_h^2, L^2), ..., (g_h^n, L^n) \rangle$$

Finalizing our Formalization

The **final form of our acquisition function**:

$$A :: \langle u^1, u^2, ..., u^n \rangle \to \langle (g_h^1, L^1), (g_h^2, L^2), ..., (g_h^n, L^n) \rangle$$

This is actually just the **input-to-state map of a discrete time dynamical system!** This is familiar to:

Finalizing our Formalization

The **final form of our acquisition function**:

$$A :: \langle u^1, u^2, ..., u^n \rangle \rightarrow \langle (g_h^1, L^1), (g_h^2, L^2), ..., (g_h^n, L^n) \rangle$$

This is actually just the **input-to-state map of a discrete time dynamical system!** This is familiar to:

- Previous formalizations of **acquisition & language change**

(e.g., Niyogi et al., 1997)

Finalizing our Formalization

The **final form of our acquisition function**:

$$A :: \langle u^1, u^2, ..., u^n \rangle \rightarrow \langle (g_h^1, L^1), (g_h^2, L^2), ..., (g_h^n, L^n) \rangle$$

This is actually just the **input-to-state map of a discrete time dynamical system!** This is familiar to:

- Previous formalizations of **acquisition & language change**

  (e.g., Niyogi et al., 1997)

- **Constructivist approaches** which view acquisition is viewed as a form of **self-organization**

  (Karpf, 1990; De Boer, 2000, 2005; Dressler et al., 2019; Dressler and Payne, to appear)

Finalizing our Formalization

The **final form of our acquisition function** is an input-to-state map of a discrete-time dynamical system:

$$A :: \langle u^1, u^2, ..., u^n \rangle \rightarrow \langle (g_h^1, L^1), (g_h^2, L^2), ..., (g_h^n, L^n) \rangle$$

Finalizing our Formalization

The **final form of our acquisition function** is an input-to-state map of a discrete-time dynamical system:

$$A :: \langle u^1, u^2, ..., u^n \rangle \rightarrow \langle (g_h^1, L^1), (g_h^2, L^2), ..., (g_h^n, L^n) \rangle$$

But language acquisition happens **online**, so for our purposes it makes more sense to look at the **state update function**:

$$A :: u^i \rightarrow (g_h^i, L^i) \rightarrow (g_h^{i+1}, L^{i+1})$$

Finalizing our Formalization

The **final form of our acquisition function** is an input-to-state map of a discrete-time dynamical system:

$$A :: \langle u^1, u^2, ..., u^n \rangle \rightarrow \langle (g_h^1, L^1), (g_h^2, L^2), ..., (g_h^n, L^n) \rangle$$

But language acquisition happens **online**, so for our purposes it makes more sense to look at the **state update function**:

$$A :: u^i \rightarrow (g_h^i, L^i) \rightarrow (g_h^{i+1}, L^{i+1})$$

This tells us how we go from a **single input instance** and the **current state of the grammar & lexicon** to the **next state of the grammar & lexicon**.

Outline

The Role of Subcomponents

Under most linguistic theories, the grammar is not an **undifferentiated whole**, but divided into **subcomponents**:

- **Structure** vs. **meaning**
- **Phonological, morphological, syntactic, etc.** structure

(de Saussure, 1916; Chomsky, 1957; Jackendoff, 1972; Goldberg, 2006; Sadock, 2012, i.a.)

The Role of Subcomponents

Under most linguistic theories, the grammar is not an **undifferentiated whole**, but divided into **subcomponents**:

- **Structure** vs. **meaning**
- **Phonological, morphological, syntactic, etc.** structure

(de Saussure, 1916; Chomsky, 1957; Jackendoff, 1972; Goldberg, 2006; Sadock, 2012, i.a.)

For ease of exposition, we will characterize this view as treating each g_h^i as **a set of subcomponent grammars:**

$$g_h^i = \{g_P^i, g_M^i, g_S^i, ...\}$$

The Acquisition of a Subcomponent

Each subcomponent is learned from **the same sequence of inputs and lexicon** by a **different (albeit related) acquisition function**:

- The phonological subcomponent g_P is learned by A_P, the syntactic subcomponent g_S is learned by A_S, and so on.

The Acquisition of a Subcomponent

Each subcomponent is learned from **the same sequence of inputs and lexicon** by a **different (albeit related) acquisition function**:

- The phonological subcomponent g_P is learned by A_P, the syntactic subcomponent g_S is learned by A_S, and so on.

This approach assumes a level of **independence** between the acquisition of each subcomponent:

- In line with **modular approaches to the mind** from cognitive science (Fodor 1983; Cosmides and Tooby 1992; Sperber 1994; Pinker 1997; Samuels 1998; Carruthers 2002, i.a.)

The Acquisition of a Subcomponent

Each subcomponent is learned from **the same sequence of inputs and lexicon** by a **different (albeit related) acquisition function**:

- The phonological subcomponent g_P is learned by A_P, the syntactic subcomponent g_S is learned by A_S, and so on.

This approach assumes a level of **independence** between the acquisition of each subcomponent:

- In line with **modular approaches to the mind** from cognitive science (Fodor 1983; Cosmides and Tooby 1992; Sperber 1994; Pinker 1997; Samuels 1998; Carruthers 2002, i.a.)
- Allows us to retain a level of **agnosticism regarding the exact subcomponents** of g_h^i

The Acquisition of a Subcomponent

Each subcomponent is learned from **the same sequence of inputs and lexicon** by a **different (albeit related) acquisition function**:

- The phonological subcomponent g_P is learned by A_P, the syntactic subcomponent g_S is learned by A_S, and so on.

This approach assumes a level of **independence** between the acquisition of each subcomponent:

- In line with **modular approaches to the mind** from cognitive science (Fodor 1983; Cosmides and Tooby 1992; Sperber 1994; Pinker 1997; Samuels 1998; Carruthers 2002, i.a.)
- Allows us to retain a level of **agnosticism regarding the exact subcomponents** of g_h^i

But it would still be easy to adapt this to **include interaction between the subcomponents** while remaining in line with our formalism!

The Acquisition of a Subcomponent

For some **subcomponent** C, the role of A_C is to:

The Acquisition of a Subcomponent

For some **subcomponent** C, the role of A_C is to:

- **Select the relevant information** from the input utterance u^i and lexicon

The Acquisition of a Subcomponent

For some **subcomponent** C, the role of A_C is to:

- **Select the relevant information** from the input utterance u^i and lexicon
- Use this information to **map from the input** utterance u^i, current grammar g_C^i, and newly updated lexicon L^{i+1} **to the next state of the grammar** g_C^{i+1}

The Acquisition of a Subcomponent

For some **subcomponent** C, the role of A_C is to:

- **Select the relevant information** from the input utterance u^i and lexicon
- Use this information to **map from the input** utterance u^i, current grammar g_C^i, and newly updated lexicon L^{i+1} **to the next state of the grammar** g_C^{i+1}

A_C is thus given by:

$$A_C :: g_C^i \to u^i \to L^{i+1} \to g_C^{i+1}$$

The Acquisition of a Subcomponent

For some **subcomponent** C, the role of A_C is to:

- **Select the relevant information** from the input utterance u^i and lexicon
- Use this information to **map from the input** utterance u^i, current grammar g_C^i, and newly updated lexicon L^{i+1} **to the next state of the grammar** g_C^{i+1}

A_C is thus given by:

$$A_C :: g_C^i \to u^i \to L^{i+1} \to g_C^{i+1}$$

This is another **state update function** for a **discrete time dynamical system!**

Outline

The Lexical Acquisition Function A_L

We have the following definition of A_C:

$$A_C :: g_C^i \rightarrow u^i \rightarrow L^{i+1} \rightarrow g_C^{i+1}$$

The Lexical Acquisition Function A_L

We have the following definition of A_C:

$$A_C :: g_C^i \to u^i \to L^{i+1} \to g_C^{i+1}$$

For all subcomponents C, the A_Cs learn from **the same lexicon L^{i+1}**

The Lexical Acquisition Function A_L

We have the following definition of A_C:

$$A_C :: g_C^i \to u^i \to L^{i+1} \to g_C^{i+1}$$

For all subcomponents C, the A_Cs learn from **the same lexicon L^{i+1}**

We also want a function A_L that **updates the lexicon** to the next state L^{i+1} given u^i and L^i:

$$A_L :: L^i \to u^i \to L^{i+1}$$

The Lexical Acquisition Function A_L

We have the following definition of A_C:

$$A_C :: g_C^i \to u^i \to L^{i+1} \to g_C^{i+1}$$

For all subcomponents C, the A_Cs learn from **the same lexicon** L^{i+1}

We also want a function A_L that **updates the lexicon** to the next state L^{i+1} given u^i and L^i:

$$A_L :: L^i \to u^i \to L^{i+1}$$

This is yet another **state update function** for a **discrete time dynamical system!**

The Lexical Acquisition Function A_L

Two crucial **computational-level properties of A_L**:

The Lexical Acquisition Function A_L

Two crucial **computational-level properties of A_L**:

- **Input frequency** is a strong driver of **order of acquisition**

(Palermo and Eberhart 1968; Goodman et al. 2008; Swingley and Humphrey 2018; Braginsky et al. 2019)

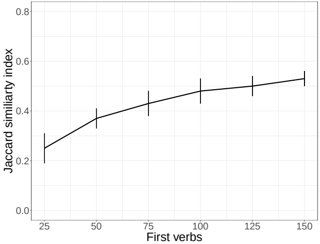
The Lexical Acquisition Function A_L

Two crucial **computational-level properties of A_L**:

- **Input frequency** is a strong driver of **order of acquisition**

  (Palermo and Eberhart 1968; Goodman et al. 2008; Swingley and Humphrey 2018; Braginsky et al. 2019)

- **Variation** in the contents of early lexicons

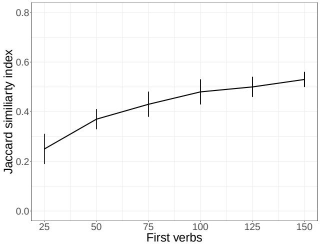


(Demuth et al., 2006; Yang et al., in prep)

The Lexical Acquisition Function A_L

Two crucial **computational-level properties of A_L**:

- **Input frequency** is a strong driver of **order of acquisition**
(Palermo and Eberhart 1968; Goodman et al. 2008; Swingley and Humphrey 2018; Braginsky et al. 2019)
- **Variation** in the contents of early lexicons



(Demuth et al., 2006; Yang et al., in prep)

A number of **algorithmic-level implementations** of A_L exist

(Yu and Smith 2007; Goodman et al. 2007; Frank et al. 2009; Fazly et al. 2010; Trueswell et al. 2013; Roembke and McMurray 2016; Stevens et al. 2017; Berens et al. 2018; Roembke et al. 2023; Yue et al. 2023; i.a.)

Putting it all Together

Having defined our **overall acquisition function** A, our **lexical acquisition function** A_L , and our **subcomponent acquisition functions** A_C, we can now define their relationship:

Putting it all Together

Having defined our **overall acquisition function** A, our **lexical acquisition function** A_L , and our **subcomponent acquisition functions** A_C, we can now define their relationship:

procedure $A(u^i, (g_h^i, L^i))$
  $L^{i+1} \leftarrow A_L(L^i, u^i)$
  $g_h^{i+1} \leftarrow \bigcup_{g_C^i \in g_h^i} A_C(g_C^i, u^i, L^{i+1})$
  return (g_h^{i+1}, L^{i+1})
end procedure

Outline

The Evaluation Function

How do we determine **whether an algorithmic implementation of A_C** which fits our computational description **is a plausible acquisition model?**

The Evaluation Function

How do we determine **whether an algorithmic implementation of A_C** which fits our computational description **is a plausible acquisition model?**

We'll need an **evaluation function E** which:

The Evaluation Function

How do we determine **whether an algorithmic implementation of A_C** which fits our computational description **is a plausible acquisition model?**

We'll need an **evaluation function E** which:

- Takes in an **acquisition algorithm A_C**

The Evaluation Function

How do we determine **whether an algorithmic implementation of A_C** which fits our computational description **is a plausible acquisition model?**

We'll need an **evaluation function E** which:

- Takes in an **acquisition algorithm A_C**
- Takes in some **empirical results** for the acquisition of that subcomponent R_C

The Evaluation Function

How do we determine **whether an algorithmic implementation of A_C** which fits our computational description **is a plausible acquisition model?**

We'll need an **evaluation function E** which:

- Takes in an **acquisition algorithm A_C**
- Takes in some **empirical results** for the acquisition of that subcomponent R_C
- Returns a **scalar score s** corresponding to the **plausibility of A_C**

The Evaluation Function

How do we determine **whether an algorithmic implementation of A_C** which fits our computational description **is a plausible acquisition model?**

We'll need an **evaluation function** E which:

- Takes in an **acquisition algorithm** A_C
- Takes in some **empirical results** for the acquisition of that subcomponent R_C
- Returns a **scalar score** s corresponding to the **plausibility of** A_C

$$E :: A_C \rightarrow R_C \rightarrow s$$

Outline

The Evaluation Function

$$E :: A_C \to R_C \to s$$

Problem: each A_C takes in the lexicon L^{i+1} and input utterance u^i

The Evaluation Function

$$E :: A_C \rightarrow R_C \rightarrow s$$

Problem: each A_C takes in the lexicon L^{i+1} and input utterance u^i

- So **we also need functions to evaluate the plausibility of A_L and I**

The Evaluation Function

$$E :: A_C \rightarrow R_C \rightarrow s$$

Problem: each A_C takes in the lexicon L^{i+1} and input utterance u^i

- So **we also need functions to evaluate the plausibility of A_L and I**
- Luckily, we have **a lot of empirical evidence regarding the nature of both!**

The Evaluation Function

$$E :: A_C \rightarrow R_C \rightarrow s$$

Problem: each A_C takes in the lexicon L^{i+1} and input utterance u^i

- So **we also need functions to evaluate the plausibility of A_L and I**
- Luckily, we have **a lot of empirical evidence regarding the nature of both!**

Evaluating our Input Sampling Function I

$$E_I :: I \rightarrow s$$

Scores I based on:

- **Sufficient variation** in input between learners
- **Sparse, skewed** distributions
- Appropriate **input size** n

The Evaluation Function

$$E :: A_C \to R_C \to s$$

Problem: each A_C takes in the lexicon L^{i+1} and input utterance u^i

- So **we also need functions to evaluate the plausibility of A_L and I**
- Luckily, we have **a lot of empirical evidence regarding the nature of both!**

Evaluating our Lexical Acquisition Function A_L

$$E_L :: A_L \to s$$

Scores A_L based on:

- Ability to learn a **plausibly-sized lexicon**
- Ability to exhibit **variation in lexical acquisition trajectories**
- Ability to do all this **from a plausible I!**

The Complete Evaluation Function

Given our implementations of E_I and E_L, let's **expand our original** E:

$$E :: I \to A_L \to A_C \to R_C \to s$$

Where I and A_L must have scored highly on E_I and E_L, respectively.

The Complete Evaluation Function

Given our implementations of E_I and E_L, let's **expand our original** E:

$$E :: I \rightarrow A_L \rightarrow A_C \rightarrow R_C \rightarrow s$$

Where I and A_L must have scored highly on E_I and E_L, respectively.

Taking in I allows us to **evaluate A_C on a range of input sequences**

- Crucial for **variation in acquisition trajectories!**

Outline

Implementing E

How do we implement our evaluation function E?

- It's not enough to state that **our final hypothesized grammar achieves high accuracy!**

Implementing E

How do we implement our evaluation function E?

- It's not enough to state that **our final hypothesized grammar achieves high accuracy!**
- The intermediate hypothesized grammars g_C^i provide **a wealth of information about the acquisition trajectory & plausibility of A_C**

Implementing E

How do we implement our evaluation function E?

- It's not enough to state that **our final hypothesized grammar achieves high accuracy!**
- The intermediate hypothesized grammars g_C^i provide **a wealth of information about the acquisition trajectory & plausibility of A_C**
  - **Example: English past tense** must show:
    - **Developmental regression**
    - Asymmetry between **over-regularization vs. over-irregularization**

Implementing E

How do we implement our evaluation function E?

- It's not enough to state that **our final hypothesized grammar achieves high accuracy!**
- The intermediate hypothesized grammars g_C^i provide **a wealth of information about the acquisition trajectory & plausibility of A_C**
  - **Example: English past tense** must show:
    - **Developmental regression**
    - Asymmetry between **over-regularization vs. over-irregularization**
  - Rumelhart and McClelland (1986) and Pinker and Prince (1988) **agreed on this computational-level specification!**

Implementing E

How do we evaluate this sequence of hypothesized grammars?

Implementing E

How do we evaluate this sequence of hypothesized grammars?

Intensional Equivalence is a non-starter

- The **internal structure of the grammar** is the **main open question in theoretical linguistics!**

Implementing E

How do we evaluate this sequence of hypothesized grammars?

Intensional Equivalence is a non-starter

- The **internal structure of the grammar** is the **main open question in theoretical linguistics!**

Strict Extensional Equivalence is also a non-starter

Implementing E

How do we evaluate this sequence of hypothesized grammars?

Intensional Equivalence is a non-starter

- The **internal structure of the grammar** is the **main open question in theoretical linguistics!**

Strict Extensional Equivalence is also a non-starter

- The language of the child's final grammar **may not be identical** to the language(s) of the grammar(s) of their caretaker(s)
- This is particularly true in cases of **language change**

(e.g., Niyogi et al., 1997; Kodner, 2020, 2022)

Implementing E

How do we evaluate this sequence of hypothesized grammars?

Intensional Equivalence is a non-starter

- The **internal structure of the grammar** is the **main open question in theoretical linguistics!**

Strict Extensional Equivalence is also a non-starter

- The language of the child's final grammar **may not be identical** to the language(s) of the grammar(s) of their caretaker(s)
- This is particularly true in cases of **language change**

  (e.g., Niyogi et al., 1997; Kodner, 2020, 2022)
- For any two learners A and B, **we don't actually expect strict extensional equivalence** between their grammars at time i

Distributional Equivalence

Consider the **intermediate grammars** of **two learners of the English past tense**.

Distributional Equivalence

Consider the **intermediate grammars** of **two learners of the English past tense**.

Learner A

Has intermediate morphological grammar g_M^A such that:

- $g_M^A(go) \rightarrow goed$
- $g_M^A(feel) \rightarrow felt$

Distributional Equivalence

Consider the **intermediate grammars** of **two learners of the English past tense**.

Learner A

Has intermediate morphological grammar g_M^A such that:

- $g_M^A(go) \rightarrow goed$
- $g_M^A(feel) \rightarrow felt$

Learner B

Has intermediate morphological grammar g_M^B such that:

- $g_M^B(go) \rightarrow went$
- $g_M^A(feel) \rightarrow feeled$

Distributional Equivalence

Consider the **intermediate grammars** of **two learners of the English past tense**.

Learner A

Has intermediate morphological grammar g_M^A such that:

- $g_M^A(go) \rightarrow goed$
- $g_M^A(feel) \rightarrow felt$

Learner B

Has intermediate morphological grammar g_M^B such that:

- $g_M^B(go) \rightarrow went$
- $g_M^A(feel) \rightarrow feeled$

Which one of these is a **better model** of over-regularization?

Distributional Equivalence

Consider the **intermediate grammars** of **two learners of the English past tense**.

Learner A

Has intermediate morphological grammar g_M^A such that:

- $g_M^A(go) \rightarrow goed$
- $g_M^A(feel) \rightarrow felt$

Learner B

Has intermediate morphological grammar g_M^B such that:

- $g_M^B(go) \rightarrow went$
- $g_M^A(feel) \rightarrow feeled$

Which one of these is a **better model** of over-regularization?

Obviously, neither.

Outline

Distributional Equivalence

Intuition: we don't care about the **exact** outputs of our intermediate grammars for each input. Rather, we care about the **distribution** of these outputs.

Distributional Equivalence

Intuition: we don't care about the **exact** outputs of our intermediate grammars for each input. Rather, we care about the **distribution** of these outputs.

- Because learners A and B will receive **distinct input sequences** sampled by some I and build **distinct lexicons** based on some A_L, they **will not be identical in their behavior in terms of specific lexical items**

Distributional Equivalence

Intuition: we don't care about the **exact** outputs of our intermediate grammars for each input. Rather, we care about the **distribution** of these outputs.

- Because learners A and B will receive **distinct input sequences** sampled by some I and build **distinct lexicons** based on some A_L, they **will not be identical in their behavior in terms of specific lexical items**
- What matters is that both learner A and learner B **show expected tendencies in their output distributions:**
  - **Asymmetry** between over-regularization and over-irregularization
  - **Developmental regression**

Distributional Equivalence

Intuition: we don't care about the **exact** outputs of our intermediate grammars for each input. Rather, we care about the **distribution** of these outputs.

- Because learners A and B will receive **distinct input sequences** sampled by some I and build **distinct lexicons** based on some A_L, they **will not be identical in their behavior in terms of specific lexical items**
- What matters is that both learner A and learner B **show expected tendencies in their output distributions:**
  - **Asymmetry** between over-regularization and over-irregularization
  - **Developmental regression**

We'll refer to this notion as **distributional equivalence.**

Implementing Distributional Equivalence

This notion isn't new, but so far **its implementation has been "vibes-based"**

Implementing Distributional Equivalence

This notion isn't new, but so far **its implementation has been "vibes-based"**

- **Visual inspection of learner output** to examine trajectory & comparison to relevant plots from acquisition

Implementing Distributional Equivalence

This notion isn't new, but so far **its implementation has been "vibes-based"**

- **Visual inspection of learner output** to examine trajectory & comparison to relevant plots from acquisition
- This actually makes a lot of sense given that **so many models don't even pass the vibe check!**
  - No NN, for example, has **ever** shown developmental regression

Implementing Distributional Equivalence

This notion isn't new, but so far **its implementation has been "vibes-based"**

- **Visual inspection of learner output** to examine trajectory & comparison to relevant plots from acquisition
- This actually makes a lot of sense given that **so many models don't even pass the vibe check!**
  - No NN, for example, has **ever** shown developmental regression

But to provide rigorous, quantified evaluations, we must go beyond this!

(Rumelhart and McClelland 1986; Pinker and Prince 1988; Belth et al. 2021; Payne and Kodner 2025; Kodner et al. 2025; Yang et al. in prep; Dressler and Payne to appear)

Implementing Distributional Equivalence

To **quantify distributional equivalence**, we want to have a metric that:

Implementing Distributional Equivalence

To **quantify distributional equivalence**, we want to have a metric that:

- Quantifies **the similarity of two categorical distributions**
  - In this case, that of g_C^i and R_C^i

Implementing Distributional Equivalence

To **quantify distributional equivalence**, we want to have a metric that:

- Quantifies **the similarity of two categorical distributions**
  - In this case, that of g_C^i and R_C^i
- Can do so over **a series of time points**

To **quantify distributional equivalence**, we want to have a metric that:

- Quantifies **the similarity of two categorical distributions**
  - In this case, that of g_C^i and R_C^i
- Can do so over **a series of time points**

Some candidates include:

- **Significance Testing:**

Implementing Distributional Equivalence

To **quantify distributional equivalence**, we want to have a metric that:

- Quantifies **the similarity of two categorical distributions**
  - In this case, that of g_C^i and R_C^i
- Can do so over **a series of time points**

Some candidates include:

- **Significance Testing:**
  - **G-Test:** likelihood-ratio test which measures how probable the distribution of g_C^i is if it came from the same distribution as R_C^i.

Implementing Distributional Equivalence

To **quantify distributional equivalence**, we want to have a metric that:

- Quantifies **the similarity of two categorical distributions**
  - In this case, that of g_C^i and R_C^i
- Can do so over **a series of time points**

Some candidates include:

- **Significance Testing:**
  - **G-Test:** likelihood-ratio test which measures how probable the distribution of g_C^i is if it came from the same distribution as R_C^i.
  - χ^2 **Test** is likely more familiar; compares squared differences rather than likelihoods but is not ideal for small or uneven distributions.

Implementing Distributional Equivalence

To **quantify distributional equivalence**, we want to have a metric that:

- Quantifies **the similarity of two categorical distributions**
  - In this case, that of g_C^i and R_C^i
- Can do so over **a series of time points**

Some candidates include:

- **Significance Testing:**
  - **G-Test:** likelihood-ratio test which measures how probable the distribution of g_C^i is if it came from the same distribution as R_C^i.
  - χ^2 **Test** is likely more familiar; compares squared differences rather than likelihoods but is not ideal for small or uneven distributions.
  - **Fisher's Exact Test** computes exact probability and is good for small samples, but doesn't scale up well.

Implementing Distributional Equivalence

To **quantify distributional equivalence**, we want to have a metric that:

- Quantifies **the similarity of two categorical distributions**
  - In this case, that of g_C^i and R_C^i
- Can do so over **a series of time points**

Some candidates include:

- **Effect Size:**

Implementing Distributional Equivalence

To **quantify distributional equivalence**, we want to have a metric that:

- Quantifies **the similarity of two categorical distributions**
  - In this case, that of g_C^i and R_C^i
- Can do so over **a series of time points**

Some candidates include:

- **Effect Size:**
  - **Cramer's V** effect size calculated from χ^2 or G-test by normalizing to a 0-1 scale.

Implementing Distributional Equivalence

To **quantify distributional equivalence**, we want to have a metric that:

- Quantifies **the similarity of two categorical distributions**
  - In this case, that of g_C^i and R_C^i
- Can do so over **a series of time points**

Some candidates include:

- **Effect Size:**
  - **Cramer's V** effect size calculated from χ^2 or G-test by normalizing to a 0-1 scale.
  - **Jensen-Shannon divergence*** symmetric measure of information difference between two distributions which ranges from 0 to 1.

Implementing Distributional Equivalence

To **quantify distributional equivalence**, we want to have a metric that:

- Quantifies **the similarity of two categorical distributions**
  - In this case, that of g_C^i and R_C^i
- Can do so over **a series of time points**

Some candidates include:

- **Effect Size:**
  - **Cramer's V** effect size calculated from χ^2 or G-test by normalizing to a 0-1 scale.
  - **Jensen-Shannon divergence*** symmetric measure of information difference between two distributions which ranges from 0 to 1.
  - **Hellinger Distance** symmetric measure of geometric distance between distributions.

Implementing Distributional Equivalence

To **quantify distributional equivalence**, we want to have a metric that:

- Quantifies **the similarity of two categorical distributions**
  - In this case, that of g_C^i and R_C^i
- Can do so over **a series of time points**

Some candidates include:

- **Effect Size:**
  - **Cramer's V** effect size calculated from χ^2 or G-test by normalizing to a 0-1 scale.
  - **Jensen-Shannon divergence*** symmetric measure of information difference between two distributions which ranges from 0 to 1.
  - **Hellinger Distance** symmetric measure of geometric distance between distributions.
  - **Total Variation Distance (L1 distance)** the fraction of probability mass that must be moved to transform between distributions.

Implementing Distributional Equivalence

To **quantify distributional equivalence**, we want to have a metric that:

- Quantifies **the similarity of two categorical distributions**
  - In this case, that of g_C^i and R_C^i
- Can do so over **a series of time points**

Some candidates include:

- **Effect Size:**
  - **Cramer's V** effect size calculated from χ^2 or G-test by normalizing to a 0-1 scale.
  - **Jensen-Shannon divergence*** symmetric measure of information difference between two distributions which ranges from 0 to 1.
  - **Hellinger Distance** symmetric measure of geometric distance between distributions.
  - **Total Variation Distance (L1 distance)** the fraction of probability mass that must be moved to transform between distributions.

What other measures do you suggest?

*Thanks to Caleb Belth for this suggestion!

Outline

Conclusion and Open Questions

In this presentation we have:

Conclusion and Open Questions

In this presentation we have:

- Provided a **formalization of the language acquisition problem** at Marr's computational level

Conclusion and Open Questions

In this presentation we have:

- Provided a **formalization of the language acquisition problem** at Marr's computational level
- Provided a **computational-level description of the corresponding evaluation function**

Conclusion and Open Questions

In this presentation we have:

- Provided a **formalization of the language acquisition problem** at Marr's computational level
- Provided a **computational-level description of the corresponding evaluation function**
- Discussed **distributional equivalence** and possible implementations

Conclusion and Open Questions

In this presentation we have:

- Provided a **formalization of the language acquisition problem** at Marr's computational level
- Provided a **computational-level description of the corresponding evaluation function**
- Discussed **distributional equivalence** and possible implementations

Open questions and next steps:

Conclusion and Open Questions

In this presentation we have:

- Provided a **formalization of the language acquisition problem** at Marr's computational level
- Provided a **computational-level description of the corresponding evaluation function**
- Discussed **distributional equivalence** and possible implementations

Open questions and next steps:

- How do we **quantify distributional equivalence**? **(suggestions welcome!)**

Conclusion and Open Questions

In this presentation we have:

- Provided a **formalization of the language acquisition problem** at Marr's computational level
- Provided a **computational-level description of the corresponding evaluation function**
- Discussed **distributional equivalence** and possible implementations

Open questions and next steps:

- How do we **quantify distributional equivalence**? **(suggestions welcome!)**
- For the measures considered here, **what's the role of variation?**

Conclusion and Open Questions

In this presentation we have:

- Provided a **formalization of the language acquisition problem** at Marr's computational level
- Provided a **computational-level description of the corresponding evaluation function**
- Discussed **distributional equivalence** and possible implementations

Open questions and next steps:

- How do we **quantify distributional equivalence**? **(suggestions welcome!)**
- For the measures considered here, **what's the role of variation?**
- How do we apply these metrics **over time series?**

Thank you!

I am grateful to Caleb Belth, Charles Yang, Katie Schuler, Jordan Kodner, Scott Nelson, Jeff Heinz, Logan Swanson, and Dwyer Bradley for discussion.

This work was supported by the Institute for Advanced Computational Science Graduate Research Fellowship and the National Science Foundation Graduate Research Fellowship Program.

References I

Marina Bedny, Jorie Koster-Hale, Giulia Elli, Lindsay Yazzolino, and Rebecca Saxe. 2019. There's more to "sparkle" than meets the eye: Knowledge of vision and light verbs among congenitally blind and sighted individuals. *Cognition*, 189:105–115.

C. Belth, S. Payne, D. Beser, Jordan Kodner, and Charles Yang. 2021. The Greedy and Recursive Search for Morphological Productivity. In *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society (CogSci)*, volume 43, pages 2869–2875.

Sam C Berens, Jessica S Horst, and Chris M Bird. 2018. Cross-situational learning is supported by propose-but-verify hypothesis testing. *Current Biology*, 28(7):1132–1136.

Mika Braginsky, Daniel Yurovsky, Virginia A Marchman, and Michael C Frank. 2019. Consistency and variability in children's word learning across languages. *Open Mind*, 3:52–67.

Martin DS Braine et al. 1971. On two types of models of the internalization of grammars. *The ontogenesis of grammar*, 1971:153–186.

Roger Brown. 1970. Derivational complexity and order of acquisition. *Cognition and Development of Language*.

Roger Brown. 1973. *A first language: The early stages*. Harvard University Press, Cambridge, MA.

References II

Peter Carruthers. 2002. The cognitive functions of language. *Behavioral and brain sciences*, 25(6):657–674.

Erwin Chan. 2008. *Structures and distributions in morphological learning*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA.

Noam Chomsky. 1957. *Syntactic structures*. Walter de Gruyter.

Stephen Pit Corder. 1967. The significance of learners' errors. *IRAL: International Review of Applied Linguistics in Language Teaching*, 5(4):161–170.

Leda Cosmides and John Tooby. 1992. Cognitive adaptations for social exchange. *The adapted mind: Evolutionary psychology and the generation of culture*, 163:163–228.

Verna Dankers, Anna Langedijk, Kate McCurdy, Adina Williams, and Dieuwke Hupkes. 2021. Generalising to german plural noun classes, from the perspective of a recurrent neural network. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 94–108.

Bart De Boer. 2000. Self-organization in vowel systems. *Journal of phonetics*, 28(4):441–465.

Bart De Boer. 2005. Self-organisation in language. *Self-organization and evolution of social systems*, pages 123–139.

References III

Ferdinand de Saussure. 1916. *Course in General Linguistics*. Columbia University Press.

Katherine Demuth, Jennifer Culbertson, and Jennifer Alter. 2006. Word-minimality, epenthesis and coda licensing in the early acquisition of english. *Language and speech*, 49(2):137–173.

Wolfgang U. Dressler, Anastasia Christofidou, Natalia Gagarina, Katharina Korecky-Kröll, and Marianne Kilani-Schoch. 2019. Morphological blind-alley developments as a theoretical challenge to both usage-based and nativist acquisition models. *The Italian Journal of Linguistics*, 31:107–140.

Wolfgang U. Dressler and S. Payne. to appear. Self-Organization in Acquisition. In *Cambridge Handbook of Natural Linguistics*. Cambridge University Press.

Afsaneh Fazly, Afra Alishahi, and Suzanne Stevenson. 2010. A probabilistic computational model of cross-situational word learning. *Cognitive science*, 34(6):1017–1063.

Jerry A Fodor. 1983. *The modularity of mind*. MIT press.

Michael C Frank, Noah D Goodman, and Joshua B Tenenbaum. 2009. Using speakers' referential intentions to model early cross-situational word learning. *Psychological science*, 20(5):578–585.

References IV

Ann C Gagliardi. 2012. *Input and intake in language acquisition*. University of Maryland, College Park.

Adele E Goldberg. 2006. *Constructions at work: The nature of generalization in language*. Oxford University Press.

Judith C Goodman, Philip S Dale, and Ping Li. 2008. Does frequency count? Parental input and the acquisition of vocabulary. *Journal of child language*, 35(3):515–531.

Noah Goodman, Joshua Tenenbaum, and Michael Black. 2007. A bayesian framework for cross-situational word-learning. *Advances in neural information processing systems*, 20.

Bruce Hayes and Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic inquiry*, 39(3):379–440.

Ray S Jackendoff. 1972. Semantic interpretation in generative grammar.

Annemarie Karpf. 1990. *Selbstorganisationsprozesse in der sprachlichen Ontogenese: Erst-und Fremdsprache (n)*, volume 352. Gunter Narr Verlag.

Christo Kirov and Ryan Cotterell. 2018. Recurrent neural networks in linguistic theory: Revisiting Pinker and Prince (1988) and the Past Tense Debate. *Transactions of the Association for Computational Linguistics*, 6:651–665.

References V

J. Kodner, S. Payne, S. Khalifa, and Z. Liu. 2025. Evaluating Learning Trajectories of Neural Morphology Acquisition Models. In *Linguistics Vanguard*. De Gruyter.

Jordan Kodner. 2020. *Language acquisition in the past*. Ph.D. thesis, University of Pennsylvania.

Jordan Kodner. 2022. Language acquisition guiding theory and diachrony: A case study from latin morphology. *Natural Language and Linguistic Theory*, 41:733–792.

Jordan Kodner and Sarah Payne. 2025. Formally defining the learning setting for child language acquisition. Talk given at the "Computational Models of Learnability and Acquisition of Morphology and Phonology" special session at the 2025 Meeting of the Linguistic Society of America.

Barbara Landau, Lila R Gleitman, and Barbara Landau. 1985. *Language and experience: Evidence from the blind child*. Harvard University Press.

Constantine Lignos and Charles Yang. 2016. Morphology and language acquisition. In Andrew Hippisley and Gregory Stump, editors, *The Cambridge handbook of morphology*, pages 743–764. Cambridge University Press, Cambridge, UK.

Brian MacWhinney. 1996. The CHILDES system. *American Journal of Speech-language Pathology*, 5(1):5–14.

References VI

Avinash Madasu and Vasudev Lal. 2023. Is multimodal vision supervision beneficial to language? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2637–2642.

Gary F Marcus. 1993. Negative evidence in language acquisition. *Cognition*, 46(1):53–85.

Gary F Marcus, Steven Pinker, Michael Ullman, Michelle Hollander, T John Rosen, Fei Xu, and Harald Clahsen. 1992. Overregularization in language acquisition. *Monographs of the society for research in child development*, 57(4):1–178.

David Marr. 1982. *Vision: A computational investigation into the human representation and processing of visual information*. MIT press.

Partha Niyogi, Robert C Berwick, et al. 1997. A dynamical systems model for language change. *Complex Systems*, 11(3):161–204.

David S Palermo and V Lynn Eberhart. 1968. On the learning of morphological rules: An experimental analogy. *Journal of Verbal Learning and Verbal Behavior*, 7(2):337–344.

S. Payne and J. Kodner. 2025. Some Innate Characteristics of Neural Models of Morphological Inflection. In *Linguistics Vanguard*. De Gruyter.

References VII

Lisa Sue Pearl. 2007. *Necessary bias in natural language learning*. Ph.D. thesis, University of Maryland, College Park, MD.

Benjamin C Pierce. 2002. *Types and programming languages*. MIT press.

Steven Pinker. 1997. How the mind works. *WW Norton & Company*.

Steven Pinker and Alan Prince. 1988. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, 28(1-2):73–193.

Tanja C Roembke and Bob McMurray. 2016. Observational word learning: Beyond propose-but-verify and associative bean counting. *Journal of memory and language*, 87:105–127.

Tanja C Roembke, Matilde E Simonetti, Iring Koch, and Andrea M Philipp. 2023. What have we learned from 15 years of research on cross-situational word learning? a focused review. *Frontiers in Psychology*, 14:1175272.

David E. Rumelhart and James L. McClelland. 1986. On learning the past tenses of English verbs. In David E. Rumelhart, James L. McClelland, and the PDP Research Group, editors, *Parallel Distributed Processing*, volume 2, pages 216–271. The MIT Press, Cambridge, MA.

References VIII

Jerrold M Sadock. 2012. *The modular architecture of grammar*. 132. Cambridge university press.

Richard Samuels. 1998. Evolutionary psychology and the massive modularity hypothesis.

Dan Sperber. 1994. The modularity of thought and the epidemiology of representations. *Mapping the mind: Domain specificity in cognition and culture*, 39:67.

Jon Scott Stevens, Lila R Gleitman, John C Trueswell, and Charles Yang. 2017. The pursuit of word meanings. *Cognitive science*, 41:638–676.

Daniel Swingley and Colman Humphrey. 2018. Quantitative linguistic predictors of infants' learning of specific English words. *Child development*, 89(4):1247–1267.

John C Trueswell, Tamara Nicol Medina, Alon Hafri, and Lila R Gleitman. 2013. Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive psychology*, 66(1):126–156.

Alex Warstadt and Samuel R Bowman. 2022. What artificial neural networks can tell us about human language acquisition. In Shalom Lappin and Jean-Philippe Bernardy, editors, *Algebraic structures in natural language*, pages 17–60. CRC Press, Boca Raton.

References IX

Charles Yang, Sarah Payne, Caleb Belth, and Jordan Kodner. in prep. An adequate discovery procedure.

Chen Yu and Linda B Smith. 2007. Rapid word learning under uncertainty via cross-situational statistics. *Psychological science*, 18(5):414–420.

Christine Soh Yue, Alexander S LaTourrette, Charles Yang, and John Trueswell. 2023. Memory as a computational constraint in cross-situational word learning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 45.