Morphology & Language Acquisition

Constantine Lignos & Charles Yang

Learnability
27 September 2022
Sarah Brogden Payne

Introduction

- Morphology varies widely across languages
- Children have to learn the morphology of their language from sparse input
- What does this input look like?
- How do children learn from it?

Outline

1. Input Distributions
2. Developmental Findings
3. Learning Models

Outline

1. **Input Distributions**
2. Developmental Findings
3. Learning Models

Distributions

Actual learning conditions:

- **Unsupervised** learning on
- **Sparse data**

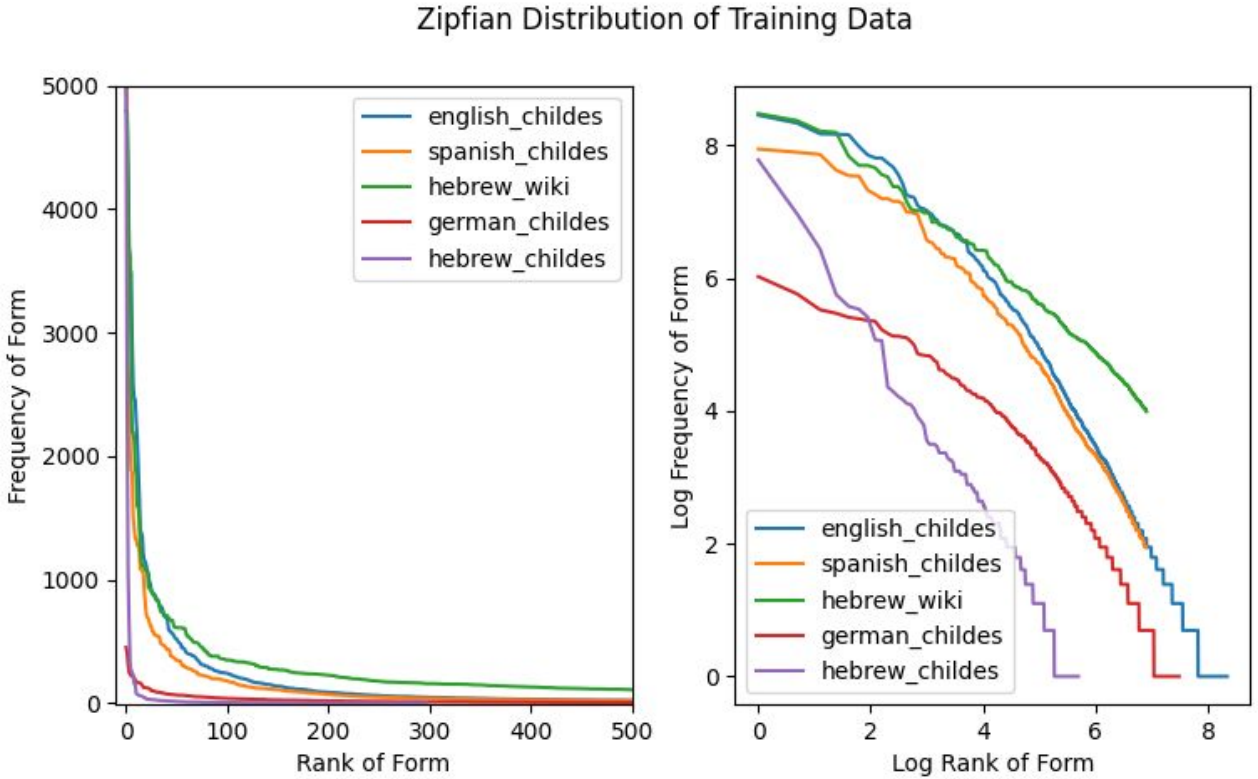
Unsupervised Learning

- Children aren’t explicitly provided with pairs of morphologically-related words (e.g. sing ~ sang)
  - They have to infer these pairs from the input and phonological relationships
- Children aren’t explicitly given the features that inflected forms realize
  - How do we know “she went to the park” is past tense? 3rd singular?
  - Distributional cues: pronouns, etc.

Sparse Input: Zipf’s Law for Words

- Frequency of a word is inversely proportional to its rank
  - A few words are used very frequently
  - Most words are rare (long tail)
- True across languages and datasets

Sparse Input: Zipf's Law for Words



From Payne (2022)

Sparse Input: Zipf's Law for Inflectional Morphology

- We can look at both the lemma rank and inflection rank
  - e.g. for *walk-s*, what is the rank of:
    - *Walk* across all inflected forms vs. other verbs across all forms **(lemma rank)**
    - 3rd singular vs. other inflections (past, etc.) **(inflection rank)**
- Of almost a million tokens of Spanish child-directed speech:
  - ~1500 lemmas, 54 inflectional categories
  - **10 most frequent lemmas = 42%** of occurrences of verbs, but **521 lemmas appear once**
  - 3rd singular present appears **37k times** but 1 & 2 imperfect subjunctive appear **once each**

Sparse Input: Zipf's Law for Inflectional Morphology

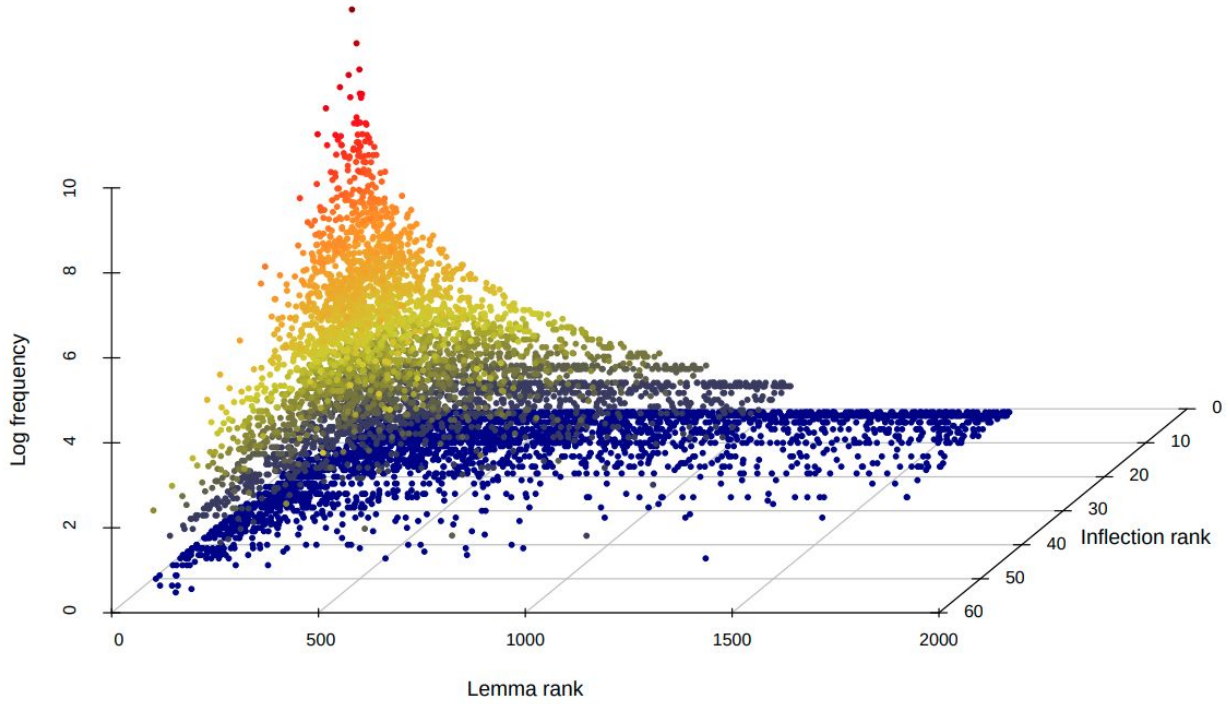


Figure 1: Frequencies of CHILDES Spanish lemmas across inflection categories.

Sparse Input: Zipf's Law for Paradigm Saturation

- Computed per lemma:
  - (# inflectional categories lemma has been seen in)/(total # inflectional categories)
  - Example: 6/30 = 20% saturation

	Present	Preterite	Imperfect	Conditional	Future
yo	amo	amé	amaba	amaría	amaré
tú	amas	amaste	amabas	amarías	amarás
él/ella/Ud.	ama	amó	amaba	amaría	amará
nosotros	amamos	amamos	amábamos	amaríamos	amaremos
vosotros	amáis	amasteis	amabais	amaríais	amaréis
ellos/ellas/Uds.	aman	amaron	amaban	amarían	amarán

Sparse Input: Zipf's Law for Paradigm Saturation

- **Findings:**
  - Most saturated Spanish form is **72% saturated**
  - **Mean saturation across lemmas is 7.9% ≅ 1/13**
  - No language besides English reaches 100% saturation

Sparse Input: Zipf's Law for Paradigm Saturation

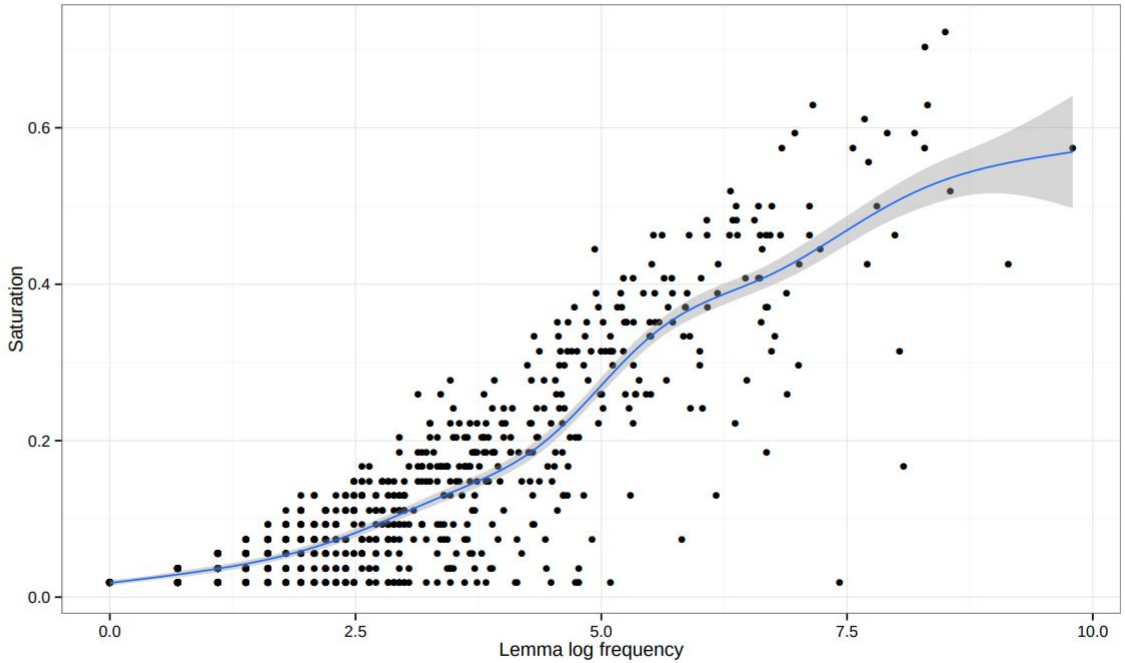
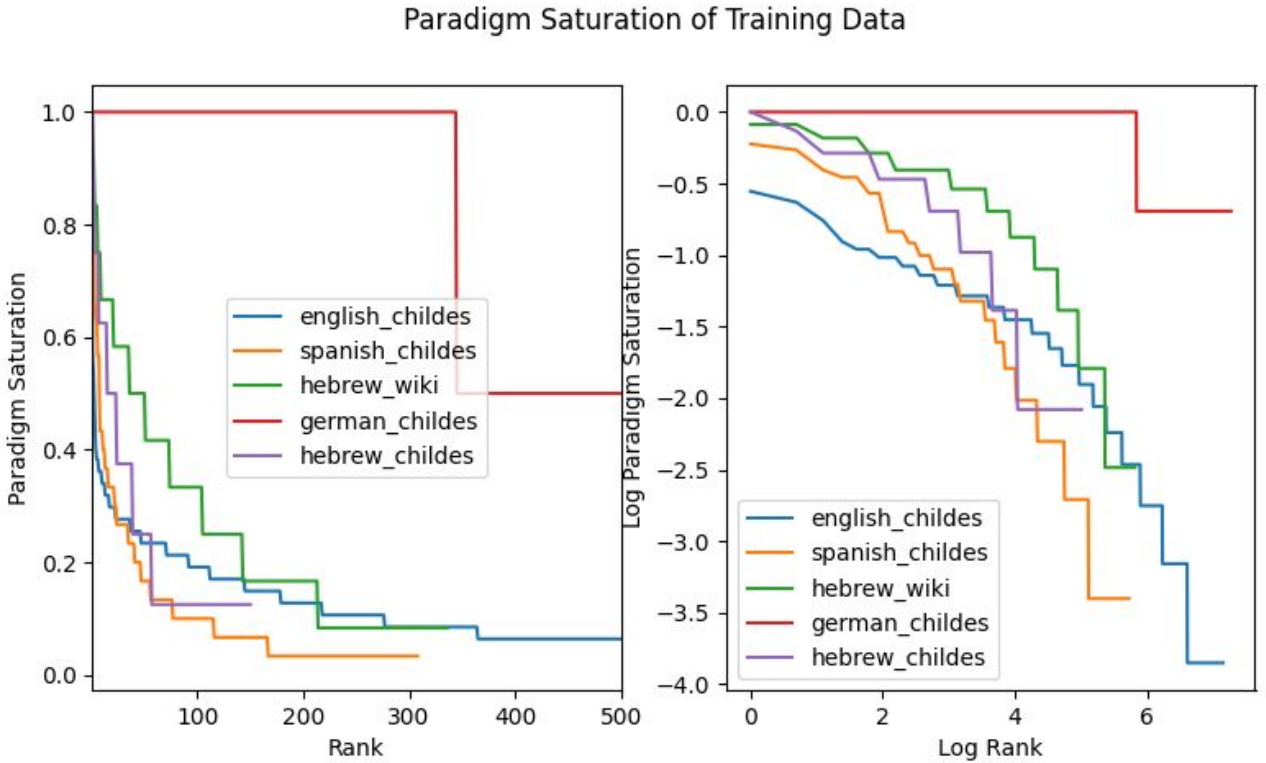


Figure 2: Saturation of CHILDES Spanish lemmas across lemma frequencies, with a GAM-derived fit line and standard error estimate.

Sparse Input: Zipf's Law for Paradigm Saturation



From Payne (2022)

Implications for Learning

- Can’t expect learner to have access to the full paradigm
- The learner needs to **“generalize aggressively”**
- Sparse distribution must be taken into account when assessing children’s productions
    - **Usage-based** people say children’s low morphological diversity in production means they’re just memorizing the input
    - But **similar low diversity happens for adults** – this is just how the distribution is!

Outline

1. Input Distributions
2. **Developmental Findings**
3. Learning Models

Development: The Past Tense Debate!

- How do we learn to productively add *-ed* to generate the past tense despite exceptions (go, see, feel, etc)?
- **Rumelhart & McClelland: Single Route**
  - Irregulars **and** regulars are learned as paired associations
  - Early connectionism
- **Pinker & Prince: Dual Route**
  - Irregulars are learned as paired associations
  - But regulars are handled by a default symbolic rule
- **Newer Work:**
  - Irregulars are handled by lexicalized rules (which still must be associated with lemmas) and regulars are handled by generalizable rules

Development: Child Errors

- **Over-regularization** (feel~feeled)
  - Quite common in child speech!
  - More frequent words are less likely to be over-regularized
  - **But** it turns out that rule frequency is a better predictor than word frequency
    - *teach* vs. *fly*
  - This supports an irregular rules account!
- **Over-irregularization** (e.g. wipe-wope)
  - Almost never happens! (0.02%)

Development: Child Errors

- Across languages, child errors are overwhelmingly either:
  - **Over-regularization** – applying a productive rule to an irregular
  - **Omission** – not inflecting the form at all
- Strong evidence for a categorical distinction between regular and irregular forms

Development: Paradigmatic Gaps

- Polish masculine genitive singular takes *-a* or *-u* but neither is productive
  - Children make few errors on this => lexicalized
- Polish masculine genitive plural takes *-ow* with some exceptions
  - Children make **more errors** here => productive rule
- **You don’t have to have a default!** (c.f. Pinker and Prince)

Development: Summary

- Clear distinction between regulars and irregulars (c.f. Rumelhart & McClelland)
- Evidence for irregular rules rather than associative memory (c.f. Pinker and Prince)
- You don’t have to have a default! (c.f. Pinker & Prince)
- Generalization!!!

This supports accounts where irregulars are handled by lexical rules and there is (optionally) a general/default productive rule

Outline

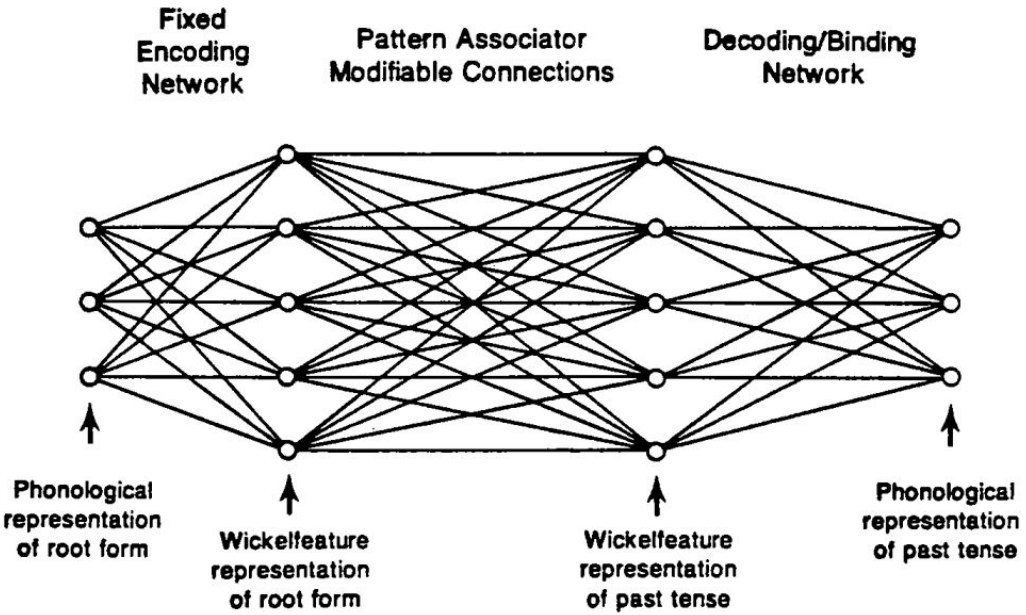
1. Input Distributions
2. Developmental Findings
3. **Learning Models**

Learning Models: Requirements

- Should **over-regularize** but not over-irregularize
- Focus on **productivity**: should generalize well from a sparse distribution
- Current models take in stem (e.g. *walk*) and produce inflected form (e.g. *walked*), but where to these pairs come from?

Learning Models: The Past Tense Debate (Again!)

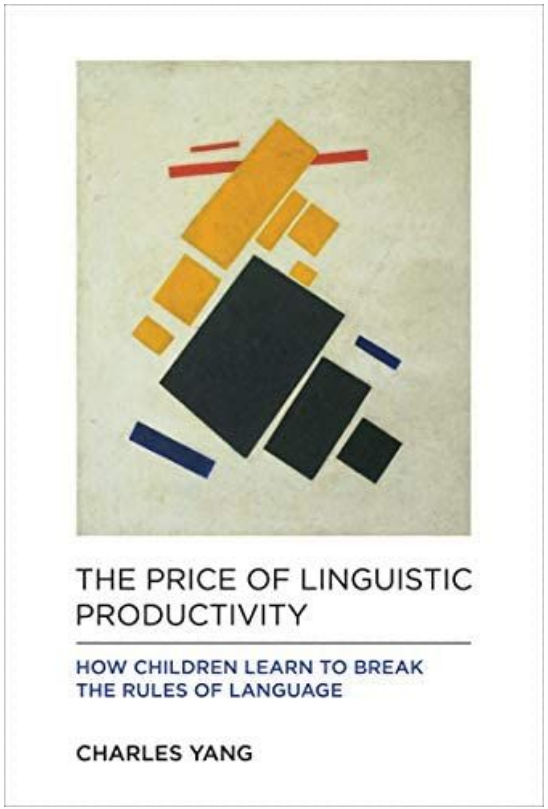
- **Rumelhart & McClelland:** all forms are paired associations, regular and irregular (connectionist)
  - Produced high rates of weird over-irregularizations (mail~membled)
  - Input not representative of children's: large # irregulars followed by regulars



Aside: Connectionism Today

- **Kirov & Cotterell (2018):** encoder-decoder NN for inflecting the past tense
  - Much more accurate and fewer weird errors than R&M – can connectionism work after all?
- K&C model sees 3500 verb types **in their complete paradigm**
  - Of the top ~3000 verb types in CHILDES, only a third appear in their complete paradigm
  - Makes 100 passes over the data
- Corkery et al (2019) find that K&C model still produces **more over-irregularizations than humans**

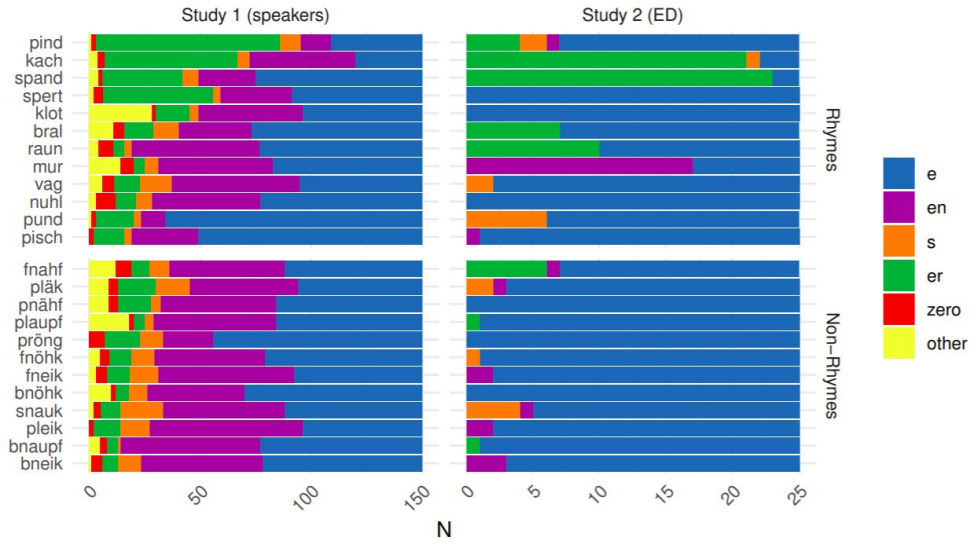
Aside: Rule-Based Models Today



- **Albright & Hayes (2002):** makes minimal generalizations over the input to learn sets of rules
- **Yang (2016):** the **Tolerance Principle** is supported by the developmental points made here
  - Assumes Zipfian frequency distribution and irregular rules
  - Provides threshold for when it will be more efficient to generalize under these conditions

Aside: The German Noun Plural Debate

- In English past tense, the statistically-dominant form is the default
- In German noun plurals, one of the least frequent forms (*-s*) is the default
- Today, models evaluated on both of these to ensure they **generalize** and don't just match the frequency distribution of their input



Learning Models: Distributional Learning

- Input to learning models is neatly organized into pairs of forms representing a single inflectional change
  - e.g. *go~went, walk~walked*
- How do children get these pairs?
  - For truly unsupervised learning, we will need models that are able to extract these pairs as well as learn from them

Learning Models: Summary

- Models should:
  - **generalize aggressively**
  - **over-regularize** more than they over-irregularize
  - successfully learn from **sparse data**
  - be able to extract input pairs themselves

Thanks!