Variation in First Language Development

Stoll & Lieven

Lecture 5:
Using corpora

Outline

- The most comprehensive crosslinguistic study to date
- Methdological issues
  - Sampling
  - Measuring development
  - Productivity
- Some examples of corpus analyses
  - Spanish verb inflections
  - Extracting and testing probabilistic grammars
  - Testing for an effect of input on me-for-I errors

The Crosslinguistic Study of Language Acquisition (Slobin Ed.) [1985, 1992, 1997]

Descriptive chapters on the acquisition of 28 languages (19 families)

- Indo-European
  - Romance: French, Italian, Portuguese, Rumanian, Spanish
  - Germanic: English, German
  - Slavic: Polish
- Semitic: Hebrew
- Finno-Ugric: Hungarian
- Ural-Altaic: Turkish
- Japanese
- Trans-New Guinea non-Austronesian: Kaluli
- Polynesian: Samoan
- Sign Language: ASL

Structure of Slobin’s volumes

- Introductory materials about the language
- Language acquisition data (errors, error-free acquisition, timing of acquisition)
- Data on the setting of language acquisition (cognitive pacesetting, linguistic pacesetting, input and adult-child interaction, individual differences)

Comparative Study (Slobin, 1985)

To study a specific question:

- Choose a group of languages that share a selection of features. Then one can focus on variation along specified dimensions.
- Keep most factors constant in order to explore the role of variation of a specific feature, e.g. test for the replication of a developmental pattern across languages.

Pye C., Pfeiler, B., de León, L., Brown, P. & Mateo, P. 2007. Roots or edges?: A comparative study of Mayan children's early verb forms.

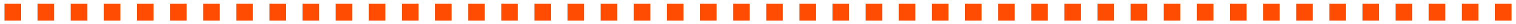
Comparative language acquisition

Issues:

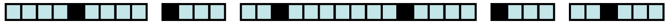
- Complexity of grammatical structure and corresponding tasks for the learning child
- Comparability (structural and developmental)
- Sampling
- Measuring productivity

Sampling: types of corpora

traditional = *1 hour per 1-2 weeks, 26-52 hours per year* = *1-2%*



high density = *5 hours per week, 260 hours per year* = *10%*
double density = *10 hours per week* = *20%*



diary **= *???***

Tomasello & Stahl, 2004
Rowland, Fletcher & Freudenthal, 2008

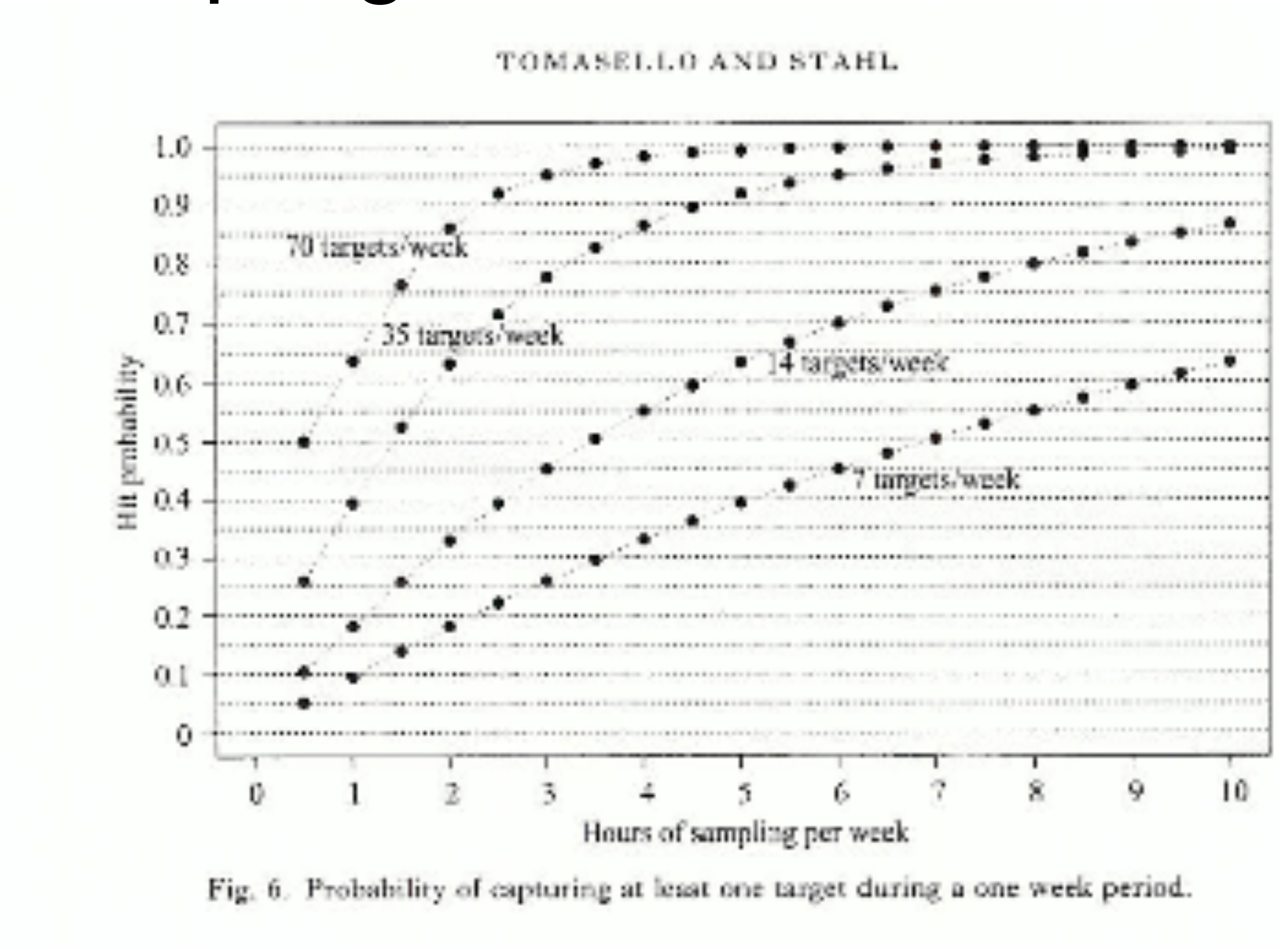
Productivity: Sampling Issues

[Tomasello & Stahl 2004]

Two important factors:

- Occurrence (frequency) of the linguistic feature
- Sample size and density of data collection:
  - Influences the probability of detecting a feature in the corpus
  - Influences the reliability of the estimates we make
  - Influences the estimated age of the first occurrence

Sampling Issues (Tomasello&Stahl 2004)



TOMASELLO AND STAHL

Fig. 6. Probability of capturing at least one target during a one week period.

How can we measure development?

- Are there stages and how can we determine them?
- Criteria for developmental stages:
    - Age
    - MLU (mean length of utterance)
    - MSL (mean syntactic length) or IPS (index of productive syntax)
    - Testing the child on standardised tests

How can we measure development?

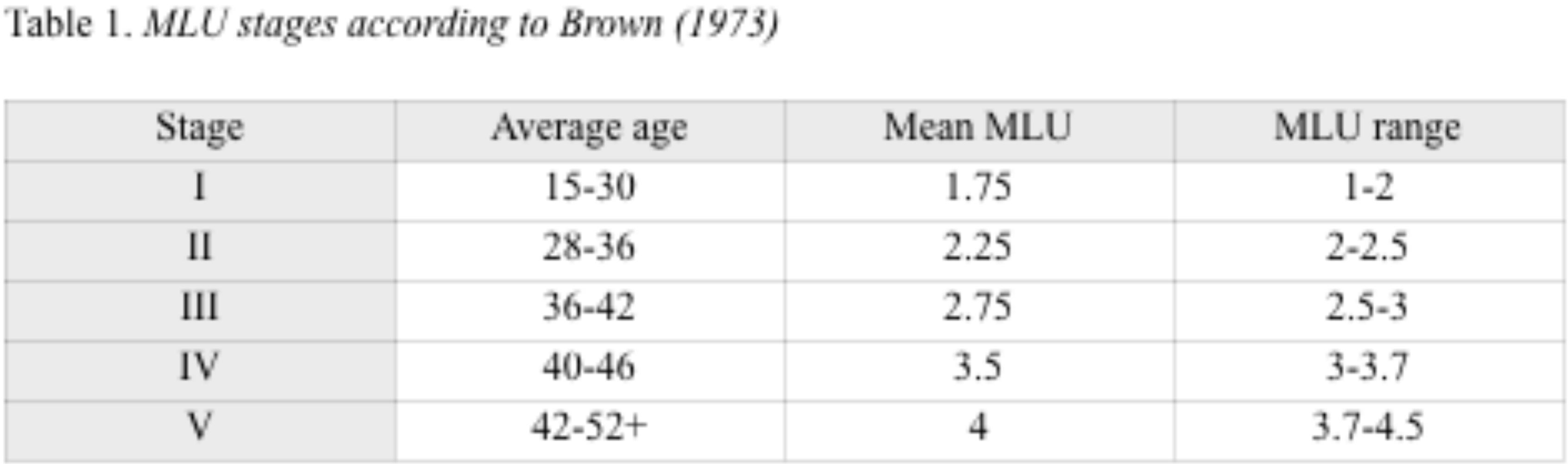
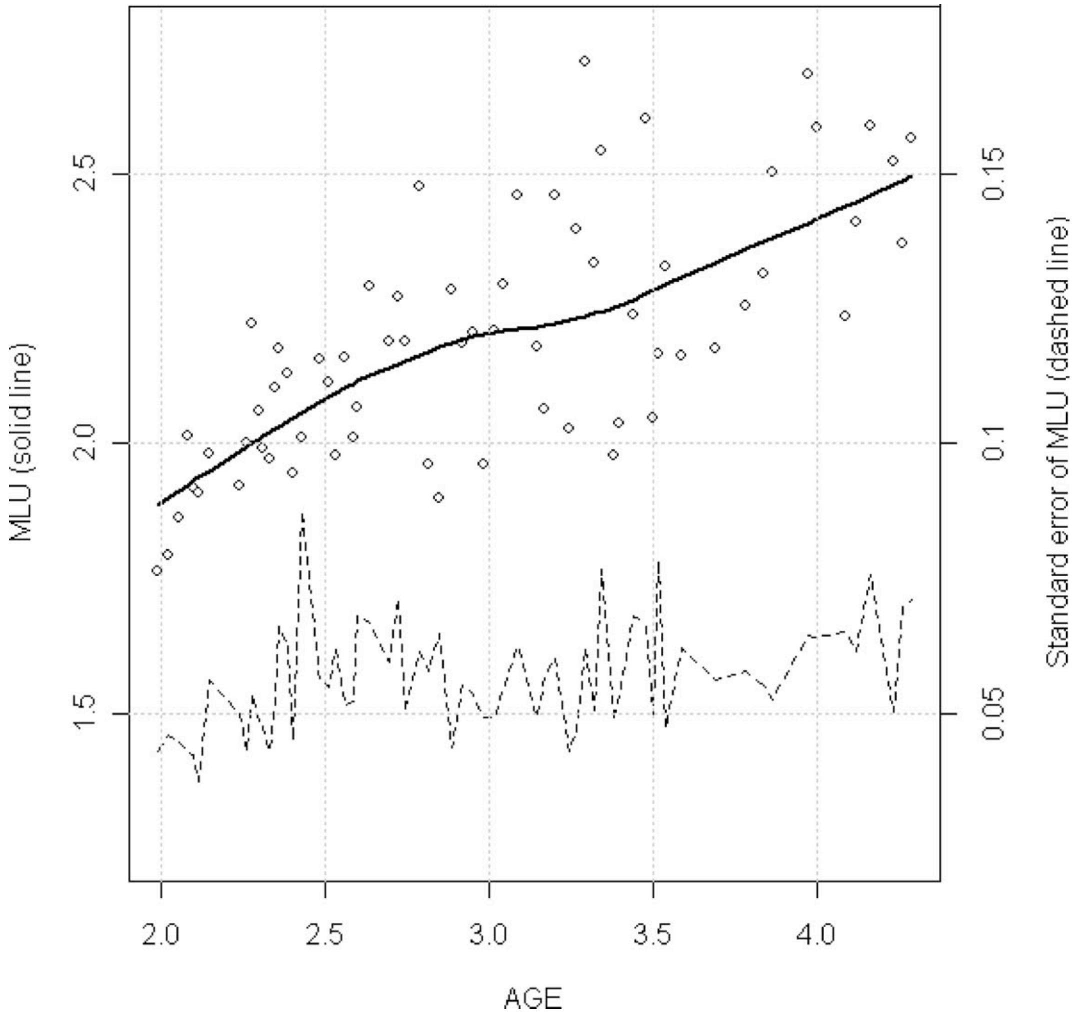


Table 1. *MLU stages according to Brown (1973)*

Stage	Average age	Mean MLU	MLU range
I	15-30	1.75	1-2
II	28-36	2.25	2-2.5
III	36-42	2.75	2.5-3
IV	40-46	3.5	3-3.7
V	42-52+	4	3.7-4.5

MLU of one Russian child



(Gries & Stoll, 2009)

Problems with MLU

- What to count? Words or morphemes?
- Extreme variation between recordings
- How can we distinguish non significant variation in MLU values that is determined by contextual variation or the shape of the child from actual development?
- There are no sound criteria to determine stages

Problems with stages

1. Developmental problem: large variation in different recordings (e.g.; recording at 2;01.12 with MLU value of 1.96 and 2;10.06 with MLU value of 1;9).

- 2. Arbitrariness problem: boundaries between MLU stages are completely arbitrary.
- 3. Subjectivity problem: group data and respect developmental problem.

- Develop methods that allow such comparisons in a principled bottom-up manner using hierarchical agglomerative clustering.
  - Variability neighborhood clustering
  - Take all MLU values of all recordings
  - Filter out random variation resulting from performance issues

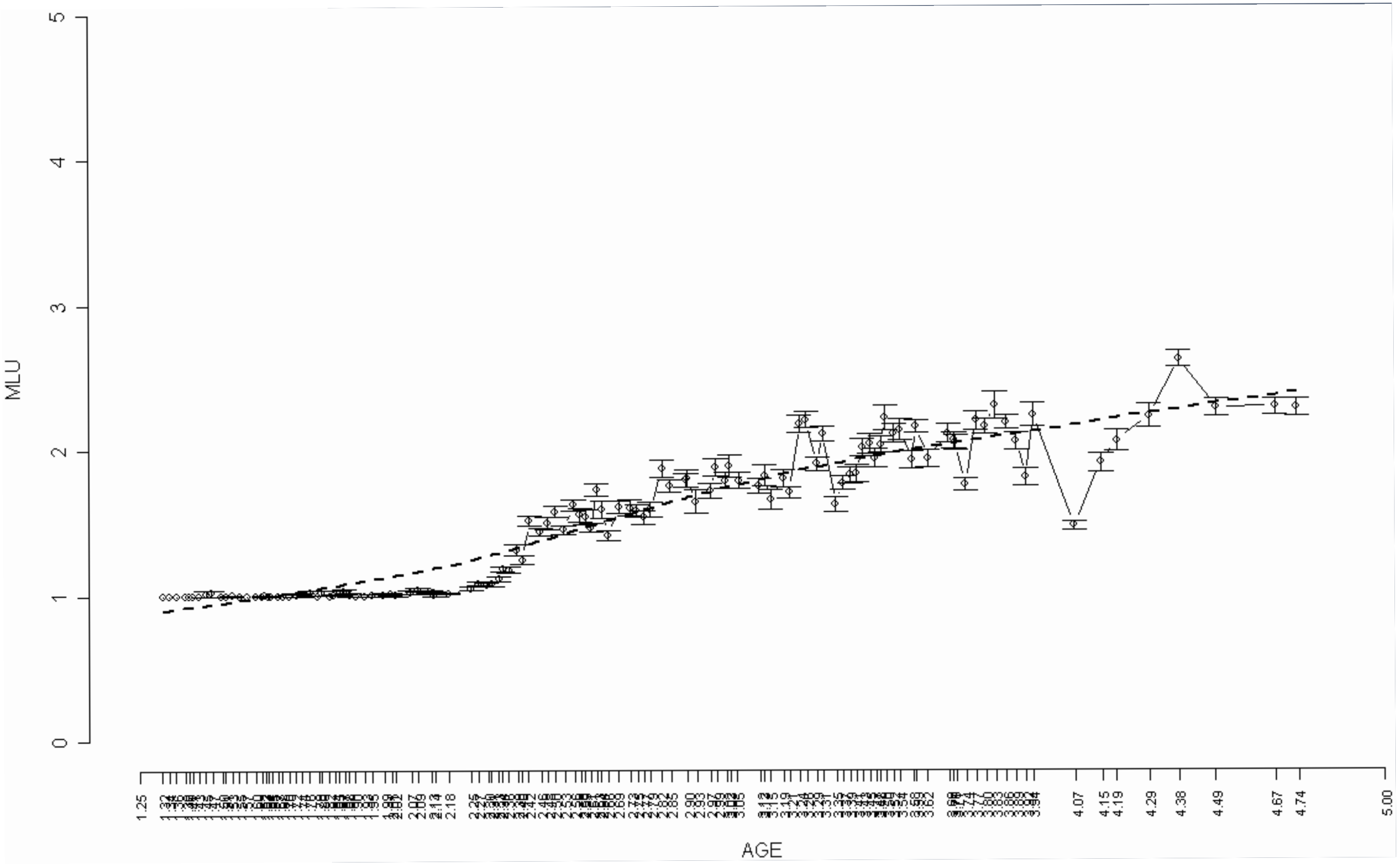
(Gries & Stoll 2009)

Hierarchical agglomerative clustering

[Gries & Stoll 2009]

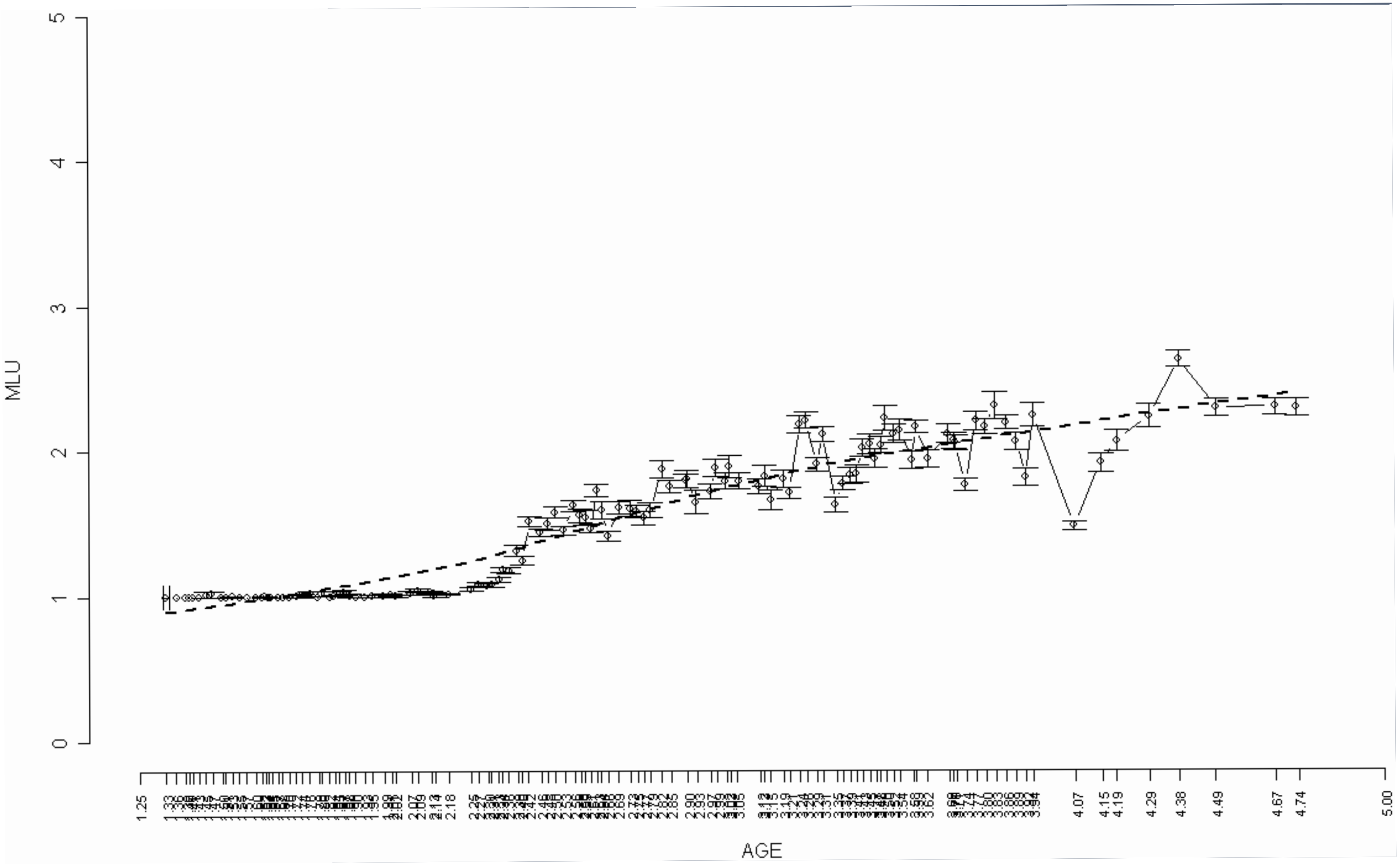
- Compute a distance, which provides the similarity of all measures on the basis of some distance measure.
- Identify the 2 elements that are most similar to each other
- Merge the 2 elements that are most similar to each other and compute new distances on the basis of this merger (until the number of elements is one)
- Draw a dendrogram that summarizes the groupings.

MLU before amalgamation starts

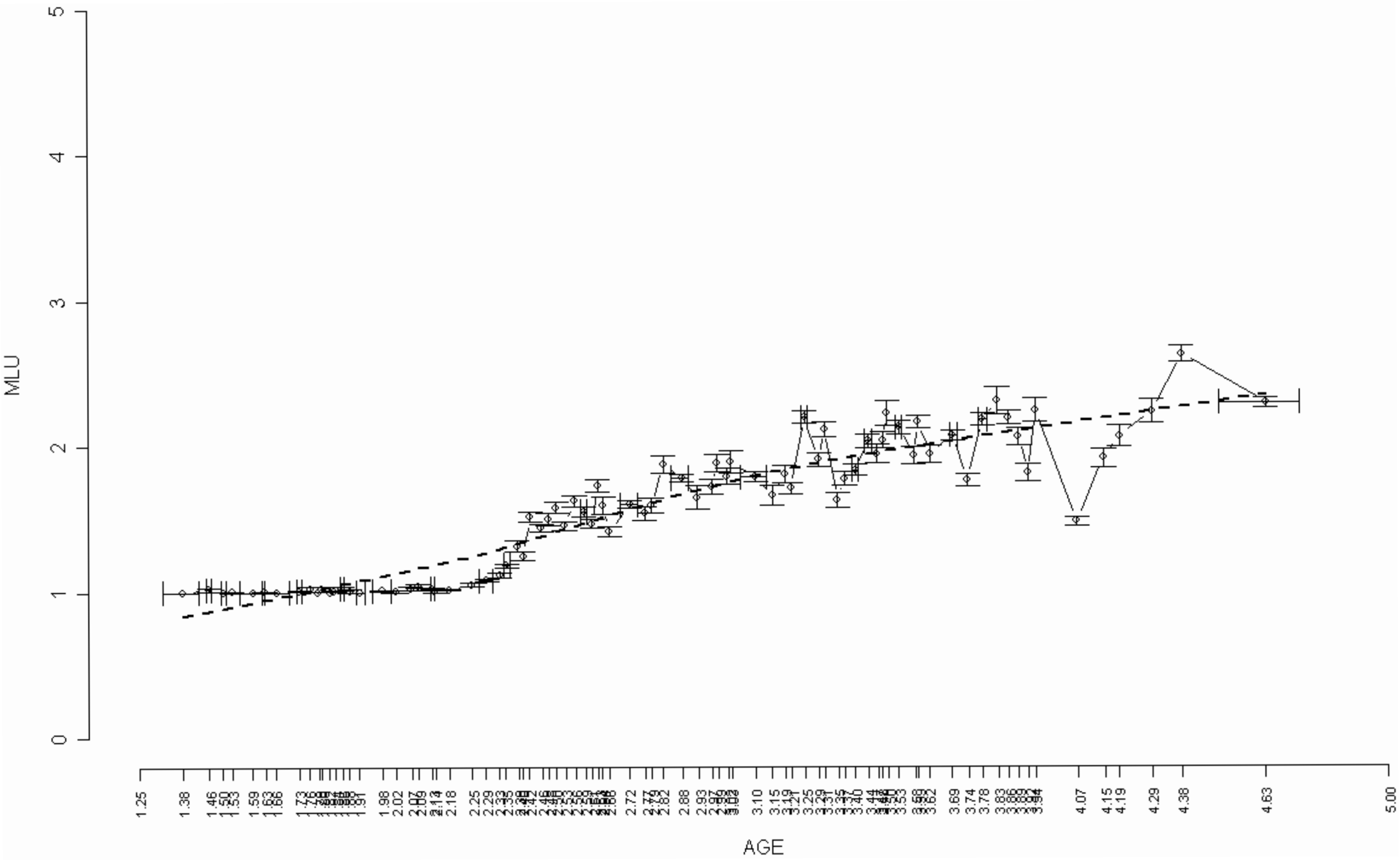


Stoll&Lieven:LSS.2010:Lecture 5

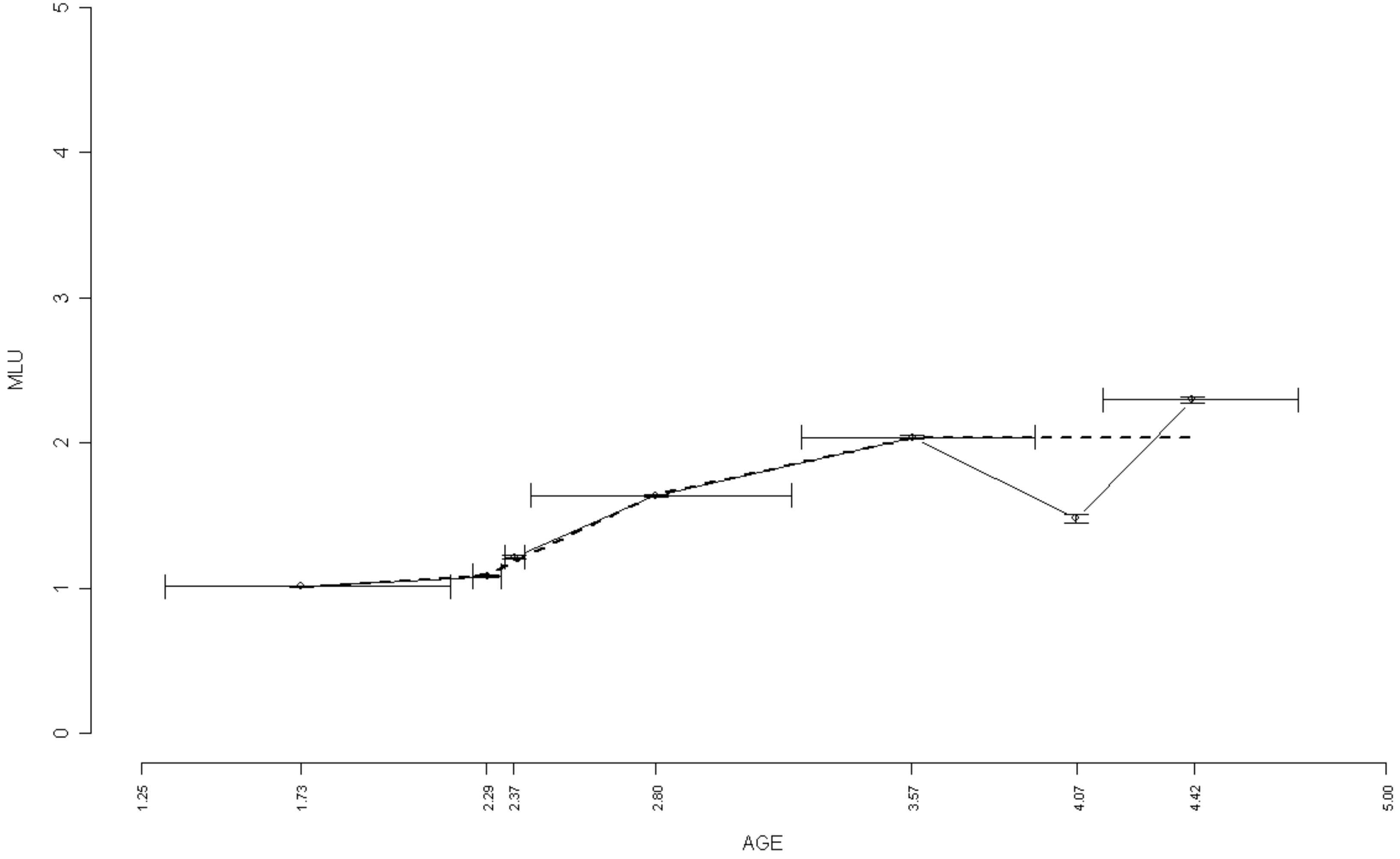
MLU stages: step 1



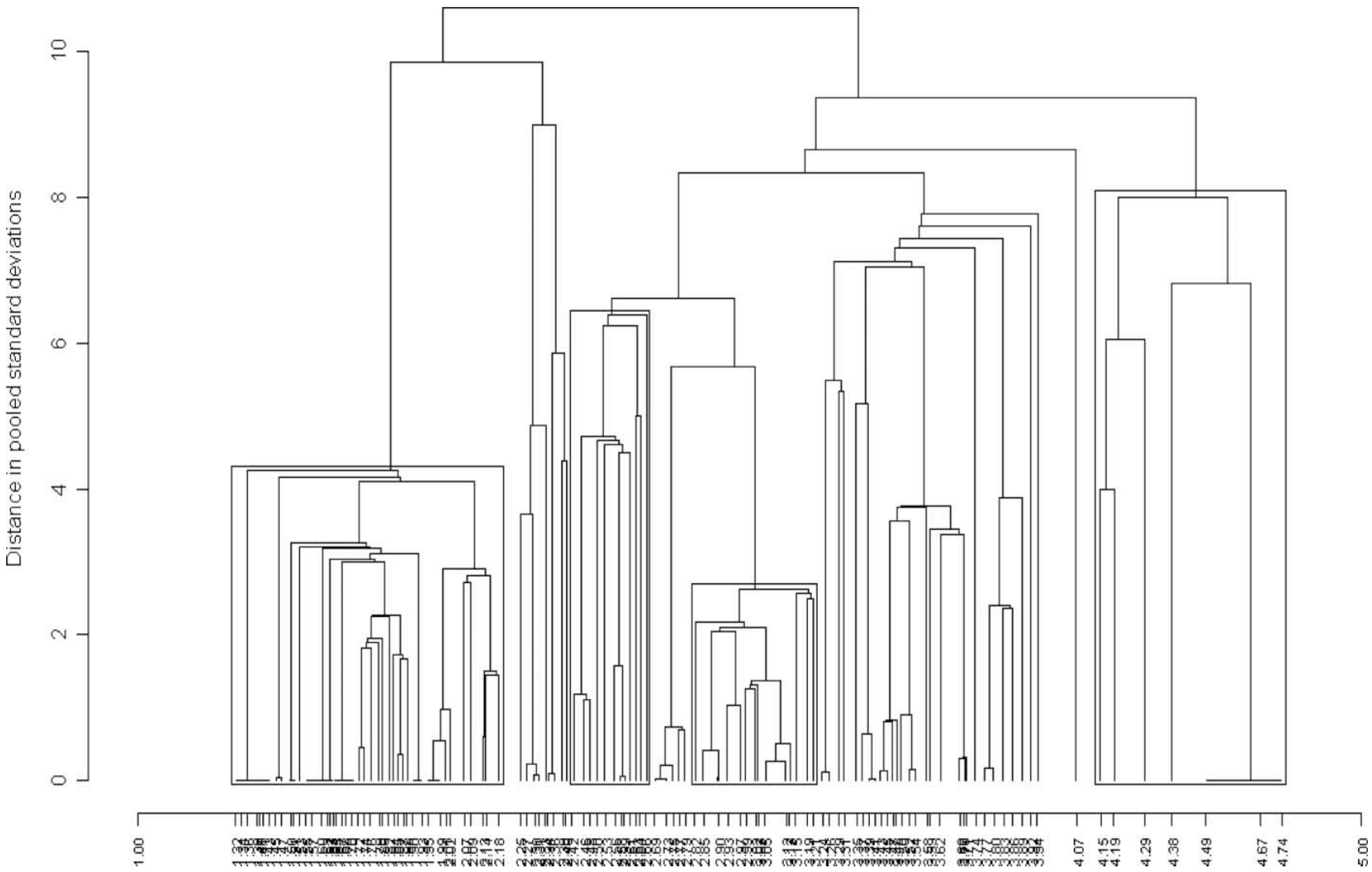
MLU stages: step 40



MLU stages: step 115



A dendrogram of the amalgamation of the 123 recordings of the child 1;03.26 and 4;09.30



Productivity – why does it matter?

Methodologically

- Unless they are clear errors and unattested in the adult language, utterances and accompanying morphology could have been rote-learned
- If its rote-learned, it should not be compared with a productive form in another language
- So we have to have ways of assessing productivity before we can make comparisons:

1. Using corpora
2. Introducing novel items and seeing what the child does with them
3. Comparing children's comprehension and production of utterances with contrasting forms
4. Modelling to see if productivity can develop as a function of input

Theoretically

Full competence model:

Children bring abstract linguistic categories to the learning of language

Why do children make errors?

- performance limitations
- late maturation
- language-specific features

Predictions:

- ➢Early abstract knowledge
- ➢Few errors with forms that they already know
- ➢Rapid development
- ➢Relatively minimal role of input

Constructivist model:

Children build abstract categories as they learn languages

Why do children make errors?

- Their representations are initially ‘low-scope’ and ‘item-based
- They extend forms they know to the wrong contexts
- They have mis-analysed the input

Predictions:

- ➢Item-based generalisations
- ➢Limited productivity, even with forms they know, at younger ages
- ➢Piecemeal development
- ➢Important role of input

Productivity

- Definition: appropriate and adult like use of a grammatical feature outside the scope of rote learned linguistic and extralinguistic contexts.
- How to measure productivity?
  - Brown (1973) 90% correct use in appropriate contexts
  - Number of different forms
  - Errors of commission and omission (overgeneralisations)
  - Comparisons with adult usage
    - semantic
    - pragmatic
    - frequency correlations
    - flexibility measures (e.g. entropy)

Level of analysis

- We start by counting at the level of specific form and string:
  - *is/are*
  - *I'm X-ing/You're Y-ing*
  - *What do X?/What can X?*
- We only count at a more abstract level, when there is evidence for it
- We do not credit the child with pre-given, abstract linguistic categories from the outset

Morphology

- Children often produce correct morphology very early
- Errors can seem relatively rare
- Often cited as evidence for underlying linguistic knowledge of inflection (tense, agreement etc.)

Spanish verb inflections

[Aguado Orea, 2004]

Nottingham corpus

- Lucia: 22 hours: 2;2.25 – 2;7.14
- Juan: 31 hours: 1;1-.21 – 2;5.28

Productivity: Spanish verb inflections

- Only verbs used by both adult and child
  - stem
  - agreement properties
- Adult sample of verb tokens randomly reduced to number found in child's speech

Aguado-Orea, PhD.
Krajewski et al. (submitted)

Number of inflections per stem

- No significant difference between parents
- Significant difference between children and parents at both tested ages
- For Juan, significant difference between first and second half of the corpus

Rates of Subject-Verb Agreement Error in Juan and Lucía’s data broken down by Person and Number

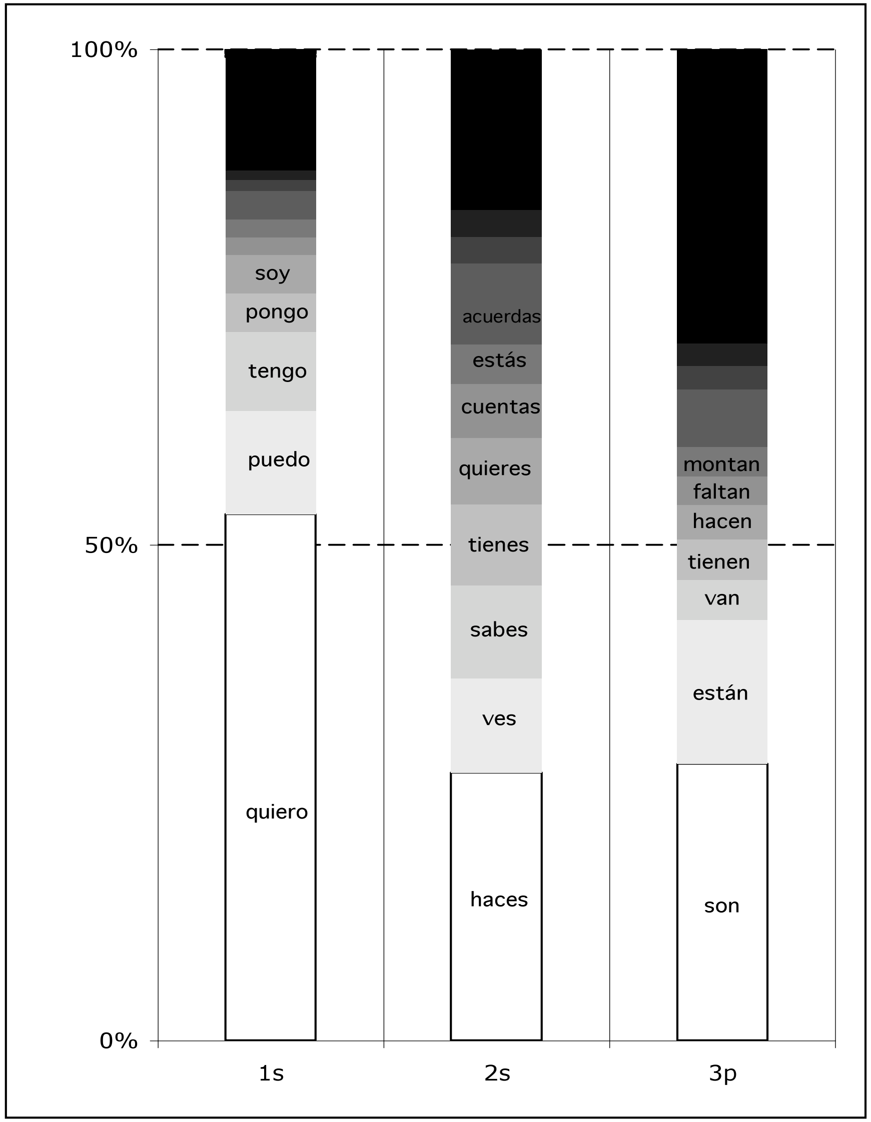
	Juan		Lucía	
Inflection	Contexts	Error Rate	Contexts	Error Rate
Overall	3152	4.5%	1672	4.4%

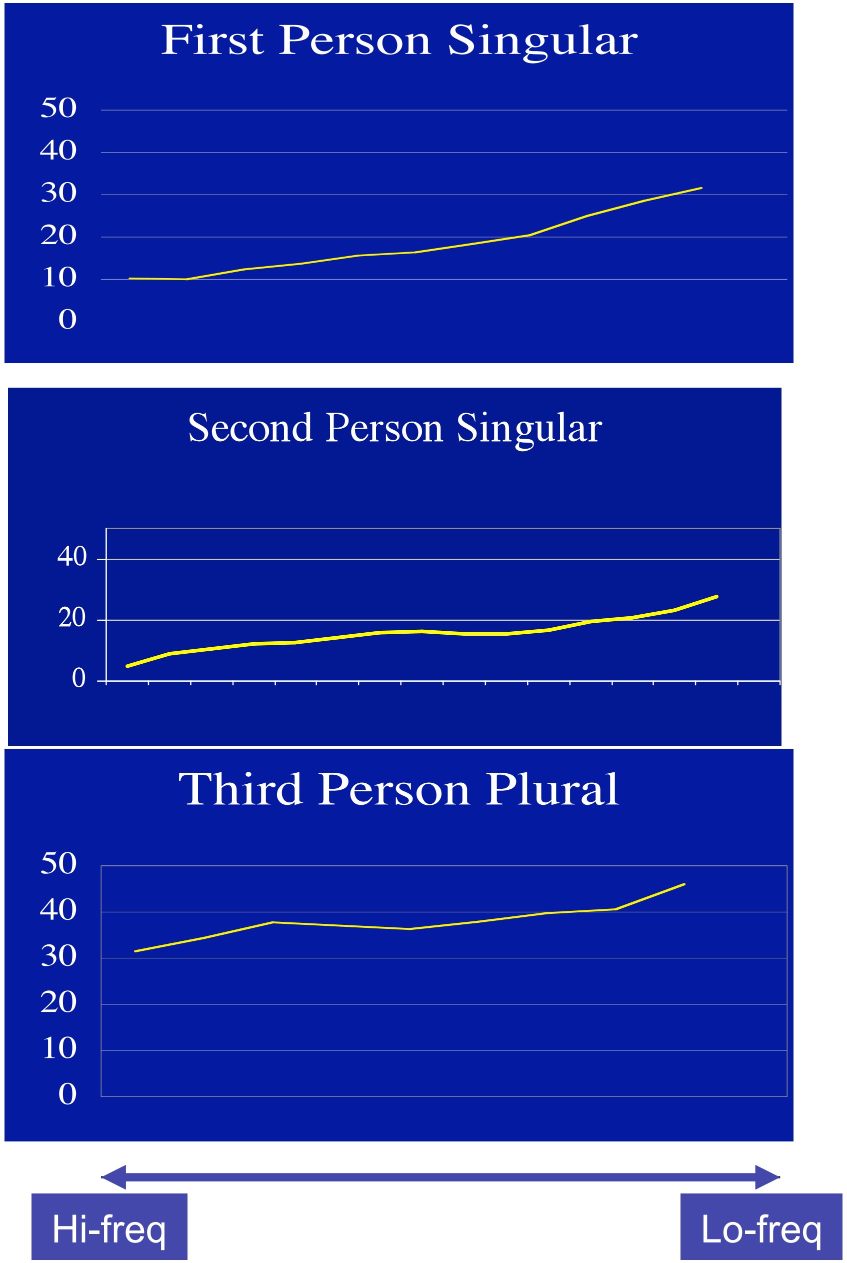
Rates of Subject-Verb Agreement Error in Juan and Lucía's data broken down by Person and Number

	Juan		Lucía	
Inflection	Contexts	Error Rate	Contexts	Error Rate
Overall	3152	4.5%	1672	4.4%
1sg	693	4.9%	496	3.0%
2sg	147	10.2%	96	22.9%
3sg	1997	0.7%	1018	0.5%
1pl	61	0	14	0
3pl	251	31.5%	48	66.7%

Pattern of error very similar across children (r = 0.99)
Not consistent with idea that errors can be disregarded as noise in the data

Rates of Subject-Verb Agreement Error in Juan's Data as a function of Lexical Frequency

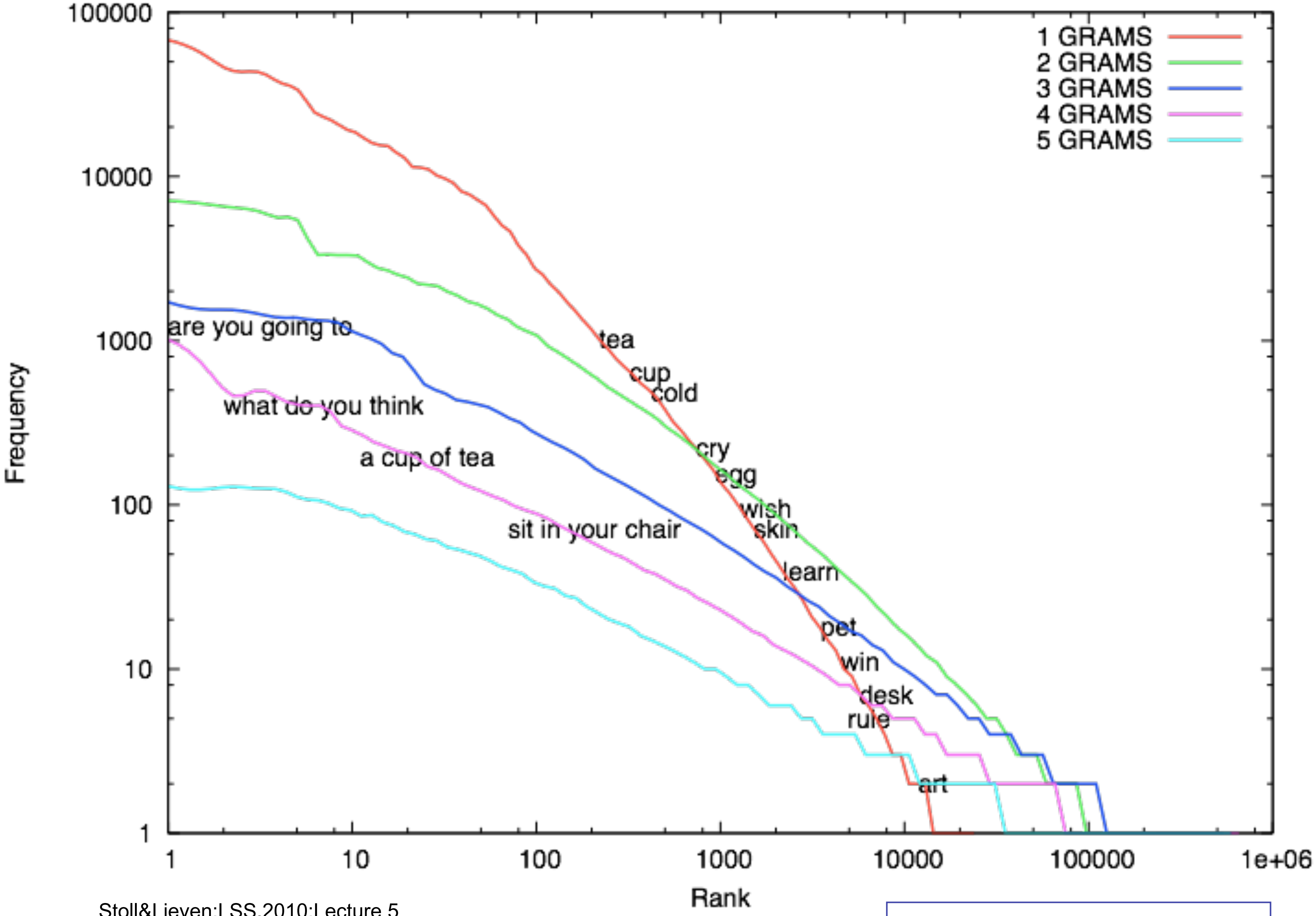




- Apparent sophistication of Spanish children’s use of verb inflection an illusion
- Low overall error rate reflects
  - Children’s knowledge of a relatively small number of high frequency forms
  - Children’s use of most frequent inflection (3rd person singular) as a ‘default’ when they don’t know what to do

Extracting grammars from corpora

The input



Bannard & Matthews, 2008,

‘Trace back’

- 20-40% of children’s utterances at 2;0 are exact repeats of what they have said before
  - But less so with increasing mlu
- 40% -50% of their utterances at 2;0 only differ from a previous utterance by one operation of substitution into a slot
  - But wider range of slots and more PROCESS slots with increasing mlu
- The vast majority of these operations are substitutions of a string into a REFERENT slot:
  - **There’s X, Want my X, X on there, Where’s the X?**

Lieven, Salomo & Tomasello, 2009

‘Trace forward’: Extracting and testing grammars with corpora

Construct grammars purely from what children actually say and then testing how well these grammars do in accounting for a new set of utterances produced by the children.

Bannard, Lieven & Tomasello, 2009

Finding candidate analyses

1. For each utterance, find all other utterances that share lexical material
2. Extract a series of concrete signs, schemas and material filled by Xs:

   e.g. ***I see Mummy*** → ***I see X, I X, X see X, X see Mummy, X Mummy, I X Mummy***

3. Take all concrete strings extracted and repeat procedure, until no further alignments possible

→ Set of candidate analyses, often very large

Extracting a grammar

- The grammars are probabilistic context-free grammars which pair rewrite rules with their corresponding probabilities given the distributions found in the corpus.

- Categories then emerge with given probabilities

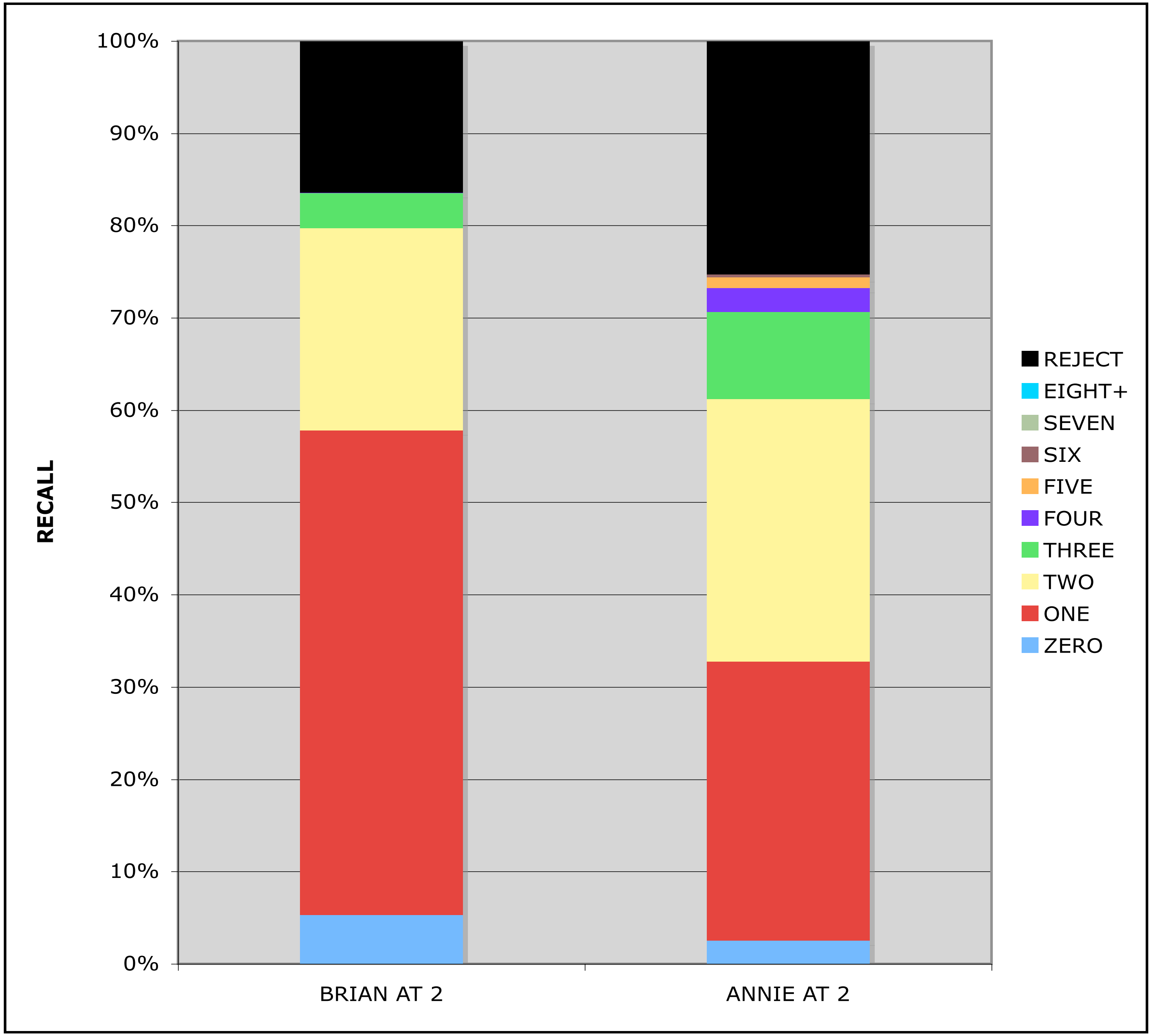
- Then use work on unsupervised grammar induction inspired by recent work in Bayesian grammar learning

- Sample probable grammars from a probability distribution over possible models

- Report the mean performance achieved when parsing with all these grammars

What proportion of the child's test utterances can our sampled grammars account for?

How complex are the analyses proposed?


RECALL
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
BRIAN AT 2
ANNIE AT 2
REJECT
EIGHT+
SEVEN
SIX
FIVE
FOUR
THREE
TWO
ONE
ZERO

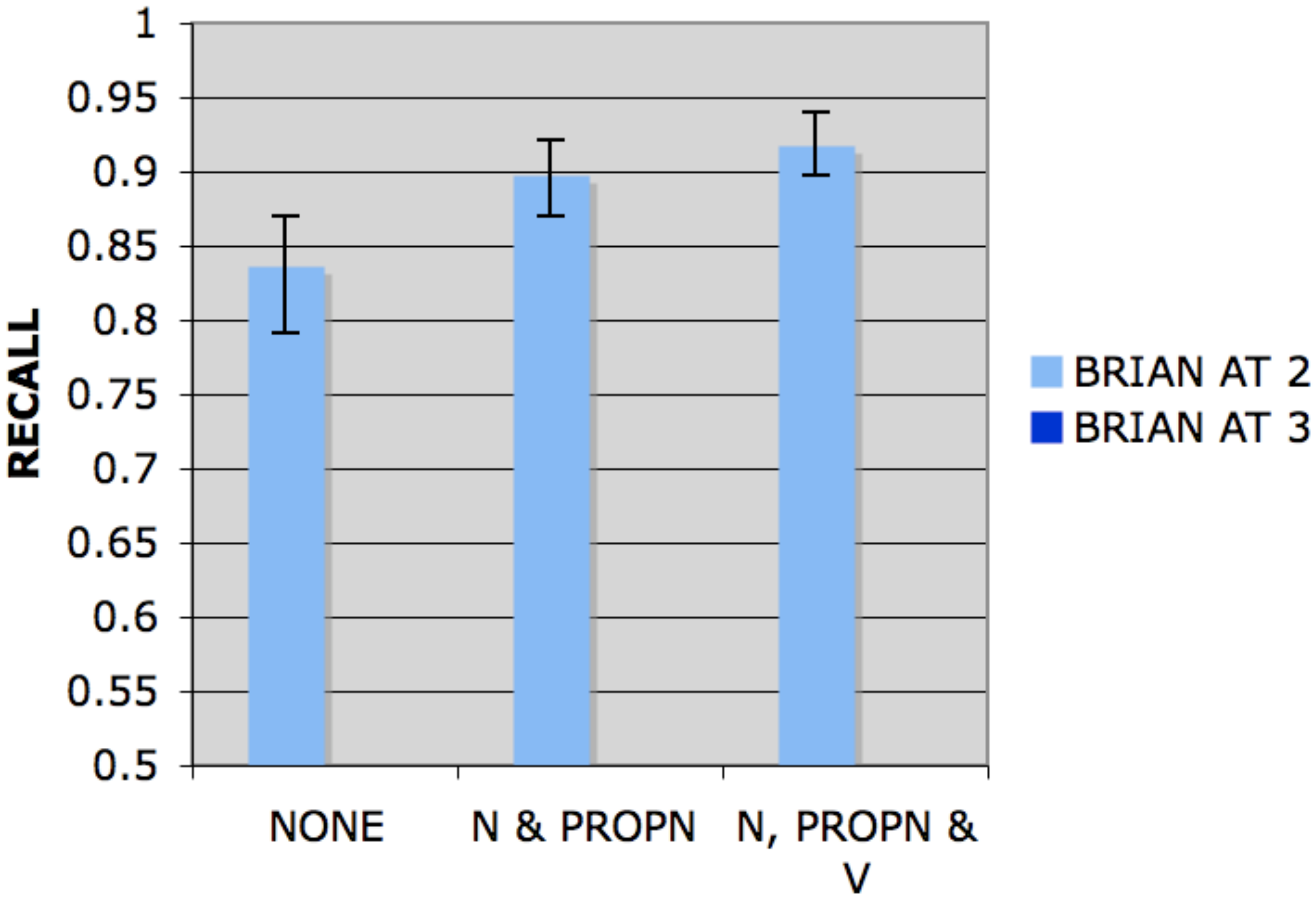
What about the utterances we couldn’t account for?

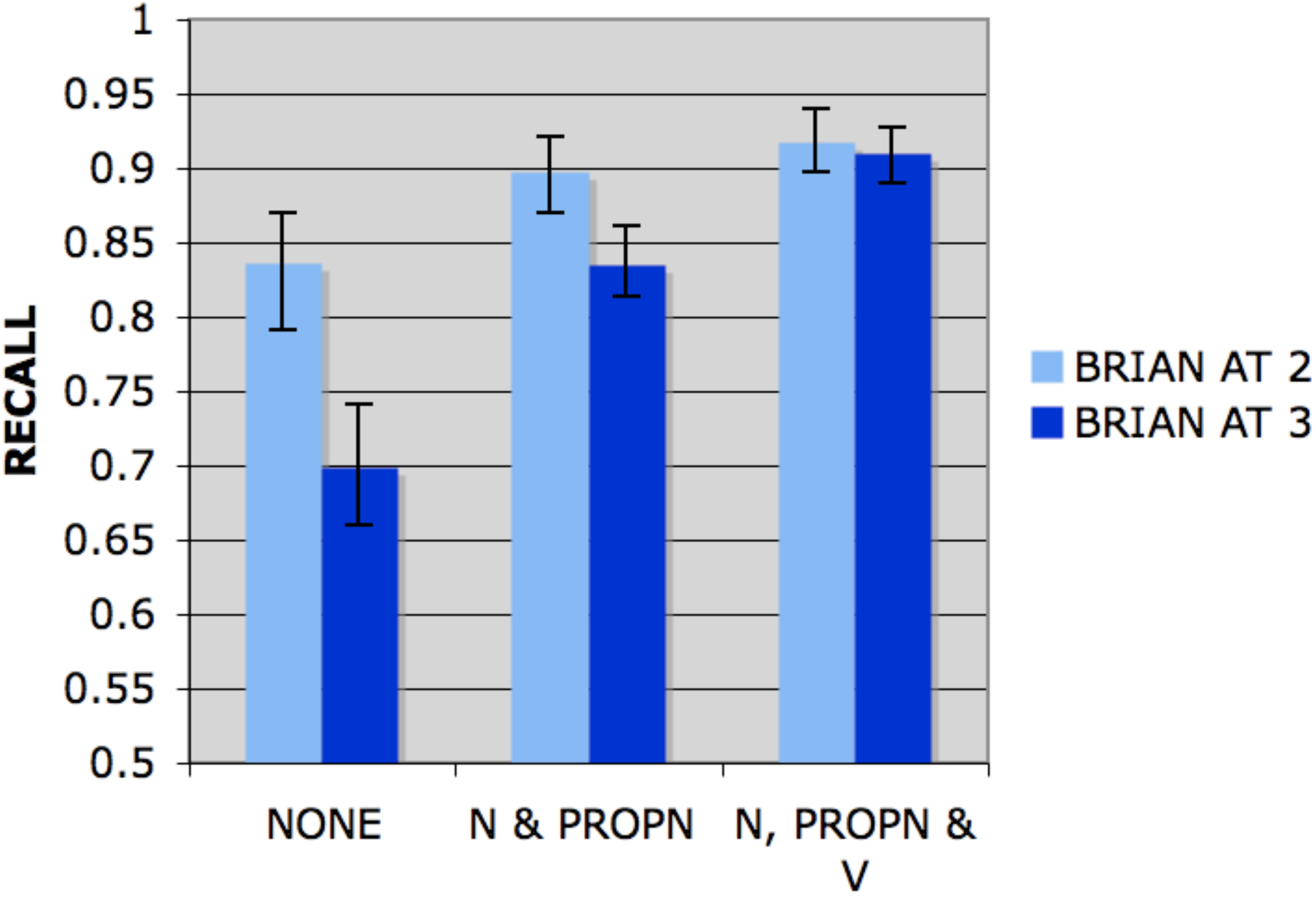
- Would positing abstract categories help to account for them?
- What if we replace each word in corpus with their categories (taken from human-tagged mor line) and induce grammar from this
- Does this improve coverage?

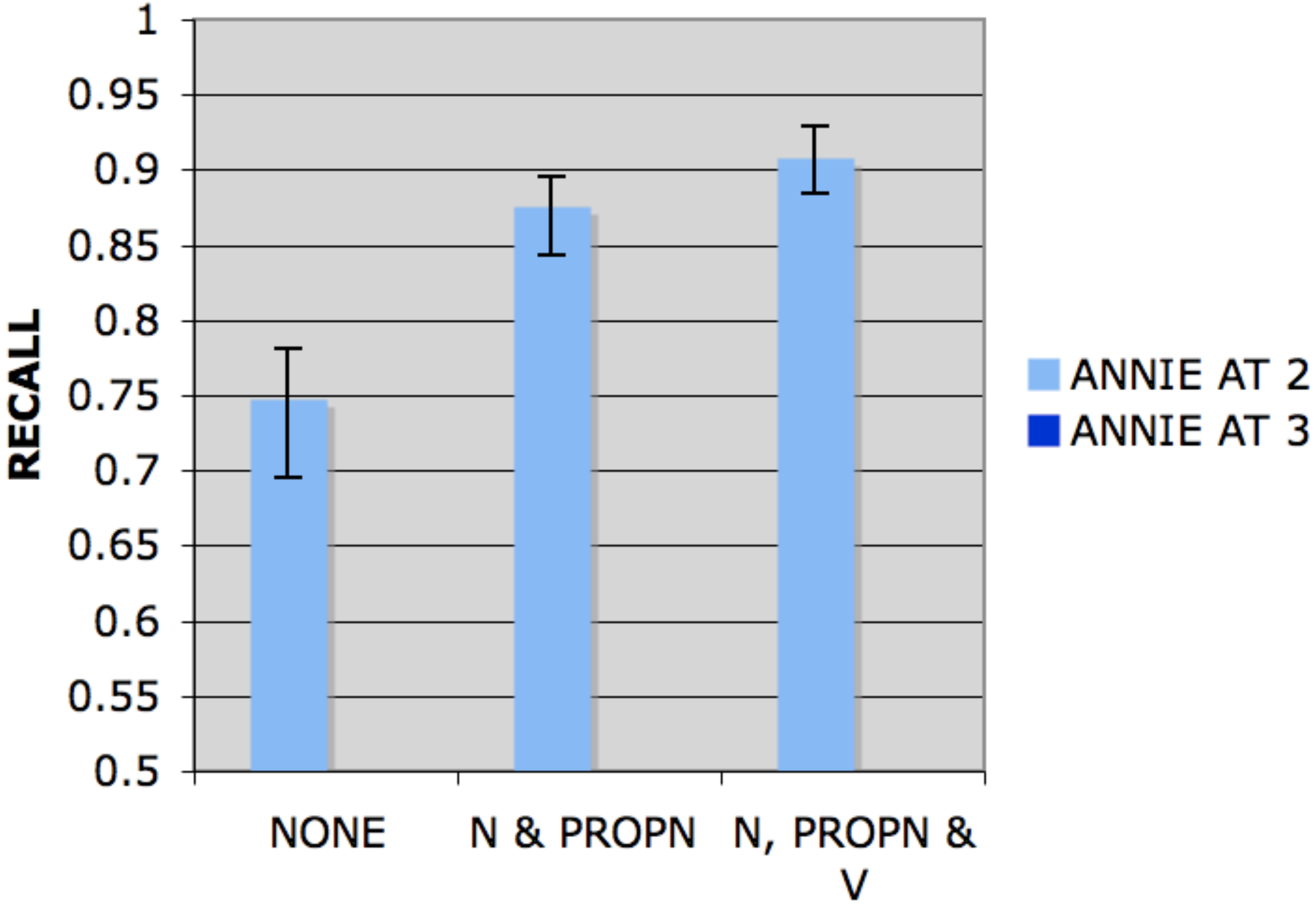
Replace nouns and proper nouns with categories:
Mummy needs a spoon = PropN *needs a* N

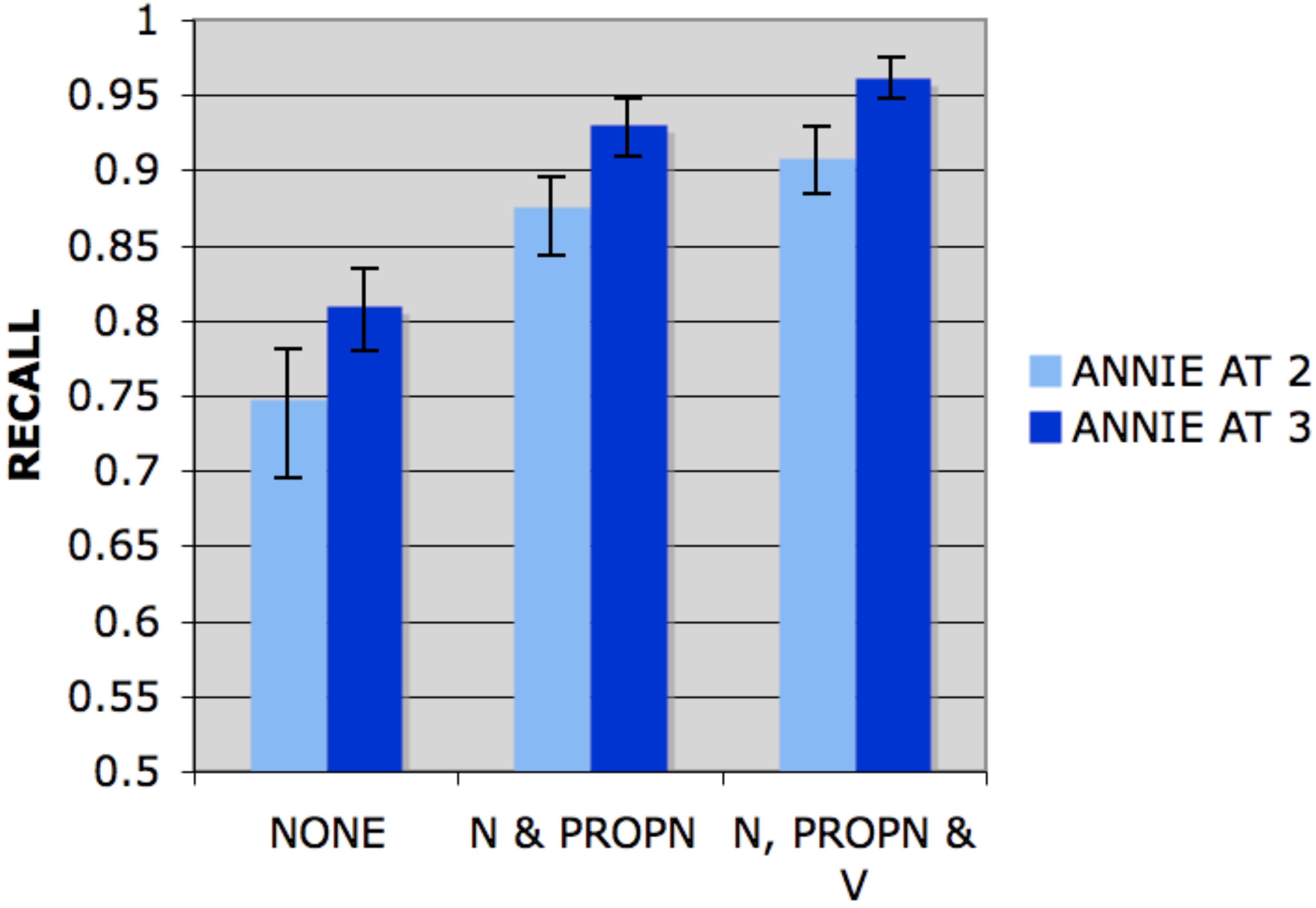
Replace verbs with category:
Mummy needs a spoon = PropN V *a* N

Then induce a grammar from this transformed corpus and parse test sentences in which replacements have also been made


RECALL
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
NONE
N & PROPN
N, PROPN & V
BRIAN AT 2
BRIAN AT 3


1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
RECALL
NONE
N & PROPN
N, PROPN & V
BRIAN AT 2
BRIAN AT 3


1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
RECALL
NONE
N & PROPN
N, PROPN & V
ANNIE AT 2
ANNIE AT 3


1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
RECALL
NONE
N & PROPN
N, PROPN & V
ANNIE AT 2
ANNIE AT 3

- Adding a noun category to our grammars has a large effect at age 2;0:
  - Consistent with claim that children at 2;0 have a nominal category (e.g. Tomasello & Olguin, 1993)

- Adding a verb category to our grammars has almost no effect at 2;0 but a large effect at 3;0 (for one of the children):
    - Consistent with claim that children do not generalize information across verbs at age 2;0 but some do at 3;0 (Olguin & Tomasello, 1993)

Can input account for errors?

- Pronoun case-marking errors in English
  - Him do it,
  - My want it
  - Him was crying

- Me-for-I errors
  - Me do it
  - Me doing it
  - Me find it

Me-for-I errors

- 17 children, longitudinal data
- 10 recordings from MLU ≥ 2.0, or from when produced NOM and ACC 1psg forms correctly if later
  - all *me-for-I* errors extracted
  - all other 1psg subjects (*I/my*) extracted.
  - % *me*-errors calculated

Kirjavainen et al., 2009

Input sample

- Input data from 5 recordings preceding child's sample
  - all complex sentences with *me+V* (non-fin) sequence extracted.
  - all *I+V* sequences extracted
  - % *me+V* sequences calculated
  - Total no. *me* and *I* in input calculated (all contexts)
  - % *me* calculated

Results

- Significant correlation between % *me-errors* in children’s speech and % *me+V* sequences in input (r (17) = 0.53, p = 0.029)

- Non-significant correlation between % *me-errors* and % *me* (*overall*) in the input (r (17) = 0.31, p = 0.23)

Lexical effects?

- Are the specific verbs in me+V sequences in the input produced in me+V errors in children’s speech?
- Children’s speech coded for:
  - Verb types in *me+V errors*
  - Control verb types in *I+V* sequences selected randomly from same files as errors

Input

- All input files prior to each target *me+V* / *I+V* verb type searched for *me+V* (complex sentences with non-fin verbs) sequences containing target verbs

- % *me+V* and *I+V* verb types found in *me+V input* sequences calculated.

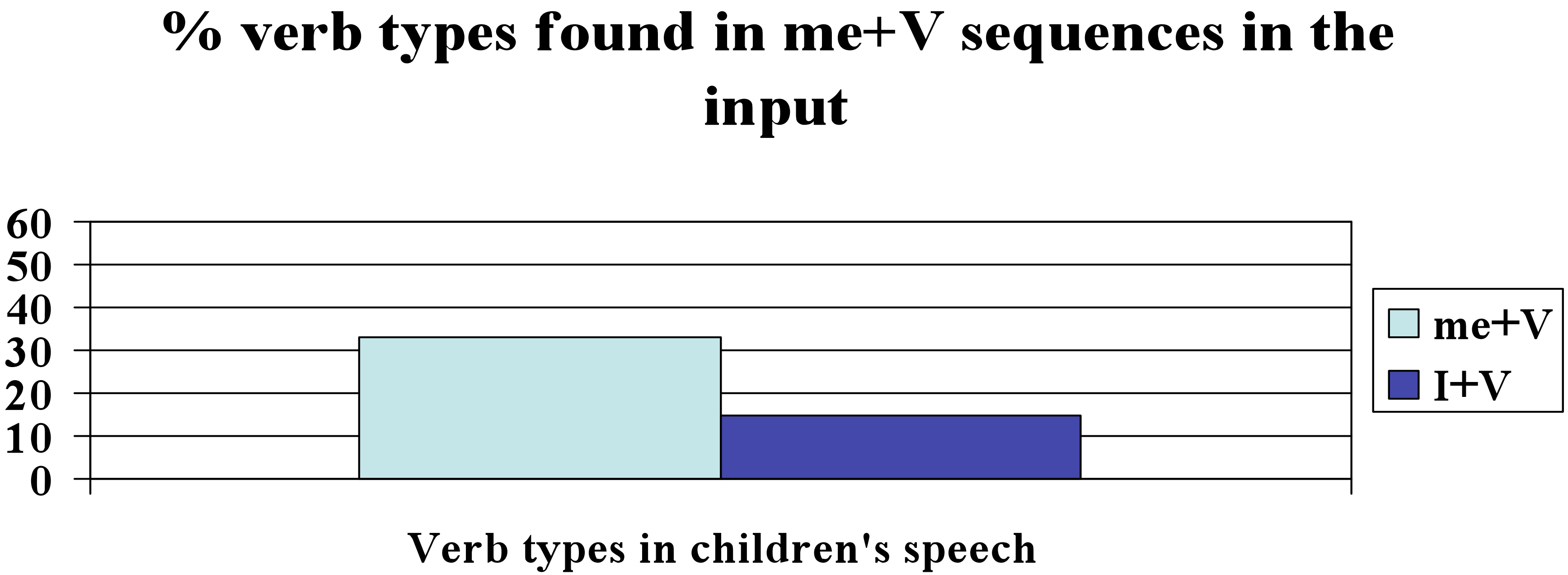
Errors and counterparts for Joel

Errors (me + verb)		Counterparts (I + verb)	
Me do it	(do)	*I found*	(found)
Me fetch it	(fetch)	*I just turn around*	(turn)
Me sing	(sing)	*I fetch it*	(fetch)
Me have some	(have)	*I don’t know*	(don’t)
Me go there	(go)	*I want to sit there*	(want)
Me read it	(read)	*I didn’t put it in*	(didn’t)

Kirjavainen et al., 2009

Results

- 3 children excluded < 4 error types
- 14 children, Range 4 to 93 *me+V* types



(t (13) = 4.5, p < 0.001)

Summary

- Proportional *me+V error rate* related to proportional use of *complex me+V sequences* in input.

- Specific verbs in me+V errors more likely to appear in me+V sequences in the input than matched set of verbs from children's I +V sequences.

Some children have abstracted a productive pattern

- Ruth with 93 *me+verb* type errors (532 tokens)
- *me got, me want, me gonna*
- *me + finite verb*
    - *She’s found me, hasn’t she?,*
    - *Don’t hit me, cries the little bear*

Final comments

We need:

- more corpora
- of more languages
- denser corpora
- tagged
- parsed
- translated into English
- with video and/or context lines!!!

- Longitudinal corpora are essential for the study of children’s language acquisition
- When using corpora the issues of sampling, level of analysis and defining productivity are critical
- Modelling studies with corpora are an important source of evidence
- Experiments should be designed in the light of usage facts gained from corpora