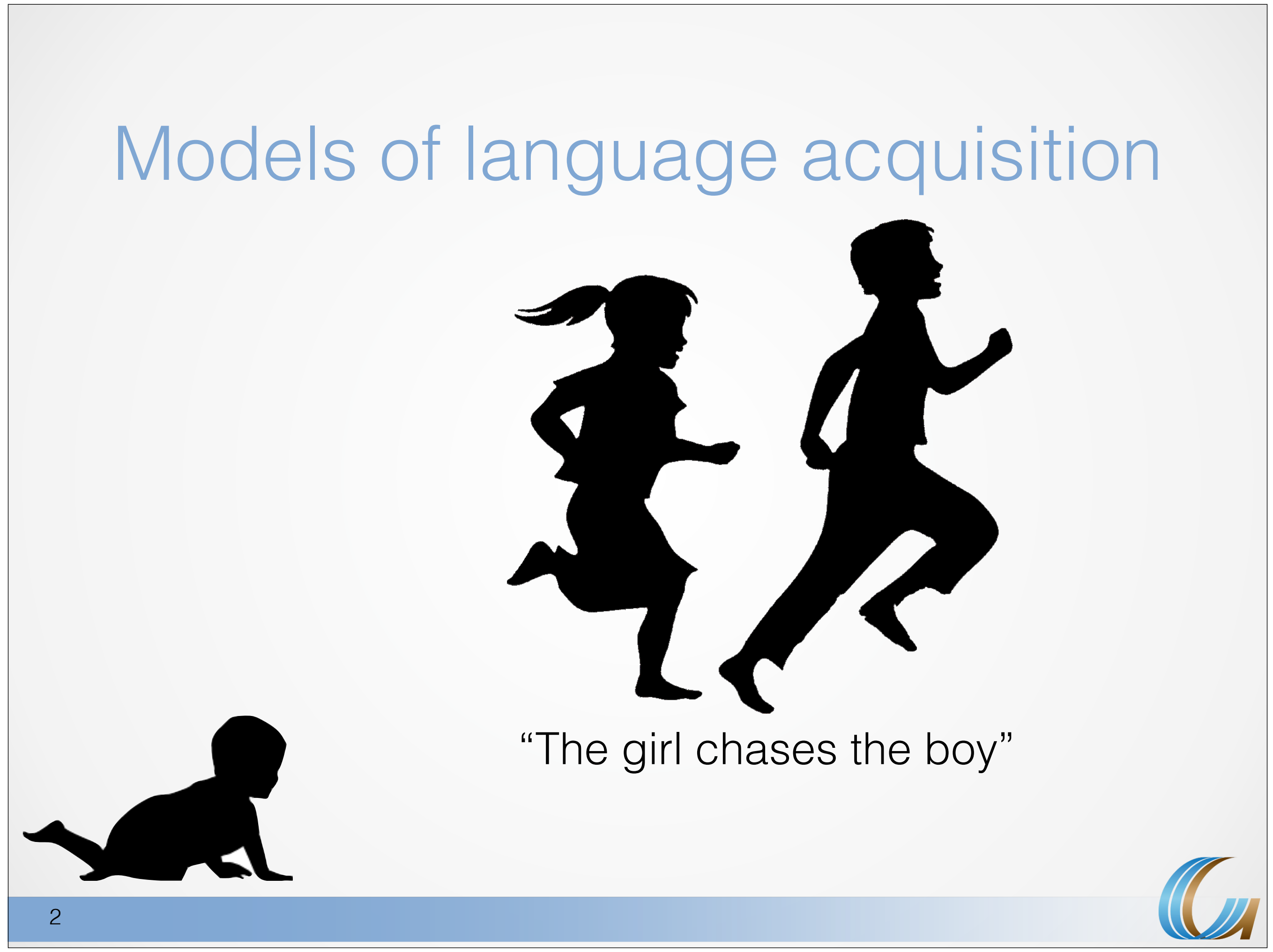
Exploring the assumptions of language acquisition models

Christos Christodoulopoulos, Cynthia Fisher and Dan Roth

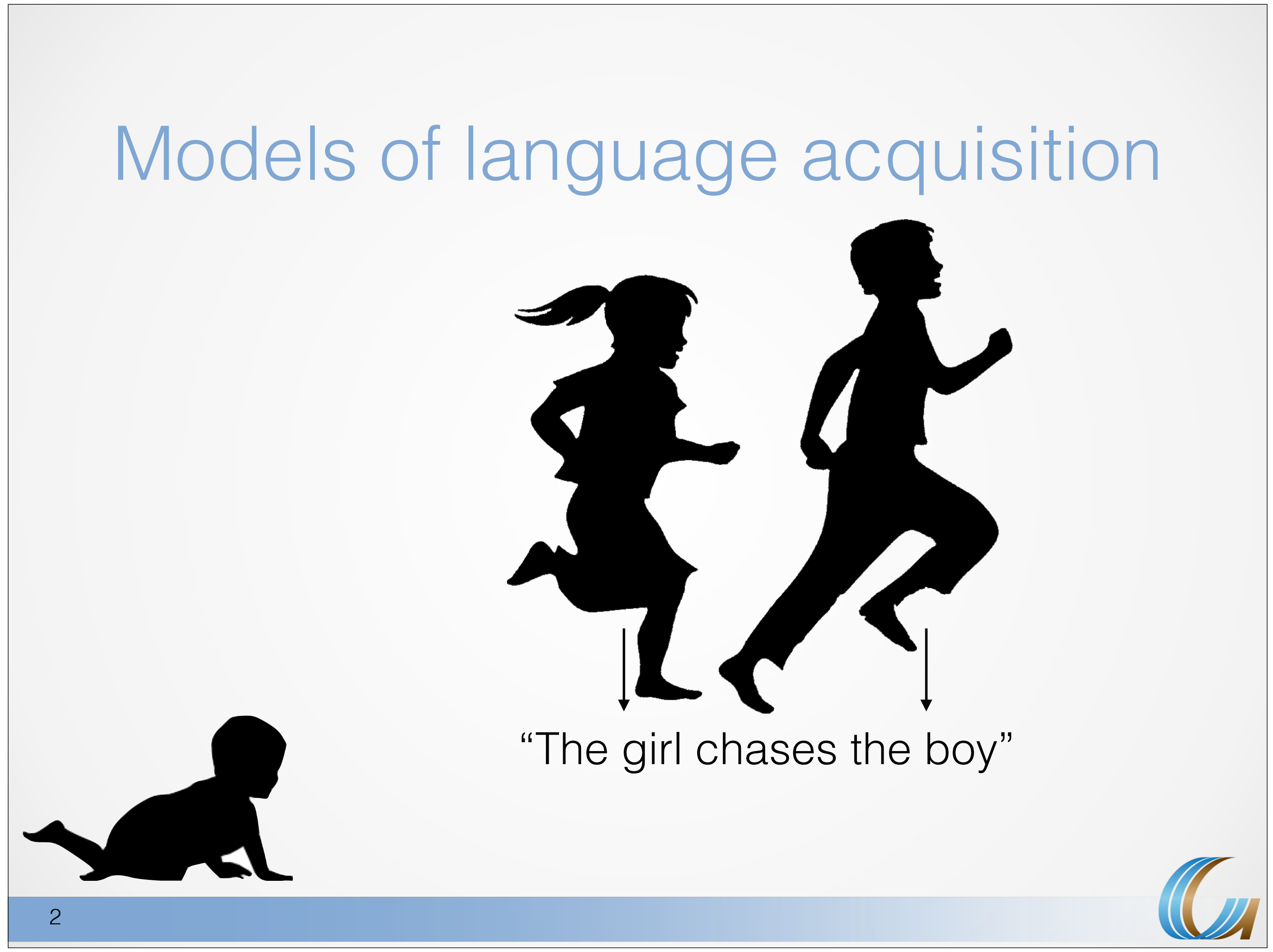
Midwest Speech & Language Days 2015

Cognitive Computation Group
University of Illinois at Urbana - Champaign



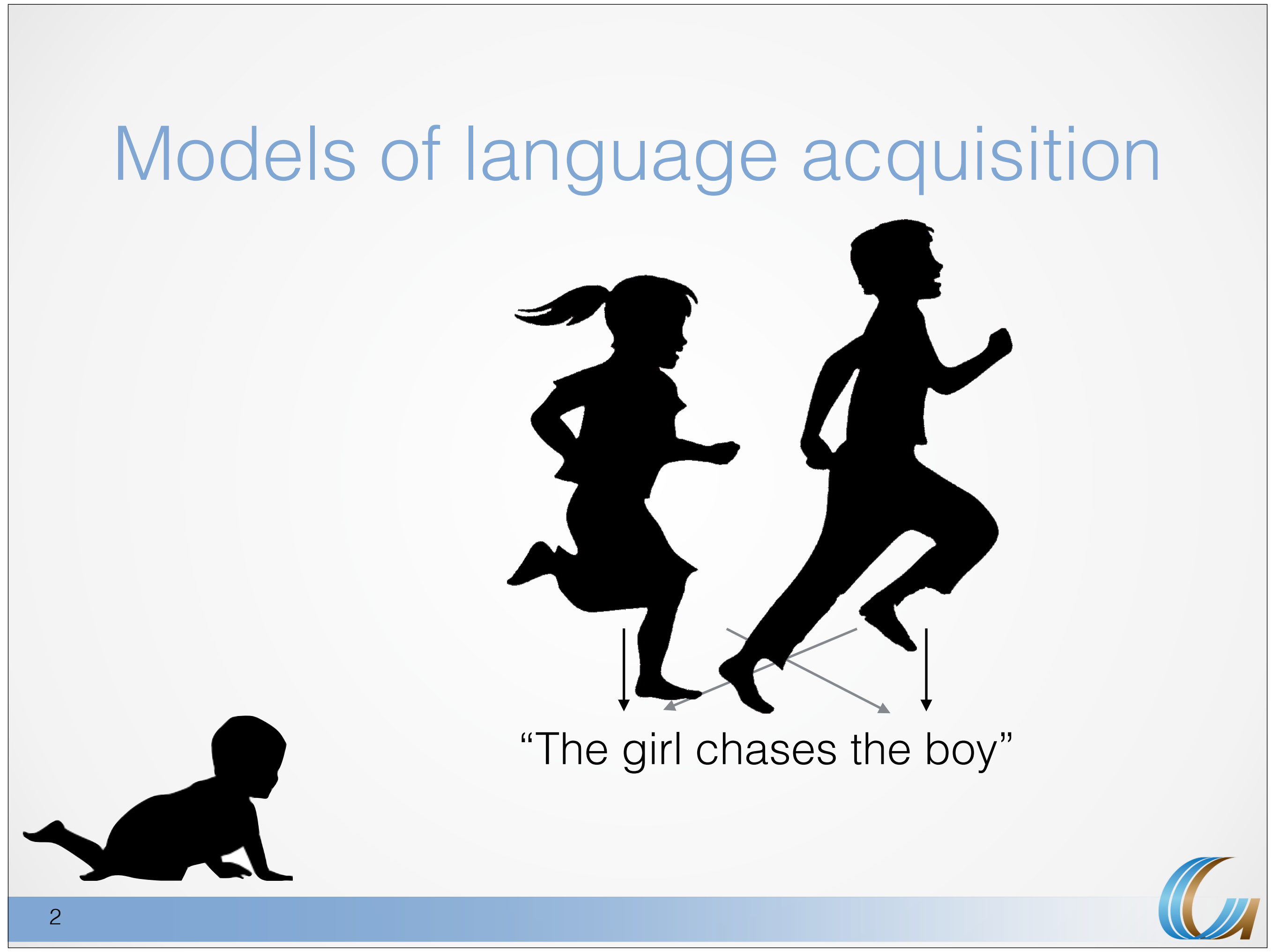
The task of language acquisition can be exemplified as follows:

- child observes a scene and hears an utterance
- it has to figure out that <SLIDE> “boy" refers to the boy and “girl” to the girl
- <SLIDE> and not the reverse
- but it also needs to somehow figure out what parts of the scene are described in the utterance, namely that the girl is chasing the boy and not <SLIDE> that boy is running



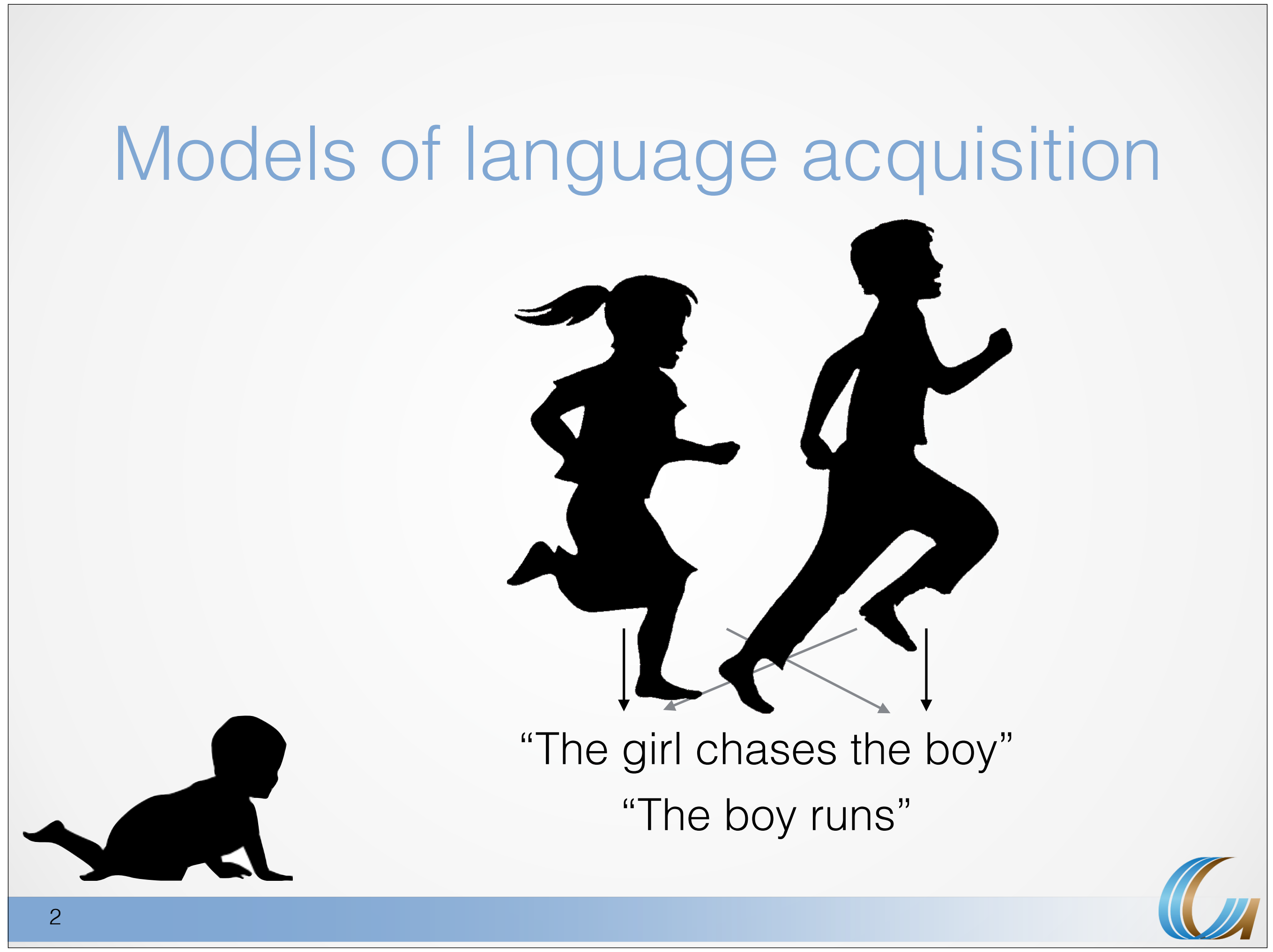
The task of language acquisition can be exemplified as follows:

- child observes a scene and hears an utterance
- it has to figure out that <SLIDE> “boy" refers to the boy and “girl” to the girl
- <SLIDE> and not the reverse
- but it also needs to somehow figure out what parts of the scene are described in the utterance, namely that the girl is chasing the boy and not <SLIDE> that boy is running



The task of language acquisition can be exemplified as follows:

- child observes a scene and hears an utterance
- it has to figure out that <SLIDE> “boy" refers to the boy and “girl” to the girl
- <SLIDE> and not the reverse
- but it also needs to somehow figure out what parts of the scene are described in the utterance, namely that the girl is chasing the boy and not <SLIDE> that boy is running



The task of language acquisition can be exemplified as follows:

- child observes a scene and hears an utterance
- it has to figure out that <SLIDE> “boy" refers to the boy and “girl” to the girl
- <SLIDE> and not the reverse
- but it also needs to somehow figure out what parts of the scene are described in the utterance, namely that the girl is chasing the boy and not <SLIDE> that boy is running

Semantic Role Labeling

PropBank corpus
[Palmer et al. 2005]

Core arguments:

- **A0** - Agent
- **A1** - Patient
- **A2** - Recipient
- …

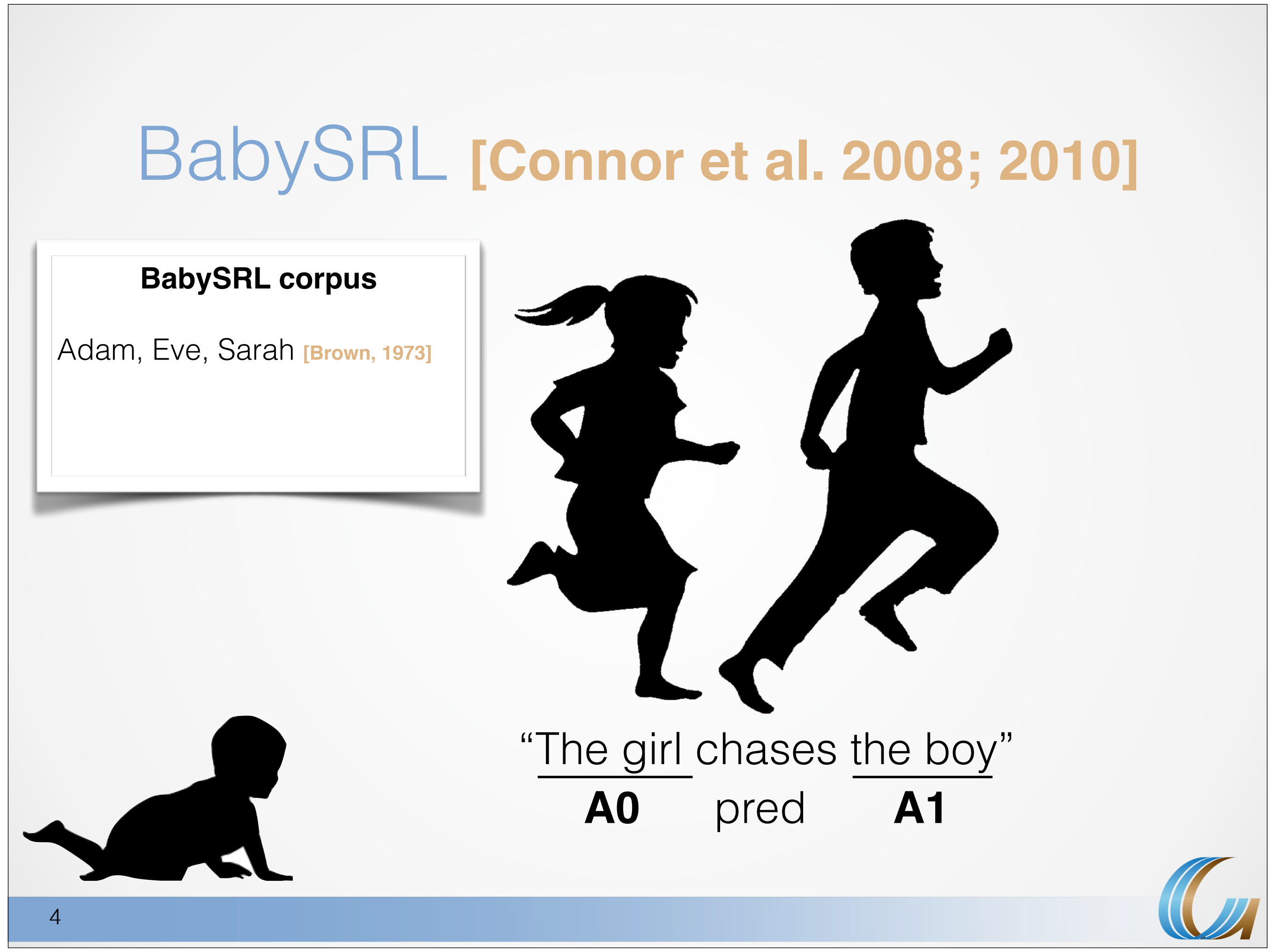
Modifiers:

- Locative
- Temporal
- Manner
- …

“The girl chases the boy”

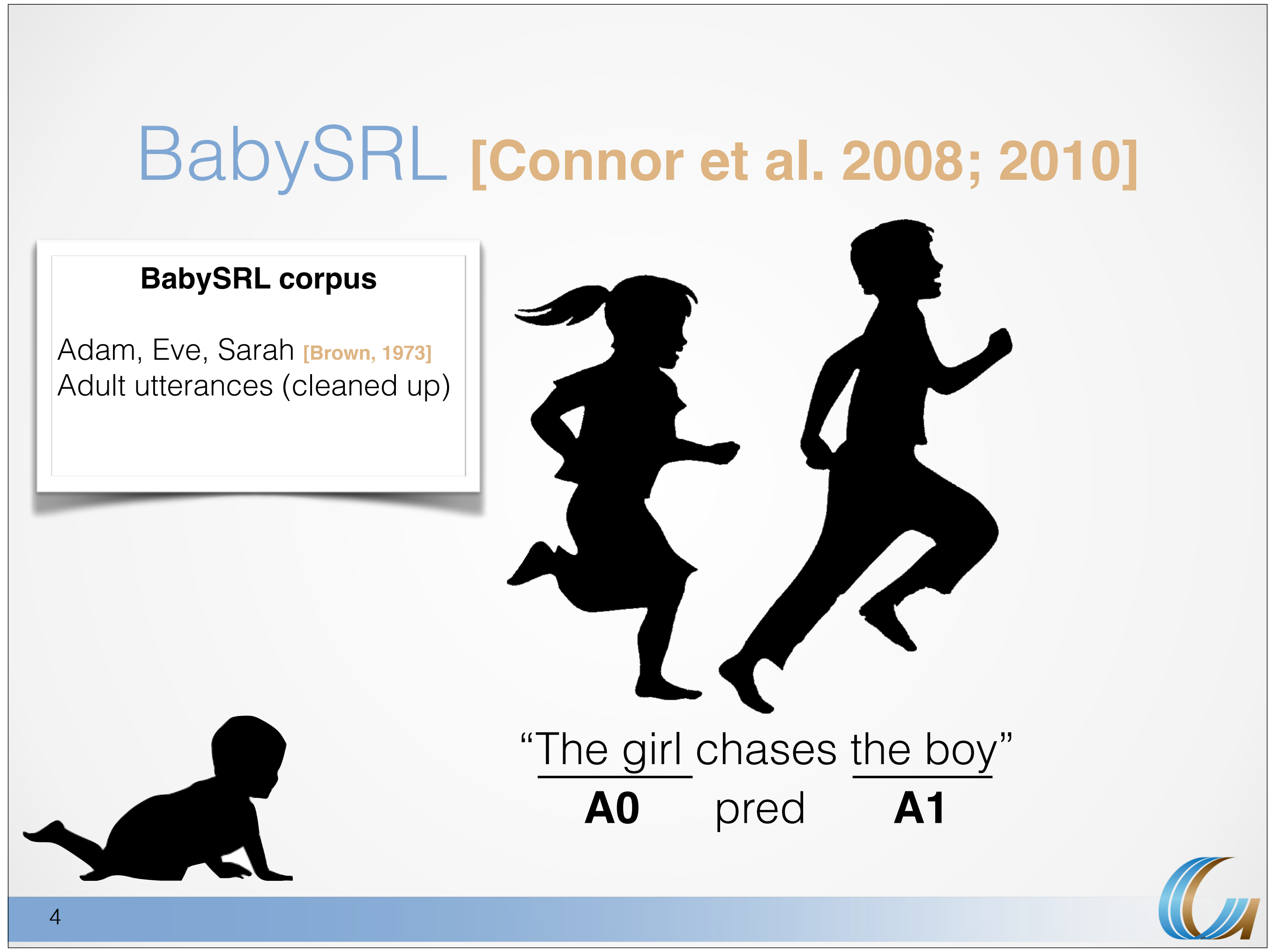
A0 pred **A1**

One way to formalise these interpretations is by using SRL, a shallow semantic representation, best described in the PropBank corpus annotation.
In there, numbered labels indicate “proto-roles” like Agent, Patient, Recipient etc. and also Modifiers, like Locative Temporal etc.



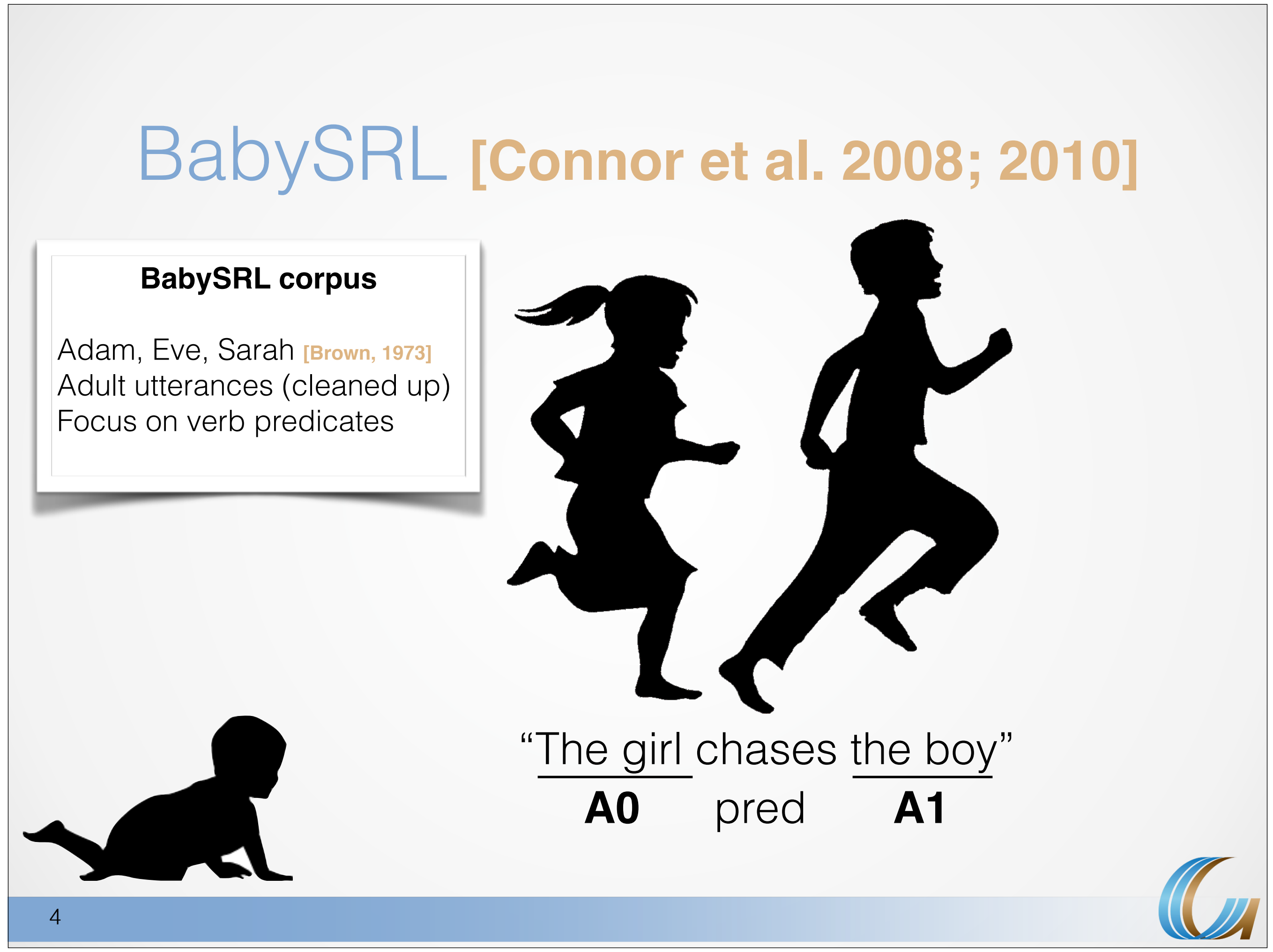
Our work is based on the BabySRL corpus, created by Mike Connor and others at Illinois. It uses part of the CHILDES corpus and <SLIDE> it provides an SRL annotation of the adult utterances —slightly cleaned up to avoid parsing errors— with the focus being on verb predicates and the prototypical proposition in one with 1 verb and 2 noun arguments but notice that that’s less than a 3rd of the total utterances, so it’s not an easy corpus.

So in our work we wanted to model this part of language acquisition and as a starting point we want to ask…



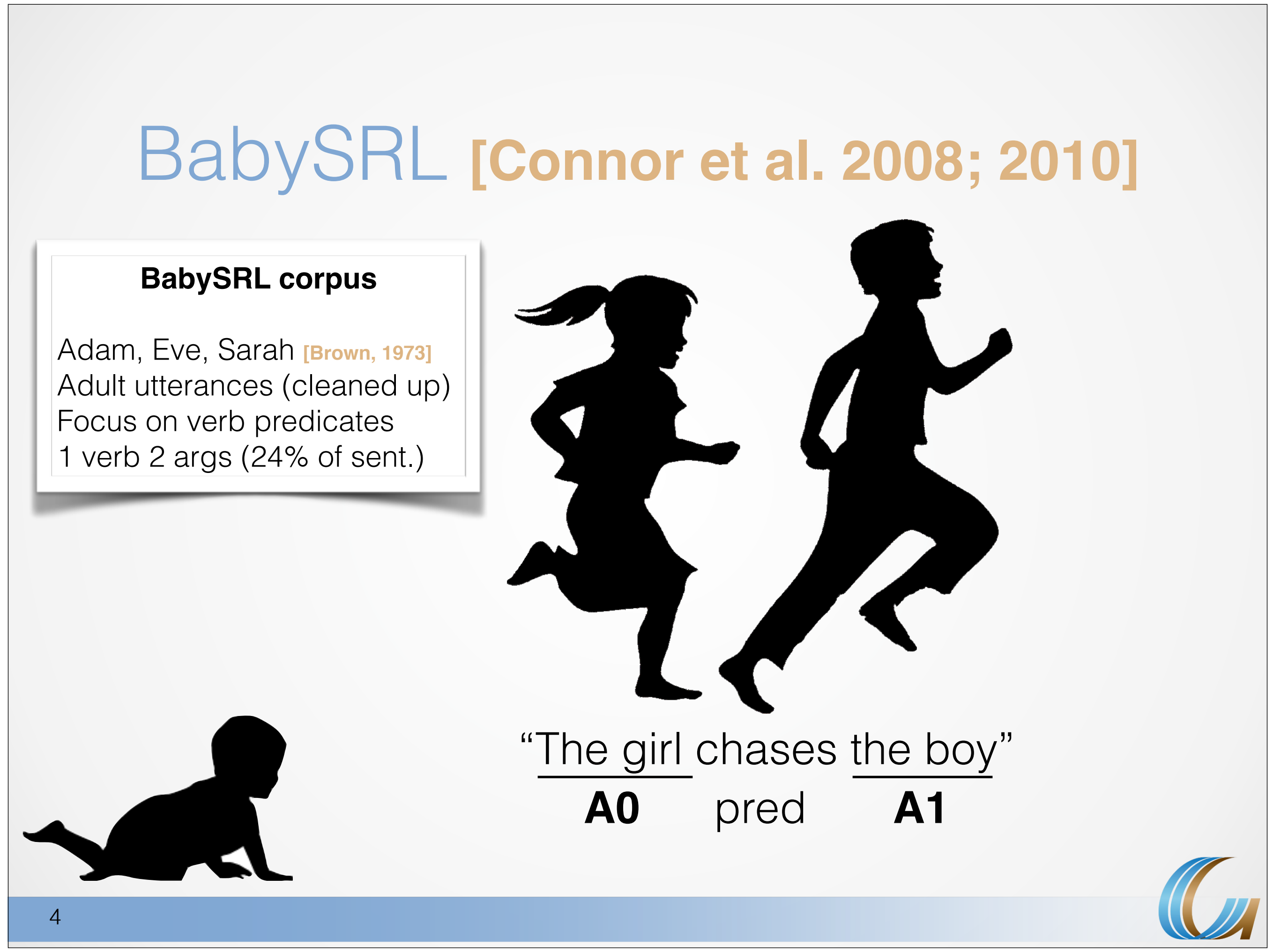
Our work is based on the BabySRL corpus, created by Mike Connor and others at Illinois. It uses part of the CHILDES corpus and <SLIDE> it provides an SRL annotation of the adult utterances —slightly cleaned up to avoid parsing errors— with the focus being on verb predicates and the prototypical proposition in one with 1 verb and 2 noun arguments but notice that that’s less than a 3rd of the total utterances, so it’s not an easy corpus.

So in our work we wanted to model this part of language acquisition and as a starting point we want to ask…



Our work is based on the BabySRL corpus, created by Mike Connor and others at Illinois. It uses part of the CHILDES corpus and <SLIDE> it provides an SRL annotation of the adult utterances —slightly cleaned up to avoid parsing errors— with the focus being on verb predicates and the prototypical proposition in one with 1 verb and 2 noun arguments but notice that that’s less than a 3rd of the total utterances, so it’s not an easy corpus.

So in our work we wanted to model this part of language acquisition and as a starting point we want to ask…



Our work is based on the BabySRL corpus, created by Mike Connor and others at Illinois. It uses part of the CHILDES corpus and <SLIDE> it provides an SRL annotation of the adult utterances —slightly cleaned up to avoid parsing errors— with the focus being on verb predicates and the prototypical proposition in one with 1 verb and 2 noun arguments but notice that that’s less than a 3rd of the total utterances, so it’s not an easy corpus.

So in our work we wanted to model this part of language acquisition and as a starting point we want to ask…

Experiment 1: Supervised learning

Given **perfect feedback**, do simple, **bottom-level** features capture anything useful about semantic roles/verb preferences?

Experiment 1: Supervised learning

- Supervised classifier (average perceptron)
  - LBJava [Rizzolo and Roth, 2010]
- Train on BabySRL corpus
- Test on novel verb sentences

The setup we have is based on a supervised classifier, implemented using LBJava (really cool project), trained on the BabySRL corpus and tested on “novel verbs” like <SLIDE> these, to mimic the experiments done with real infants (for the rest of the talk most of the results will be using the transitive case).

Experiment 1: Supervised learning

- Supervised classifier (average perceptron)
    - LBJava **[Rizzolo and Roth, 2010]**
- Train on BabySRL corpus
- Test on novel verb sentences

Intransitive: "The bunny **krads**"
Transitive: "The boy **krads** the girl"
Ditransitive: "The girl **krads** the boy a bunny"

The setup we have is based on a supervised classifier, implemented using LBJava (really cool project), trained on the BabySRL corpus and tested on "novel verbs" like <SLIDE> these, to mimic the experiments done with real infants (for the rest of the talk most of the results will be using the transitive case).

Experiment 1: Features

- **Most freq**uent label

A0 **A1**
The girl chases the boy

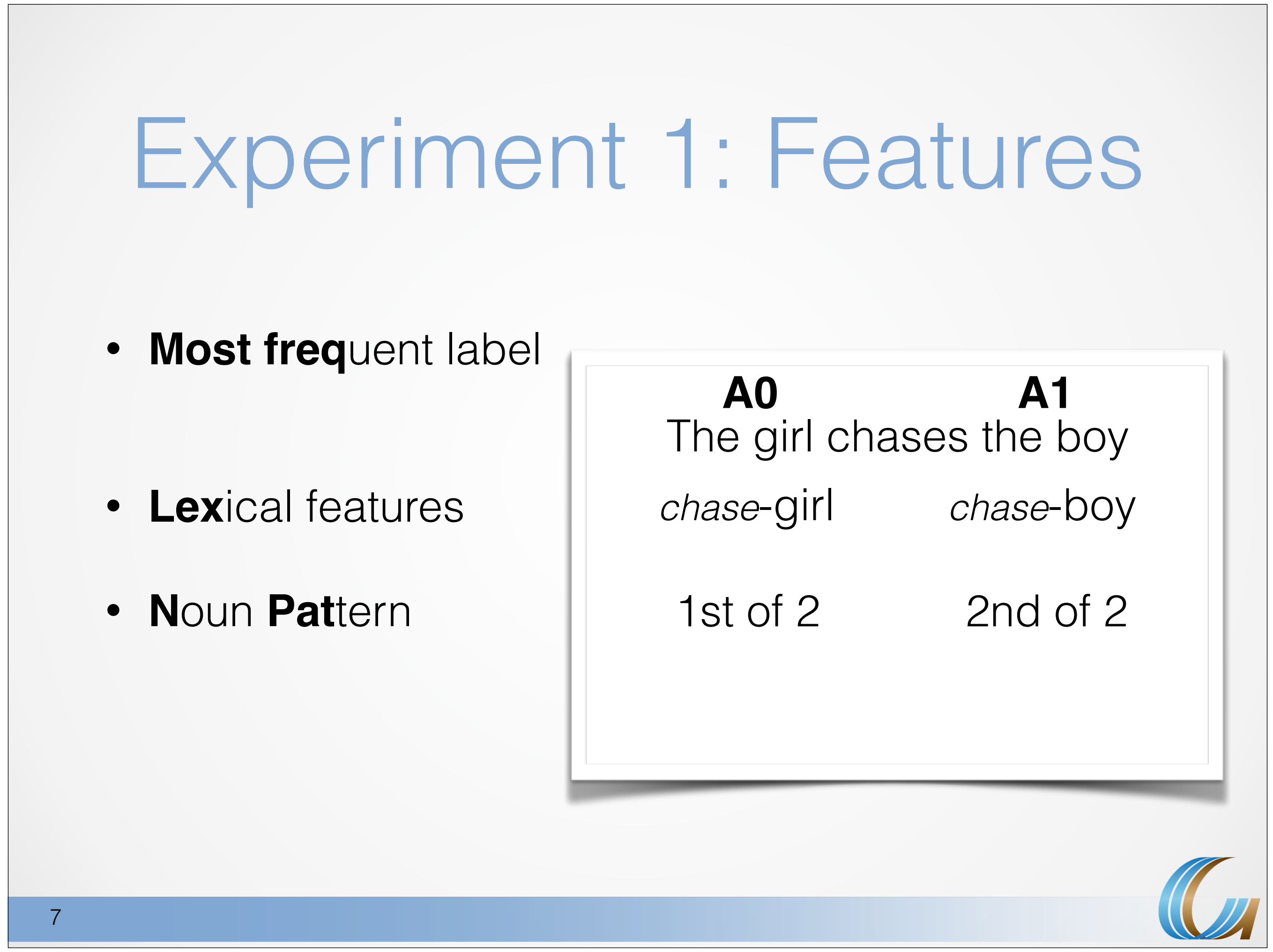
The features we use reflect really obvious phenomena at the utterance level. We start with the “most frequent label” baseline (what it says on the tin), then we have <SLIDE> the Lexical features, which are basically as concatenation of the predicate and the current argument glosses, then the <SLIDE> Noun Pattern features which simply encode the relative position of each noun and finally, <SLIDE> the Verb Position features which rely on us knowing where the predicate is and record the position of each argument relative to that predicate.

Experiment 1: Features

- **Most freq**uent label
- **Lex**ical features

A0 **A1**
The girl chases the boy
chase-girl *chase*-boy

The features we use reflect really obvious phenomena at the utterance level. We start with the “most frequent label” baseline (what it says on the tin), then we have <SLIDE> the Lexical features, which are basically as concatenation of the predicate and the current argument glosses, then the <SLIDE> Noun Pattern features which simply encode the relative position of each noun and finally, <SLIDE> the Verb Position features which rely on us knowing where the predicate is and record the position of each argument relative to that predicate.



The features we use reflect really obvious phenomena at the utterance level. We start with the “most frequent label” baseline (what it says on the tin), then we have <SLIDE> the Lexical features, which are basically as concatenation of the predicate and the current argument glosses, then the <SLIDE> Noun Pattern features which simply encode the relative position of each noun and finally, <SLIDE> the Verb Position features which rely on us knowing where the predicate is and record the position of each argument relative to that predicate.

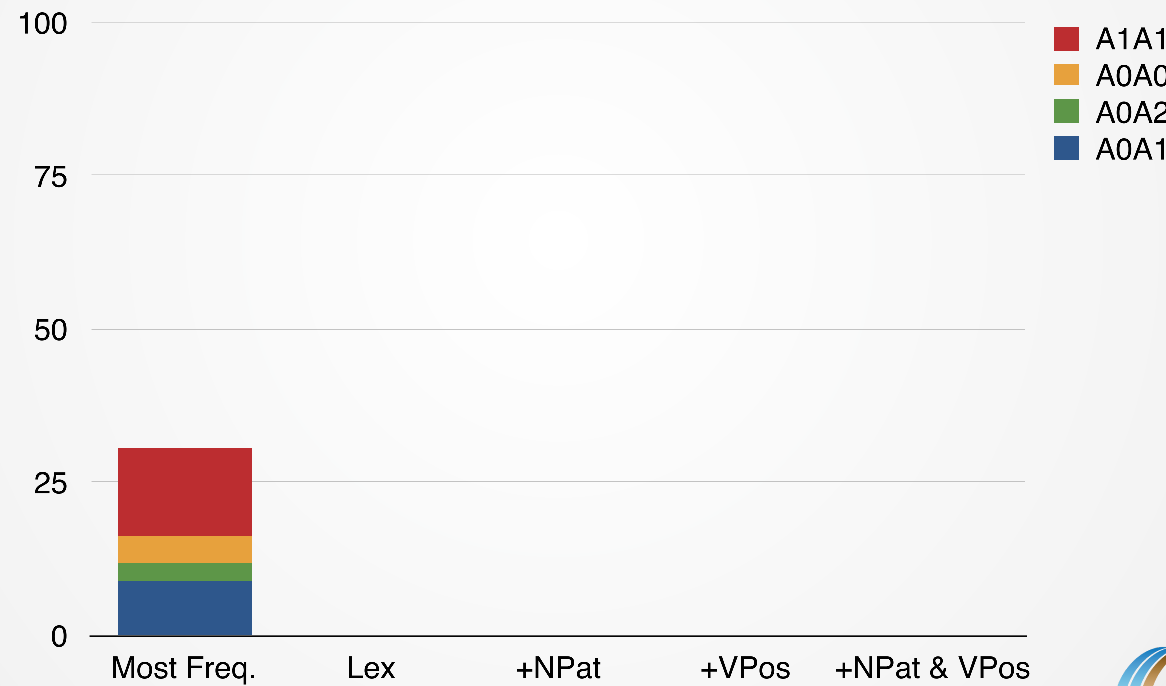
Experiment 1: Features

- **Most freq**uent label
- **Lex**ical features
- **N**oun **Pat**tern
- **V**erb **Pos**ition

	A0	**A1**
	The girl chases the boy	
Lexical features	*chase*-girl	*chase*-boy
Noun Pattern	1st of 2	2nd of 2
Verb Position	Before	After

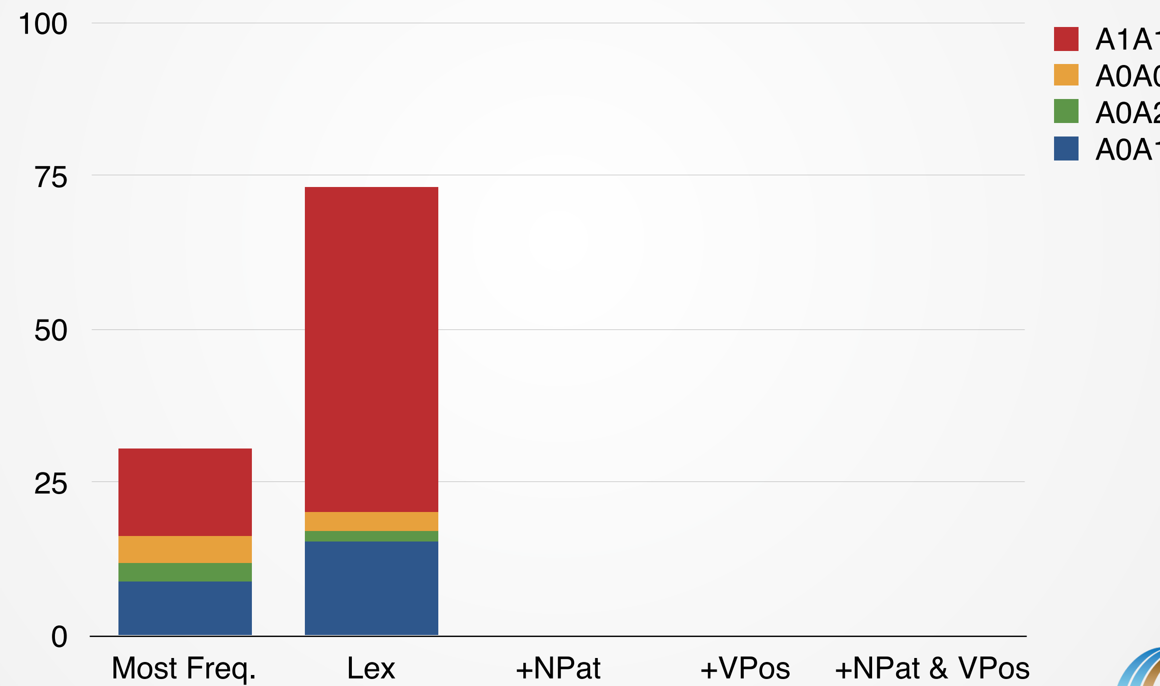
The features we use reflect really obvious phenomena at the utterance level. We start with the “most frequent label” baseline (what it says on the tin), then we have <SLIDE> the Lexical features, which are basically as concatenation of the predicate and the current argument glosses, then the <SLIDE> Noun Pattern features which simply encode the relative position of each noun and finally, <SLIDE> the Verb Position features which rely on us knowing where the predicate is and record the position of each argument relative to that predicate.

Experiment 1: Results



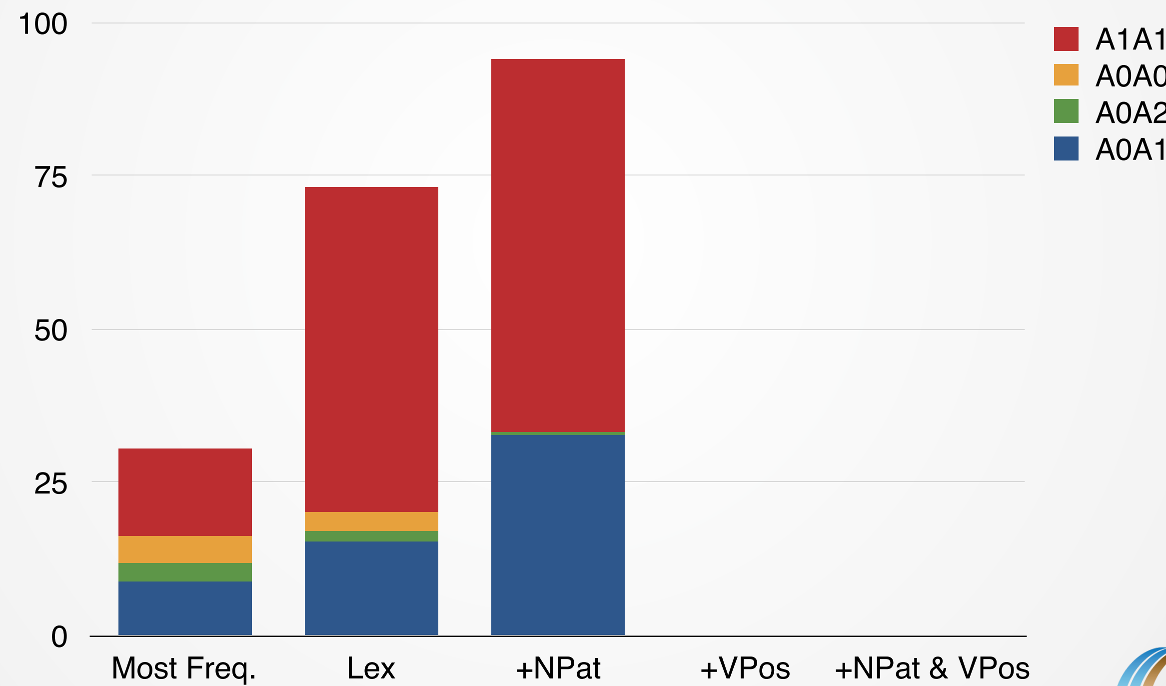
Here are the resulting patterns from each classifier. Here we're not necessarily testing the accuracy of each classifier; instead we want to know what it would predict given a sequence of nouns.

Experiment 1: Results



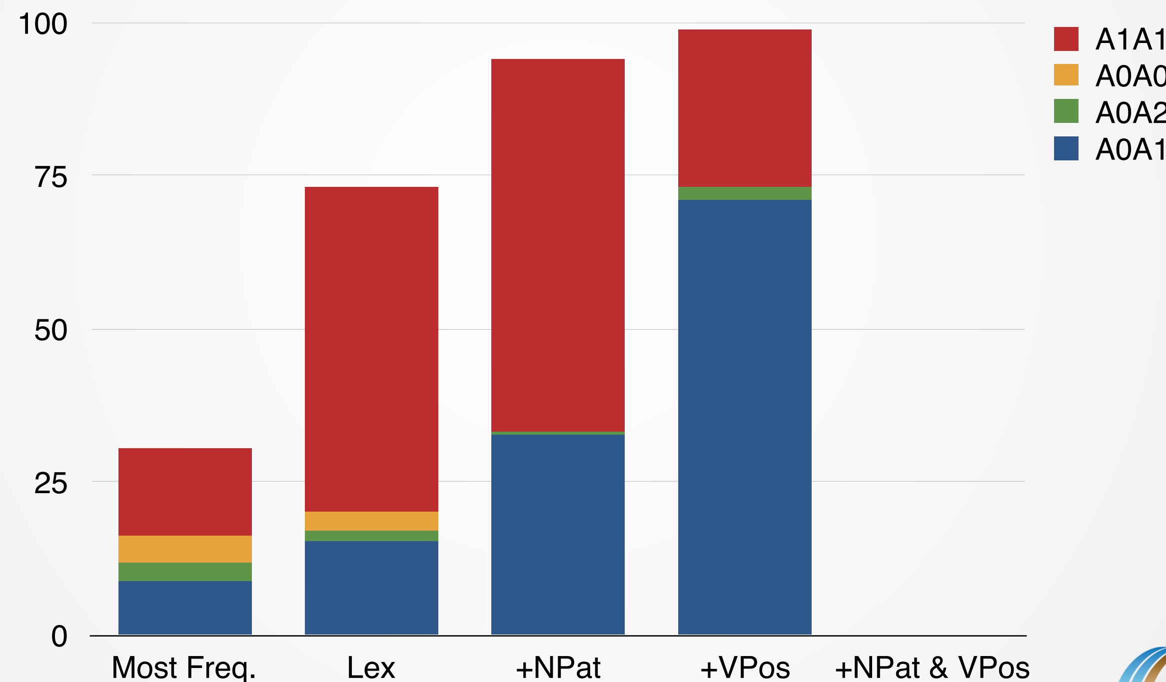
Here are the resulting patterns from each classifier. Here we're not necessarily testing the accuracy of each classifier; instead we want to know what it would predict given a sequence of nouns.

Experiment 1: Results



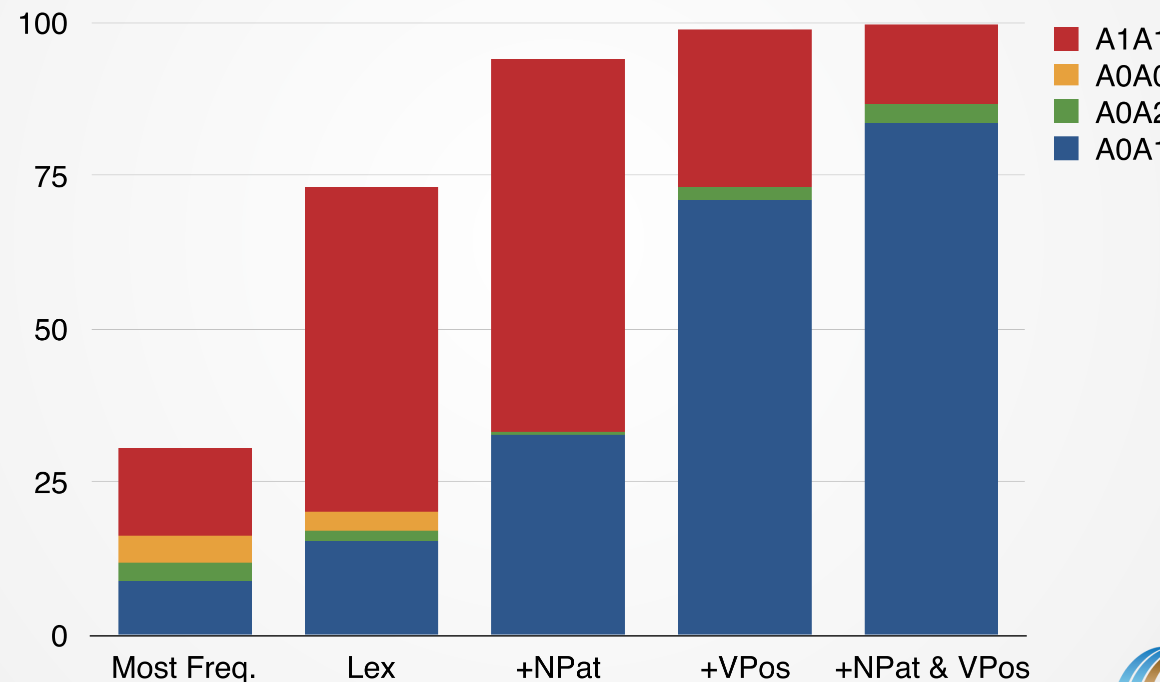
Here are the resulting patterns from each classifier. Here we're not necessarily testing the accuracy of each classifier; instead we want to know what it would predict given a sequence of nouns.

Experiment 1: Results



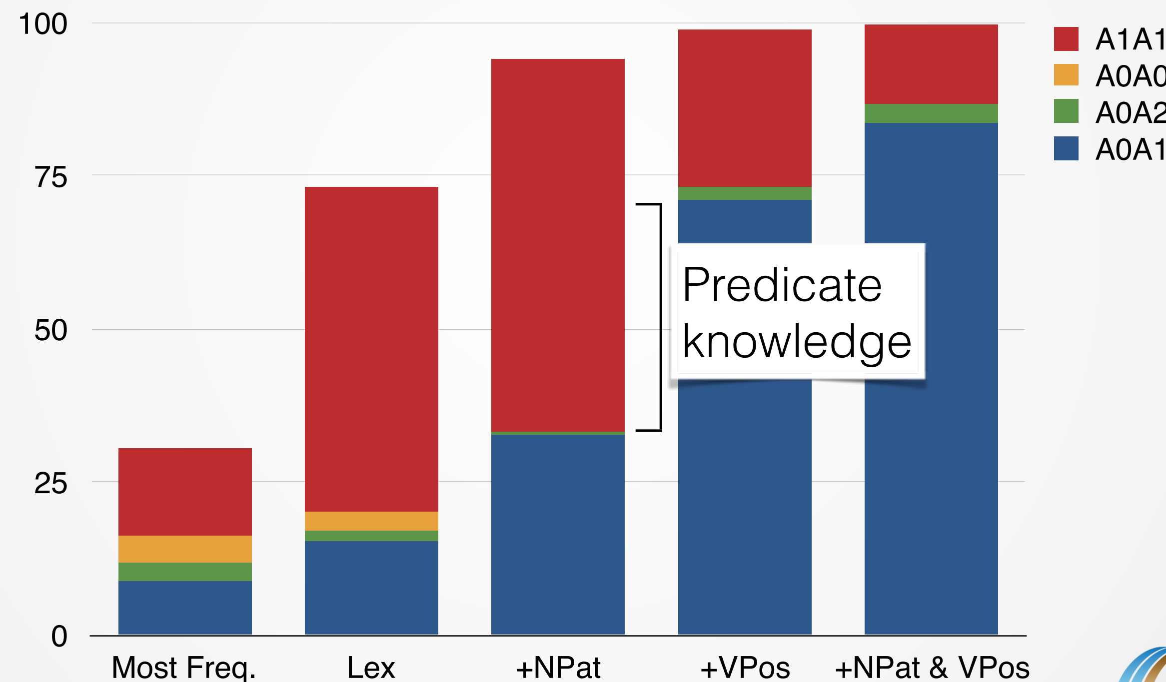
Here are the resulting patterns from each classifier. Here we're not necessarily testing the accuracy of each classifier; instead we want to know what it would predict given a sequence of nouns.

Experiment 1: Results



Here are the resulting patterns from each classifier. Here we’re not necessarily testing the accuracy of each classifier; instead we want to know what it would predict given a sequence of nouns.

Experiment 1: Results



Here are the resulting patterns from each classifier. Here we're not necessarily testing the accuracy of each classifier; instead we want to know what it would predict given a sequence of nouns.

Multiple predicates

"**Remember** how we play the surprise game?"

A1

"Remember how we **play** the surprise game?"

A0 A1

Here's another reason we need to know what the predicates are. As I already mentioned, the BabySRL is far from being a trivial dataset. And like a real child, our system is exposed to utterances with multiple predicates. The way we encode this information is to create a different proposition for each of the predicates (so for this sentence we have one proposition based on "remember", where "we" and "game" are both A1s, and a second proposition centred around "play" where "we" is now an A0, and "play" is an A1). And just to give you an idea of the significance of this, around a quarter of the corpus contains 2 verbs. So how does this affect the results of the system?

Multiple predicates

"**Remember** how we play the surprise game?"

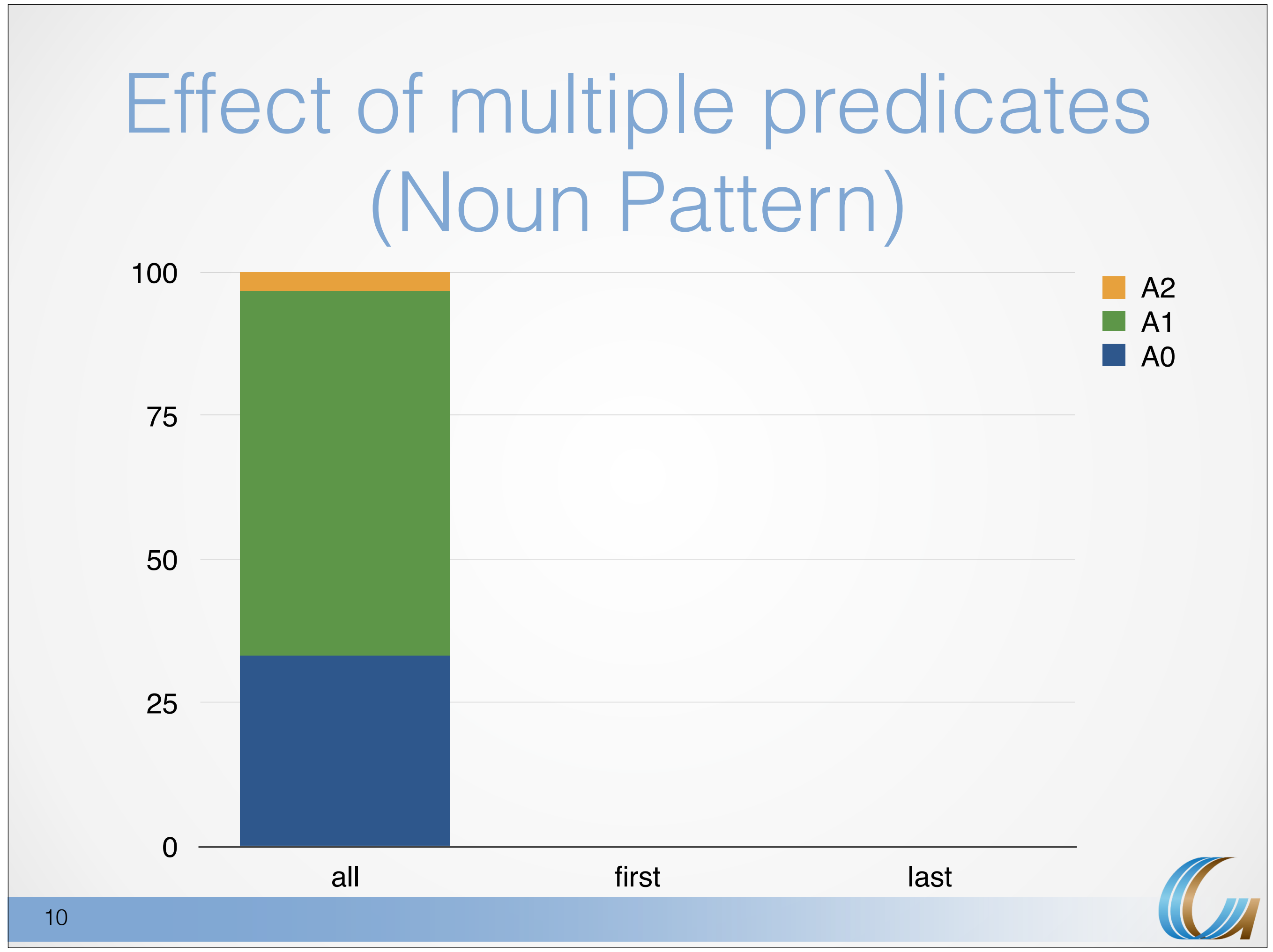
A1

"Remember how we **play** the surprise game?"

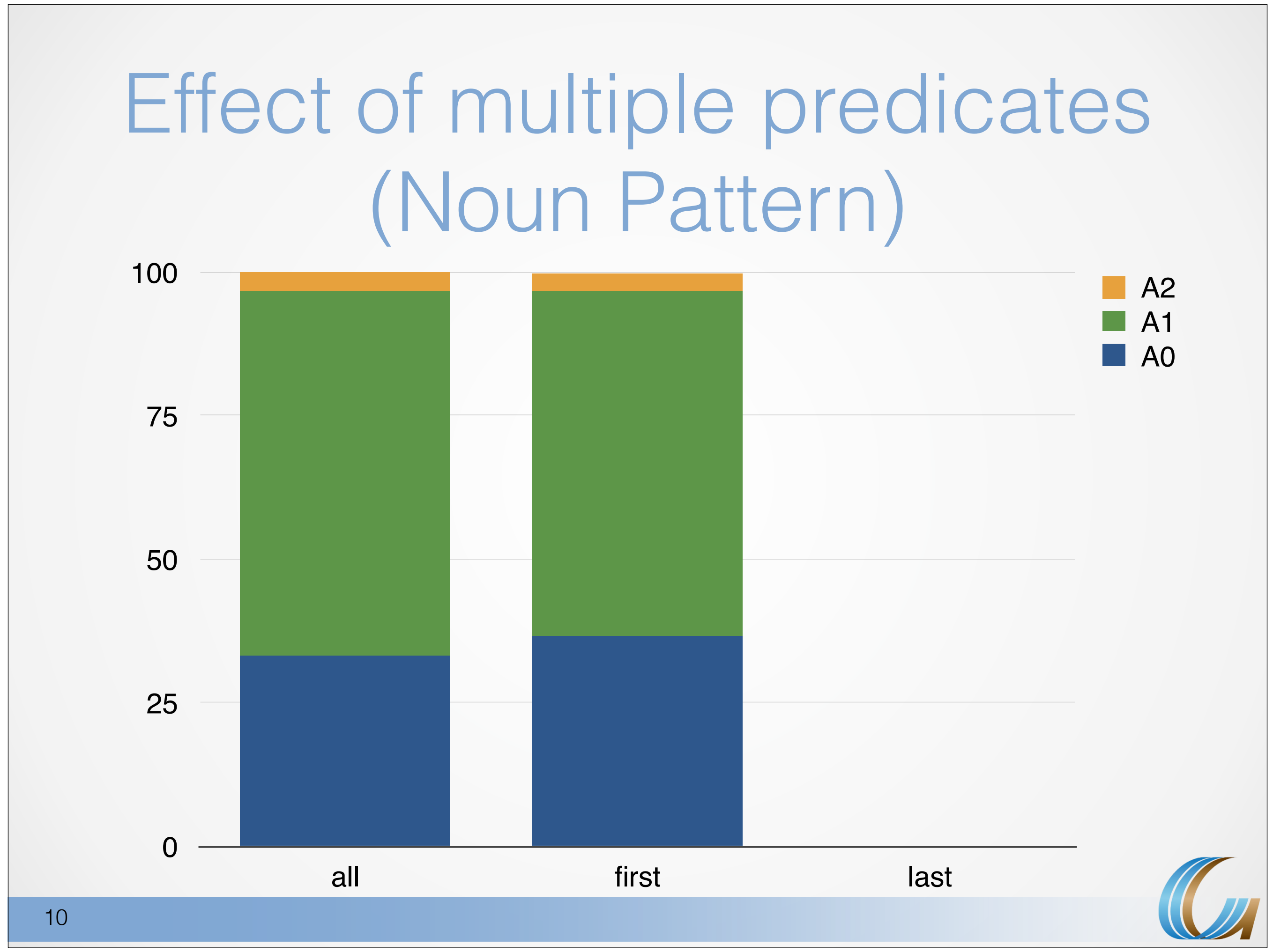
A0 A1

	# sent	%
1 verb	10,356	69.86
2 verbs	3,614	24.38

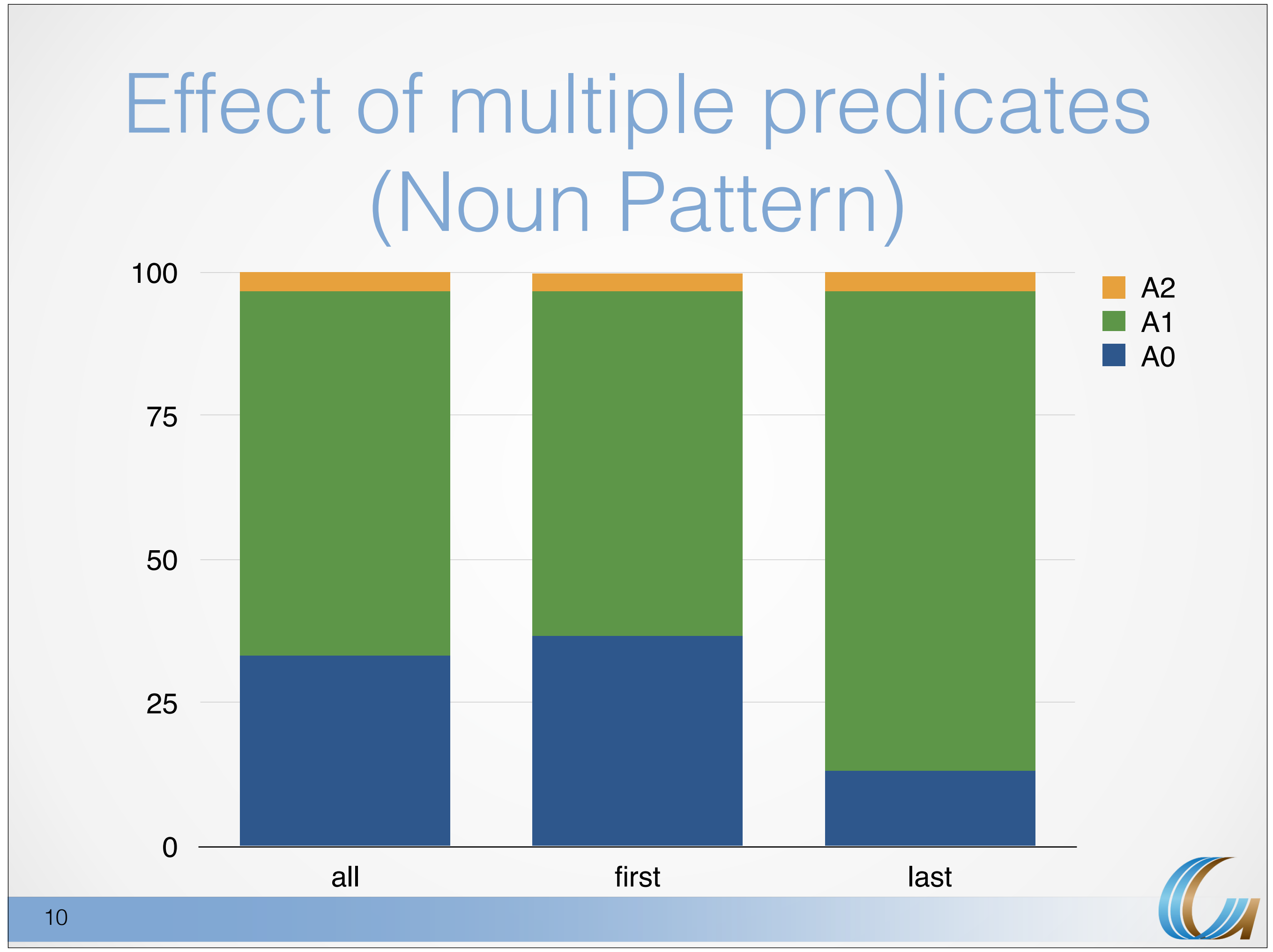
Here's another reason we need to know what the predicates are. As I already mentioned, the BabySRL is far from being a trivial dataset. And like a real child, our system is exposed to utterances with multiple predicates. The way we encode this information is to create a different proposition for each of the predicates (so for this sentence we have one proposition based on "remember", where "we" and "game" are both A1s, and a second proposition centred around "play" where "we" is now an A0, and "play" is an A1). And just to give you an idea of the significance of this, around a quarter of the corpus contains 2 verbs. So how does this affect the results of the system?



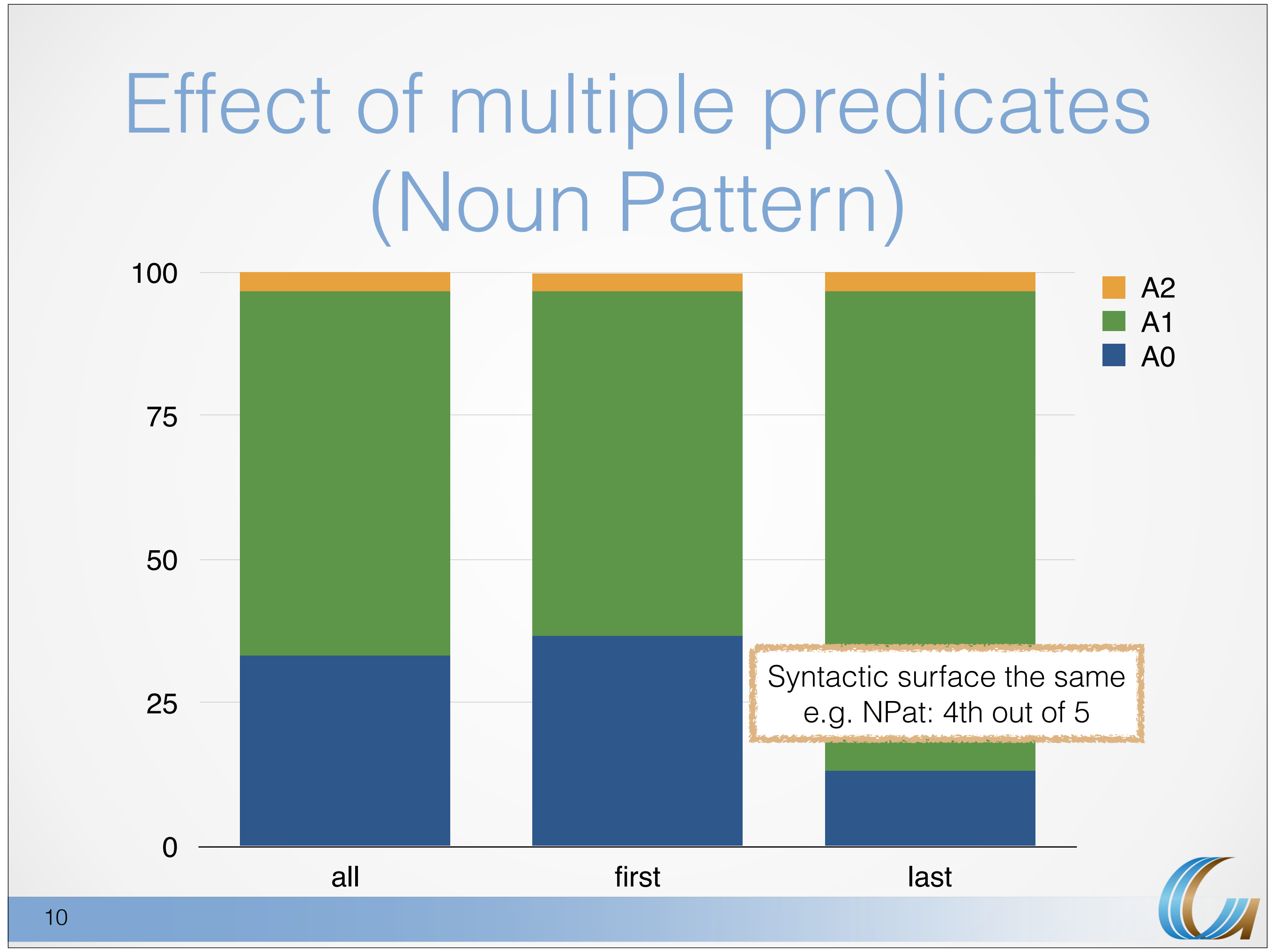
If we focus on the NounPattern feature, this is the previous score, the one where all versions of the same sentence are used as input; and here <SLIDE> is the same classifier if we only keep the fist proposition of every utterance and here <SLIDE> is if we keep the last. If this seems surprising, remember that the surface structure remains the same for the last setting. So we might get a feature like “4th out of 5 nouns” that still indicates an A0 (and little support for the 1st out of N nouns).



If we focus on the NounPattern feature, this is the previous score, the one where all versions of the same sentence are used as input; and here <SLIDE> is the same classifier if we only keep the fist proposition of every utterance and here <SLIDE> is if we keep the last. If this seems surprising, remember that the surface structure remains the same for the last setting. So we might get a feature like “4th out of 5 nouns” that still indicates an A0 (and little support for the 1st out of N nouns).



If we focus on the NounPattern feature, this is the previous score, the one where all versions of the same sentence are used as input; and here <SLIDE> is the same classifier if we only keep the fist proposition of every utterance and here <SLIDE> is if we keep the last. If this seems surprising, remember that the surface structure remains the same for the last setting. So we might get a feature like “4th out of 5 nouns” that still indicates an A0 (and little support for the 1st out of N nouns).



If we focus on the NounPattern feature, this is the previous score, the one where all versions of the same sentence are used as input; and here <SLIDE> is the same classifier if we only keep the fist proposition of every utterance and here <SLIDE> is if we keep the last. If this seems surprising, remember that the surface structure remains the same for the last setting. So we might get a feature like “4th out of 5 nouns” that still indicates an A0 (and little support for the 1st out of N nouns).

Experiment 1: Supervised learning

Given **perfect feedback**, do simple, **bottom-level** features capture anything useful about semantic roles/verb preferences?

What have we learned? <SLIDE> Yes, given, unambiguous feedback, simple bottom-level features are indeed useful, but knowledge of where the predicates are is crucial. So how do we go about discovering those predicates?

Experiment 1: Supervised learning

Given **perfect feedback**, do simple, **bottom-level** features capture anything useful about semantic roles/verb preferences?

Yes, but predicate knowledge is crucial

What have we learned? <SLIDE> Yes, given, unambiguous feedback, simple bottom-level features are indeed useful, but knowledge of where the predicates are is crucial. So how do we go about discovering those predicates?

Experiment 2: Unsupervised learning

Can we predict arguments/predicates using distributional **clusters** and a few **seed nouns**?

This is what our second experiment focuses on. Without using any form of supervision, other than a few seed nouns, can we predict which words are predicates/ arguments? <SLIDE> This experiment can be viewed as test of one of the proposed mechanisms for Syntactic Bootstrapping (this is the hypothesis that children can infer at least part of the semantic interpretation of an utterance by using the syntactic patterns they observe), called Structure-Mapping.

Structure-Mapping proposes that children are able to do that by mapping nouns to semantic arguments and using the number of arguments as a means to determine the frame of the predicate.

Experiment 2: Unsupervised learning

Can we predict arguments/predicates using distributional **clusters** and a few **seed nouns**?

Syntactic Bootstrapping via Structure-Mapping
[Gleitman, 1990; Fisher et al. 2010]

This is what our second experiment focuses on. Without using any form of supervision, other than a few seed nouns, can we predict which words are predicates/arguments? <SLIDE> This experiment can be viewed as test of one of the proposed mechanisms for Syntactic Bootstrapping (this is the hypothesis that children can infer at least part of the semantic interpretation of an utterance by using the syntactic patterns they observe), called Structure-Mapping.
Structure-Mapping proposes that children are able to do that by mapping nouns to semantic arguments and using the number of arguments as a means to determine the frame of the predicate.

Experiment 2: Unsupervised learning

- HMM over 2.2M tokens (CHILDES)
  - 80 induced clusters, list of function words
- List of seed nouns **[Dale and Fenson, 1996]**
- Noun identification

  "Cluster contains more than ***k*** seed nouns"

The setup for this experiment is the following: we use an off-the-shelf HMM to induce 80 clusters, training on a large part of the CHILDES corpus. The HMM is given a list of some function words and clusters them separately. We also assume knowledge of some seed nouns; we generated a list of 76 known nouns from a questionnaire of language development reporting (from Dale and Ferguson). Our first and easiest task is to identify all noun clusters, using the simple heuristic that a cluster is labeled as a noun if it contains more than a specific threshold k number of seed nouns (for all the experiments in this talk we use k=4).

Experiment 2: Verb Identification

	She	krads	a	red	truck
HMM	45	51	19	60	73
N Ident.	N				N
Funct.			F		

Our second task is the crucial one: identifying the verbs. We start by eliminating words that belong to either a Noun or a Function word state.

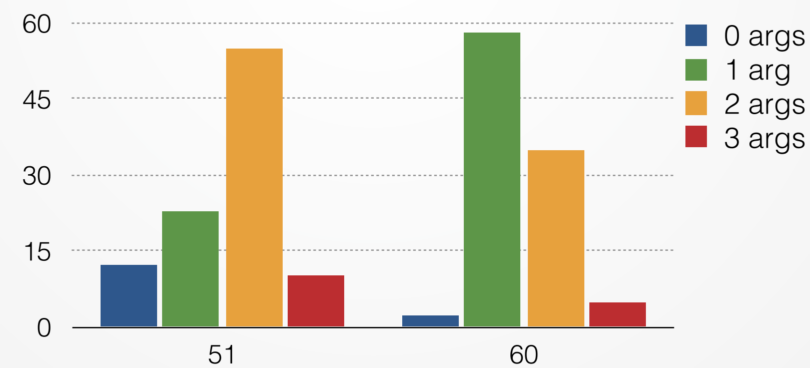
Experiment 2: Verb Identification

	She	krads	a	red	truck
HMM	45	**51**	19	**60**	73
N Ident.	N				N
Funct.			F		

We are left with two candidates (with states 51 and 60). <SLIDE> Now we look at a histogram of argument-taking frequency for each candidate. This basically tells us that words in cluster 51 really like to co-occur with 2 arguments whereas cluster 60 prefers 1. <SLIDE> We therefore chose cluster 51 as the verb cluster. (this heuristic only returns the maximum probability candidate, based on the assumption that most sentences have a single verb predicate)

Experiment 2: Verb Identification

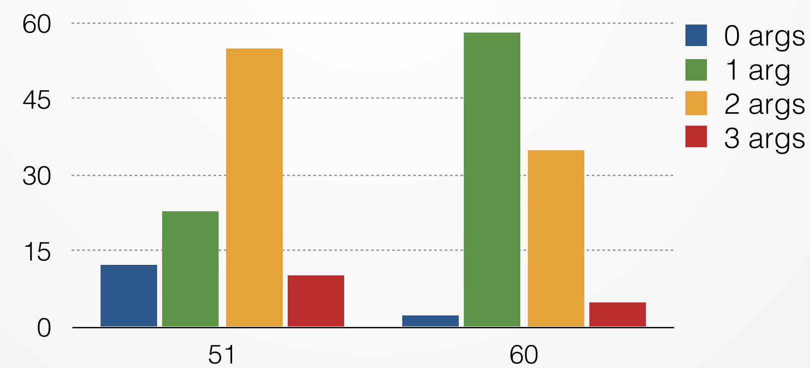
	She	krads	a	red	truck
HMM	45	**51**	19	**60**	73
N Ident.	N				N
Funct.			F		



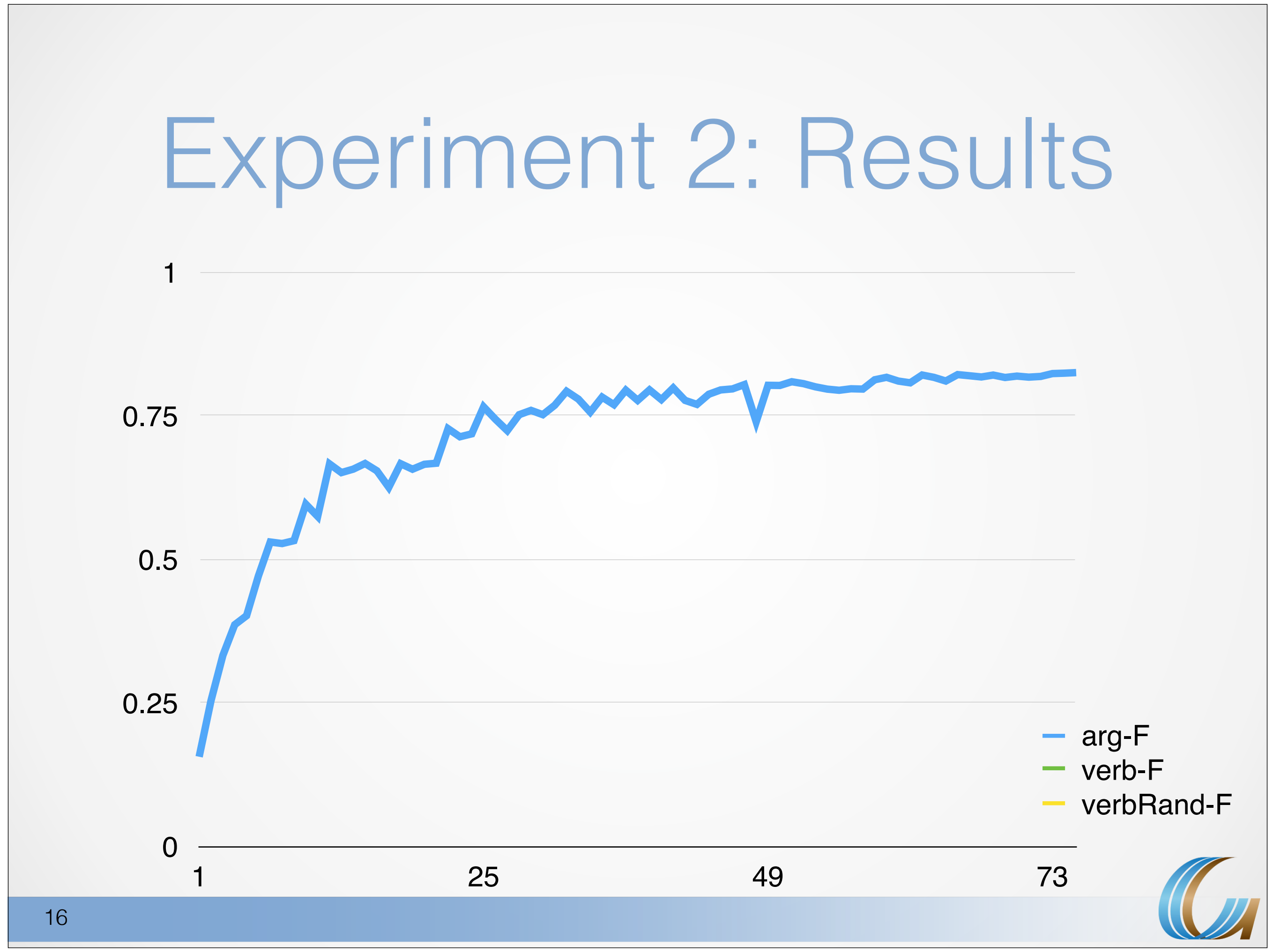
We are left with two candidates (with states 51 and 60). <SLIDE> Now we look at a histogram of argument-taking frequency for each candidate. This basically tells us that words in cluster 51 really like to co-occur with 2 arguments whereas cluster 60 prefers 1. <SLIDE> We therefore chose cluster 51 as the verb cluster. (this heuristic only returns the maximum probability candidate, based on the assumption that most sentences have a single verb predicate)

Experiment 2: Verb Identification

	She	krads	a	red	truck
HMM	45	$\boxed{\textbf{51}}$	19	**60**	73
N Ident.	N				N
Funct.			F		

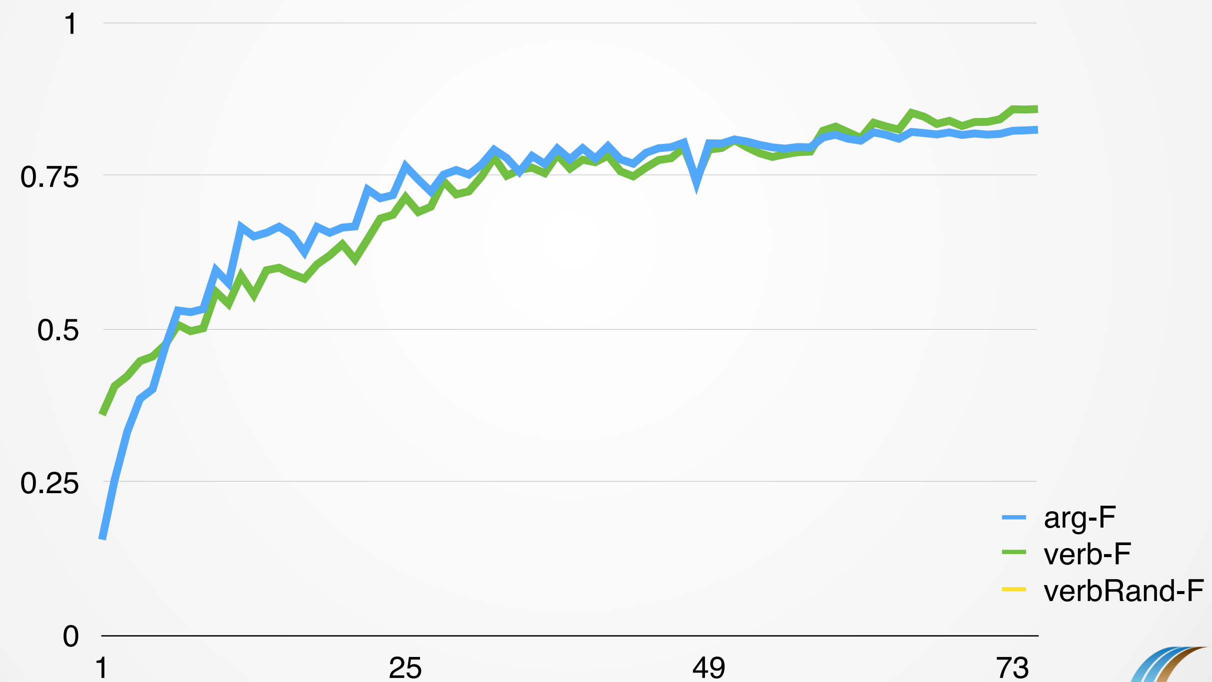


We are left with two candidates (with states 51 and 60). <SLIDE> Now we look at a histogram of argument-taking frequency for each candidate. This basically tells us that words in cluster 51 really like to co-occur with 2 arguments whereas cluster 60 prefers 1. <SLIDE> We therefore chose cluster 51 as the verb cluster. (this heuristic only returns the maximum probability candidate, based on the assumption that most sentences have a single verb predicate)

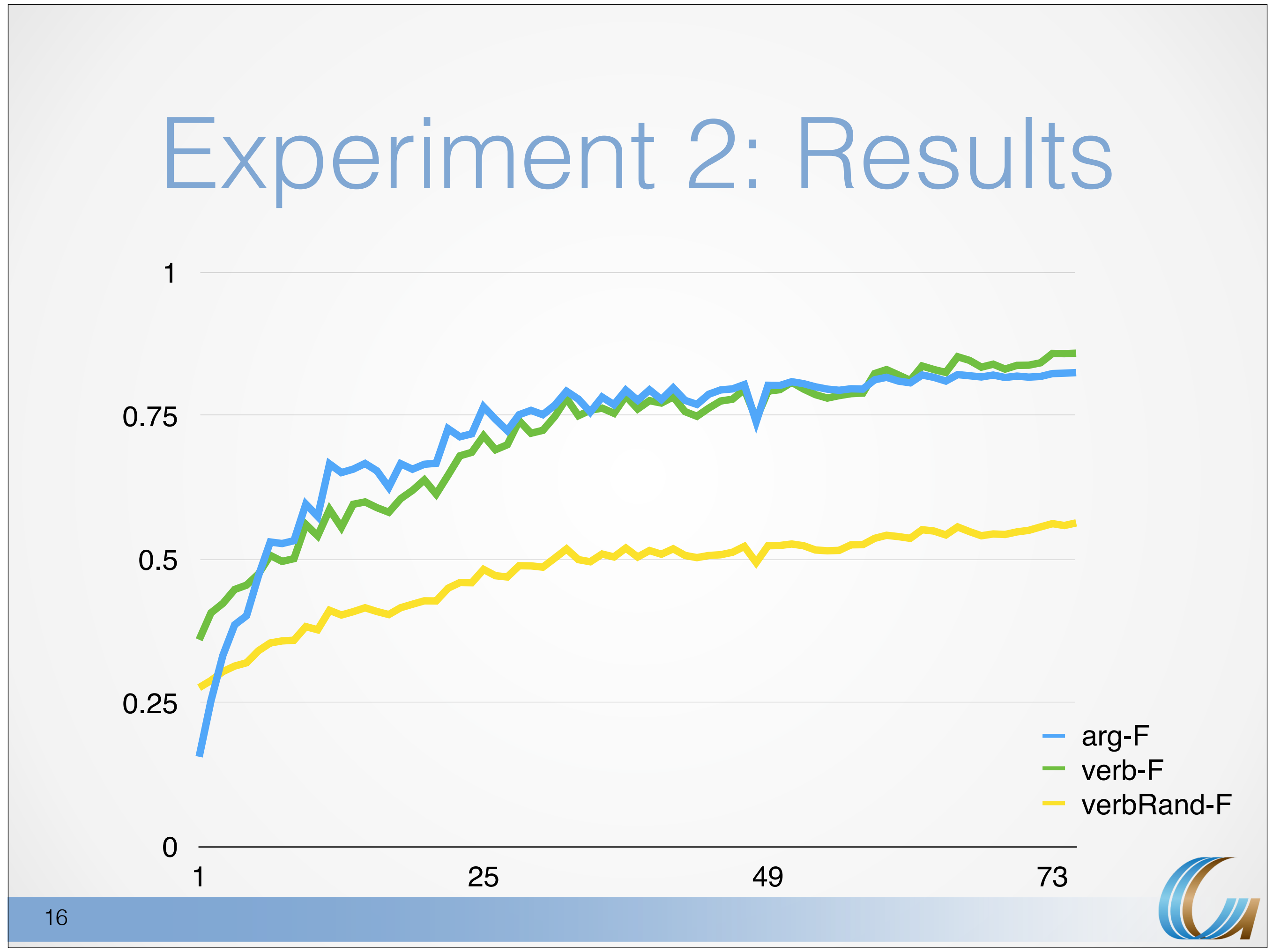


We can see here, as expected, that as the number of seed nouns we use goes up, so does the performance of our argument heuristic. <SLIDE> The same is true for the verb heuristic, which significantly outperforms the random baseline <SLIDE> (this baseline uses the same exclusion methods as the predicate heuristic —known noun states and function words are excluded— and chooses a random cluster from the remaining candidates.

Experiment 2: Results



We can see here, as expected, that as the number of seed nouns we use goes up, so does the performance of our argument heuristic. <SLIDE> The same is true for the verb heuristic, which significantly outperforms the random baseline <SLIDE> (this baseline uses the same exclusion methods as the predicate heuristic —known noun states and function words are excluded— and chooses a random cluster from the remaining candidates.



We can see here, as expected, that as the number of seed nouns we use goes up, so does the performance of our argument heuristic. <SLIDE> The same is true for the verb heuristic, which significantly outperforms the random baseline <SLIDE> (this baseline uses the same exclusion methods as the predicate heuristic —known noun states and function words are excluded— and chooses a random cluster from the remaining candidates.

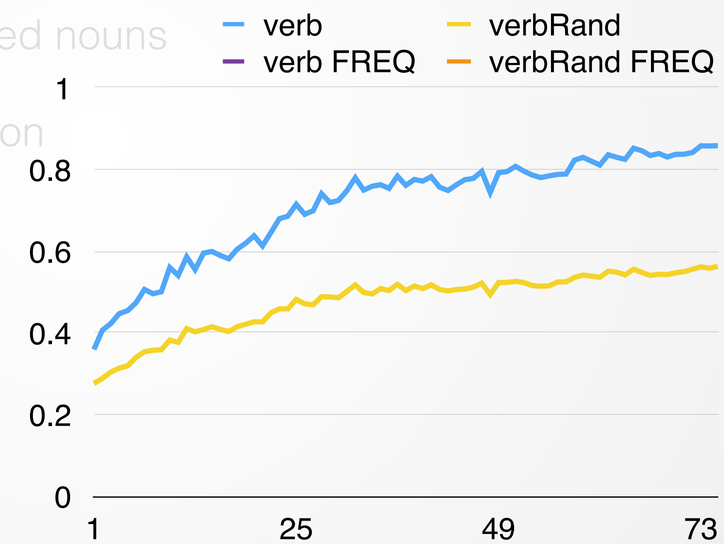
Experiment 2: Parameters

- Random/frequent seed noun selection
- Variants + plurals of seed nouns
- Verb/predicate evaluation
- Multiple predicates
- Seed noun threshold ***k***
- Null predictions
- Function words

However, behind this experiment lie a number of parameters, for which we need to either present a valid psycholinguistic account of why we are choosing a specific value, or explore their effect via experimentation. Some of these parameters are listed here. I will not go into details about every one, but we will be presenting a detailed account in our upcoming publication. I would like to focus on a couple of crucial parameters.

Experiment 2: Parameters

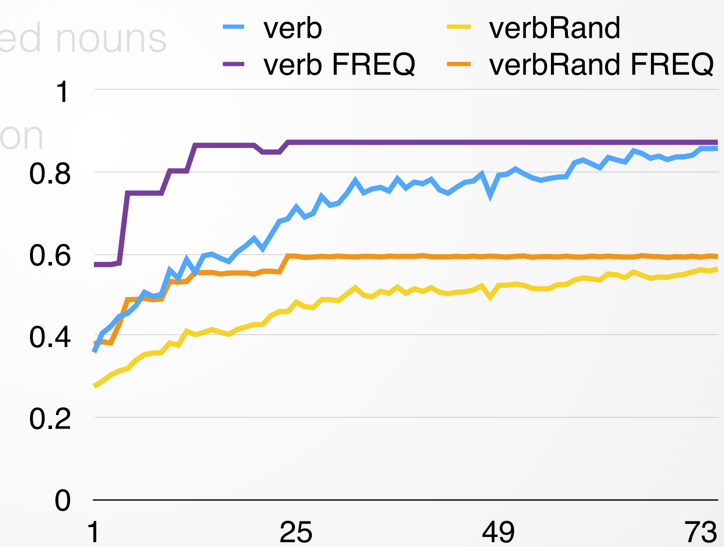
- Random/frequent seed noun selection
- Variants + plurals of seed nouns
- Verb/predicate evaluation
- Multiple predicates
- Seed noun threshold ***k***
- Null predictions
- Function words



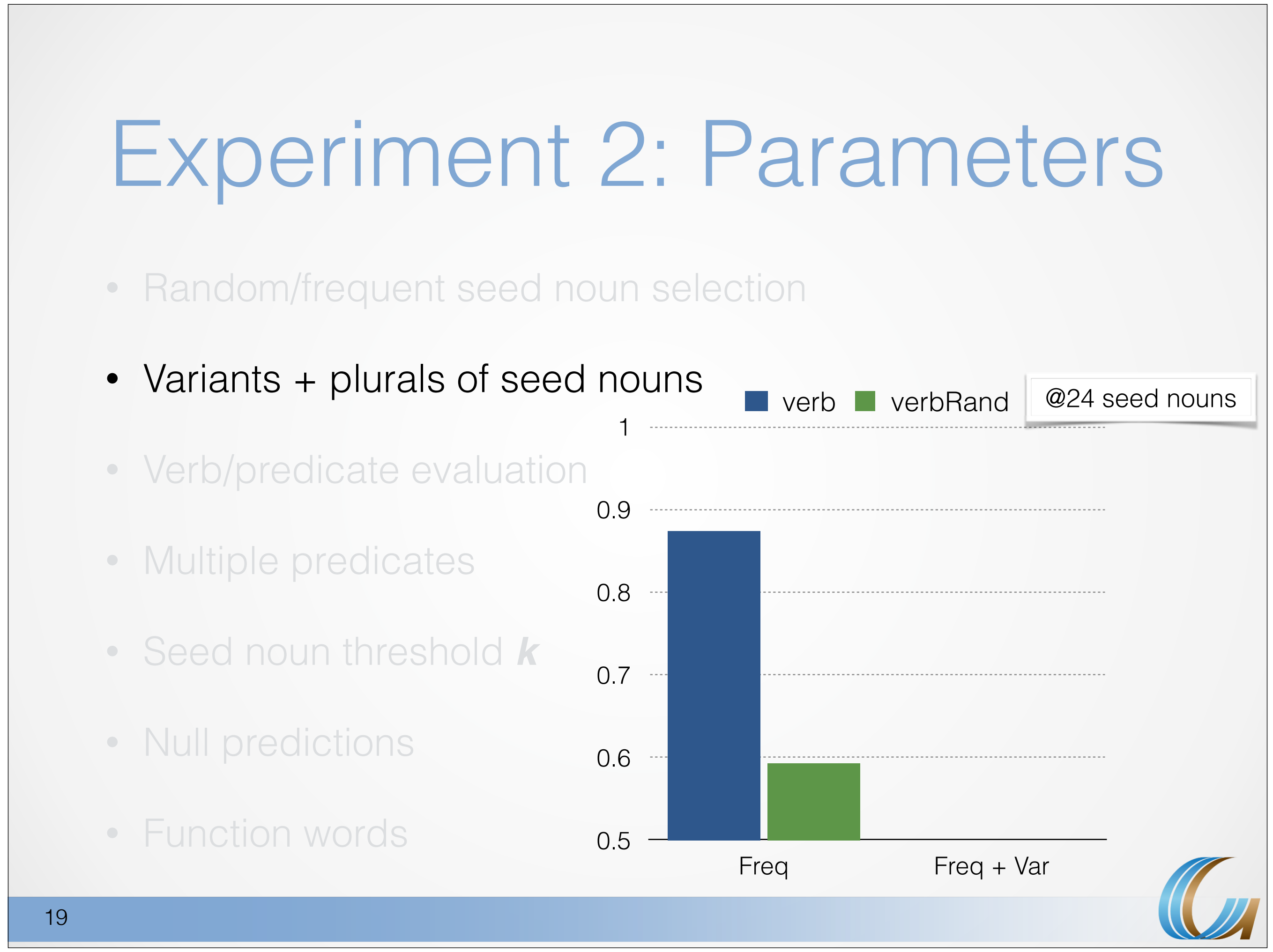
First, we need to look at our seed noun selection. The first set of results used a random subset for each number of seeds (averaged over 10 runs). But there is another source of information that we are not taking into account, and that is word frequency. If we sort the seed nouns by frequency <SLIDE> we see that even with as few as 24 nouns we can get the same performance as with the whole set of 76 nouns.

Experiment 2: Parameters

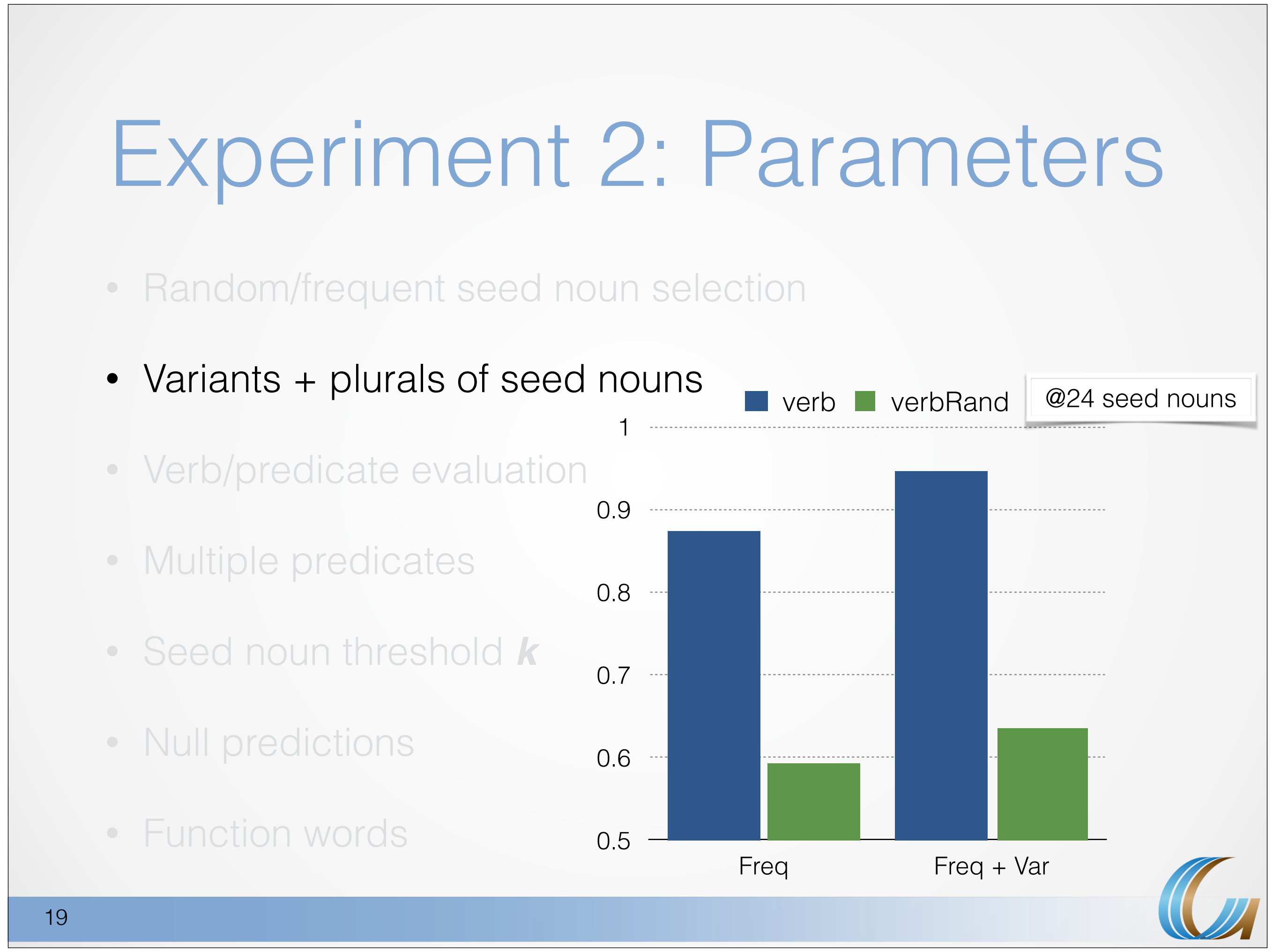
- Random/frequent seed noun selection
- Variants + plurals of seed nouns
- Verb/predicate evaluation
- Multiple predicates
- Seed noun threshold ***k***
- Null predictions
- Function words



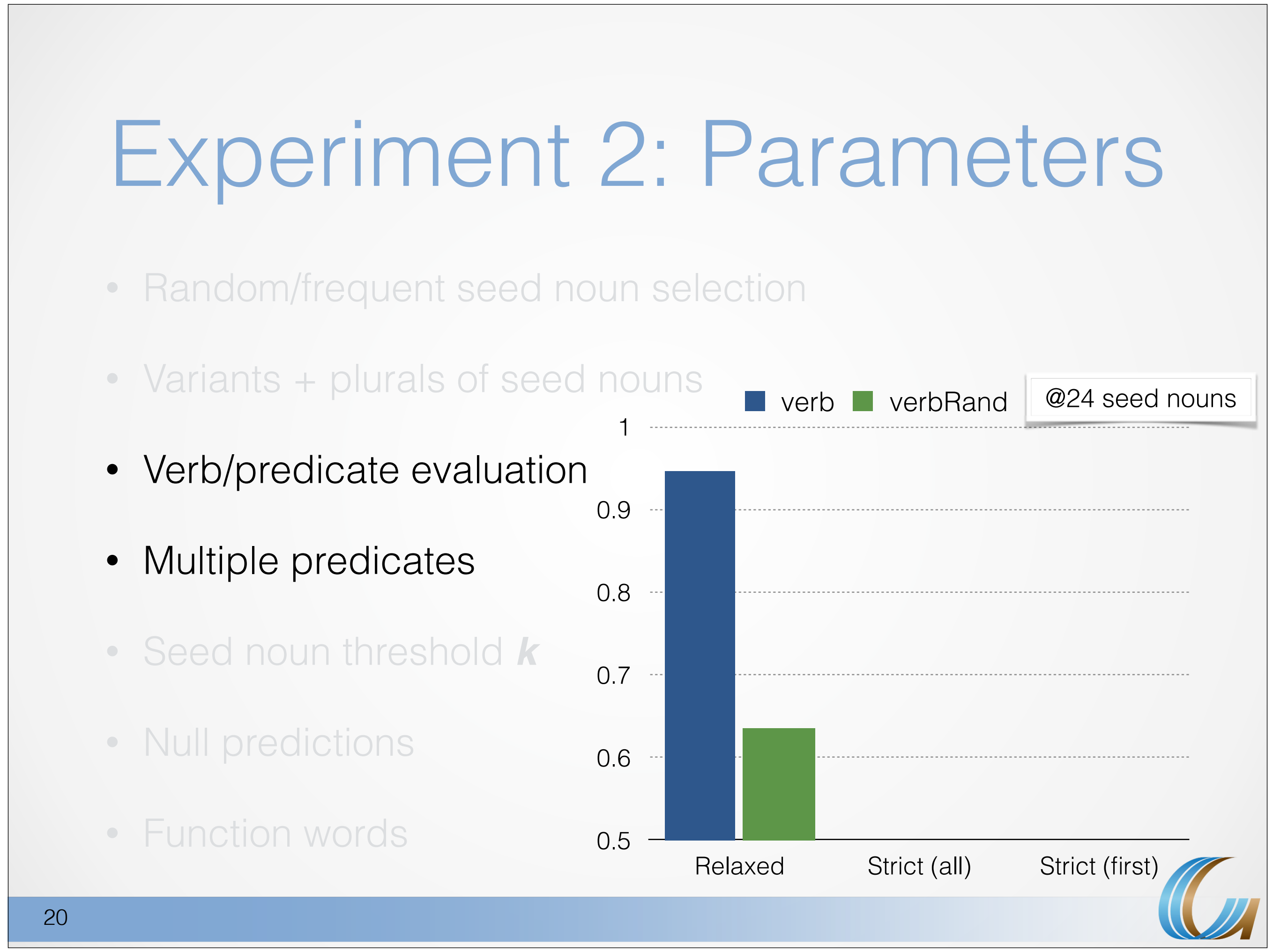
First, we need to look at our seed noun selection. The first set of results used a random subset for each number of seeds (averaged over 10 runs). But there is another source of information that we are not taking into account, and that is word frequency. If we sort the seed nouns by frequency <SLIDE> we see that even with as few as 24 nouns we can get the same performance as with the whole set of 76 nouns.



A slightly more technical parameter has to do with the representation of our input. The original seed noun list (used by Connor et al.) was based on the "normalised" questionnaire of Dale and Fenson; this meant that the list contained on the singular form of words like "toe" or "toy" whereas most of the occurrences of those words in the corpus were in the plural. Other cases of normalisation had to do with different renderings of words like "mommy" or "dollie". If we include all these variants we see that we get a significant boost in our verb identification performance —reaching close to 95%.



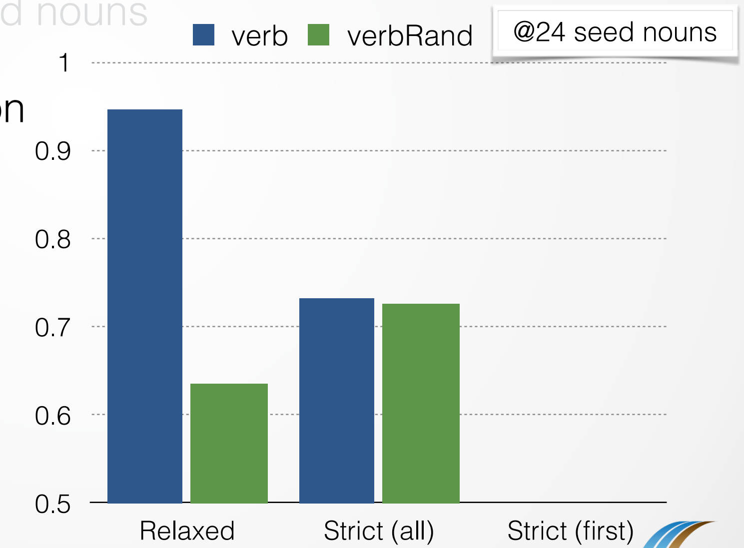
A slightly more technical parameter has to do with the representation of our input. The original seed noun list (used by Connor et al.) was based on the "normalised" questionnaire of Dale and Fenson; this meant that the list contained on the singular form of words like "toe" or "toy" whereas most of the occurrences of those words in the corpus were in the plural. Other cases of normalisation had to do with different renderings of words like "mommy" or "dollie". If we include all these variants we see that we get a significant boost in our verb identification performance —reaching close to 95%.



However, as you may remember from experiment 1, our corpus contains lots of utterances with multiple predicates. If instead of measuring the performance of our heuristic in terms of finding any verb in the utterance, we limit it to finding only the correct predicate each time <SLIDE> we see that the performance drops significantly, almost at the level of chance. However we see that the heuristic can learn if we limit our system to only using the first proposition of each utterance <SLIDE>.

Experiment 2: Parameters

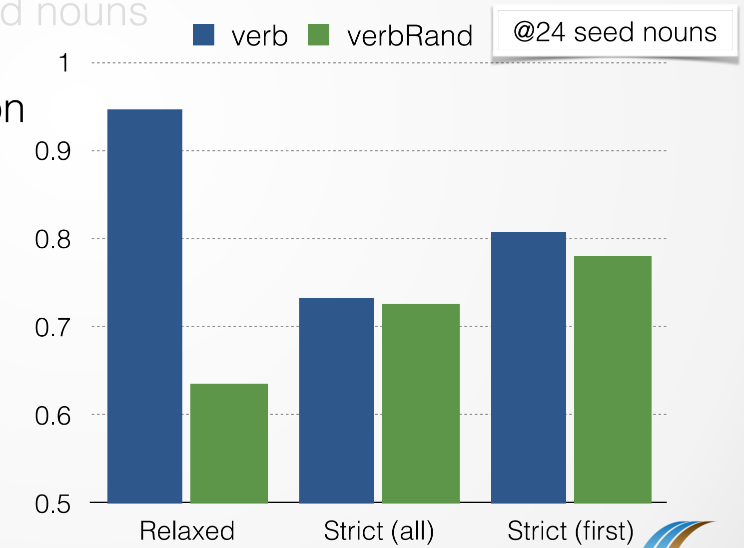
- Random/frequent seed noun selection
- Variants + plurals of seed nouns
- Verb/predicate evaluation
- Multiple predicates
- Seed noun threshold ***k***
- Null predictions
- Function words



However, as you may remember from experiment 1, our corpus contains lots of utterances with multiple predicates. If instead of measuring the performance of our heuristic in terms of finding any verb in the utterance, we limit it to finding only the correct predicate each time <SLIDE> we see that the performance drops significantly, almost at the level of chance. However we see that the heuristic can learn if we limit our system to only using the first proposition of each utterance <SLIDE>.

Experiment 2: Parameters

- Random/frequent seed noun selection
- Variants + plurals of seed nouns
- Verb/predicate evaluation
- Multiple predicates
- Seed noun threshold ***k***
- Null predictions
- Function words



However, as you may remember from experiment 1, our corpus contains lots of utterances with multiple predicates. If instead of measuring the performance of our heuristic in terms of finding any verb in the utterance, we limit it to finding only the correct predicate each time <SLIDE> we see that the performance drops significantly, almost at the level of chance. However we see that the heuristic can learn if we limit our system to only using the first proposition of each utterance <SLIDE>.

Experiment 2: Unsupervised learning

Can we predict arguments/predicates using distributional **clusters** and a few **seed nouns**?

So in terms of our original question, <SLiDE> yes, we can learn to predict predicates and arguments with as few as 24 nouns, but <SLIDE> we saw that considering multiple predicates is crucial.

Experiment 2: Unsupervised learning

Can we predict arguments/predicates using distributional **clusters** and a few **seed nouns**?

Yes, with as few as 24 seed nouns

So in terms of our original question, <SLiDE> yes, we can learn to predict predicates and arguments with as few as 24 nouns, but <SLIDE> we saw that considering multiple predicates is crucial.

Experiment 2: Unsupervised learning

Can we predict arguments/predicates using distributional **clusters** and a few **seed nouns**?

Yes, with as few as 24 seed nouns

need to consider multiple predicates

So in terms of our original question, <SLiDE> yes, we can learn to predict predicates and arguments with as few as 24 nouns, but <SLIDE> we saw that considering multiple predicates is crucial.

Conclusions

- BabySRL model of language acquisition
  - Evidence for syntactic bootstrapping
- Exploration of assumptions
  - Data representation
  - Evaluation
  - Psycholinguistic validity

In conclusion, I have presented the BabySRL model of language acquisition that, for the fist time, provides compelling evidence for the theory of syntactic bootstrapping. We explored a lot of the assumptions present in the input data, the system itself (and its evaluation) and I believe that this exploration will offer a psychologic validity to the parameter choices of this model.

Future Directions

- BabySRL from scratch [Connor et al. 2012]
- Beyond single predicates
  - Multiple verbs
  - Prepositions
- Relaxing perfect feedback (scene ambiguity)
  - Superset
  - Bootstrapped Animacy

Our first obvious next step is to combine the two experiments presented here, in an extension what Connor et al. called "BabySRL from scratch", where we go beyond single predicates (both verbal and prepositional) and where we relax the assumption of veridical top-down feedback using two different mechanisms (a superset of the gold labels or a simple animacy-based feedback, where the learned assigns the agent category to the animate object in the scene).

Future Directions

- BabySRL from scratch **[Connor et al. 2012]**
- Beyond single predicates
  - Multiple verbs
  - Prepositions
- Relaxing perfect feedback (scene ambiguity)
  - Superset
  - Bootstrapped Animacy

Thanks

Our first obvious next step is to combine the two experiments presented here, in an extension what Connor et al. called “BabySRL from scratch”, where we go beyond single predicates (both verbal and prepositional) and where we relax the assumption of veridical top-down feedback using two different mechanisms (a superset of the gold labels or a simple animacy-based feedback, where the learned assigns the agent category to the animate object in the scene).