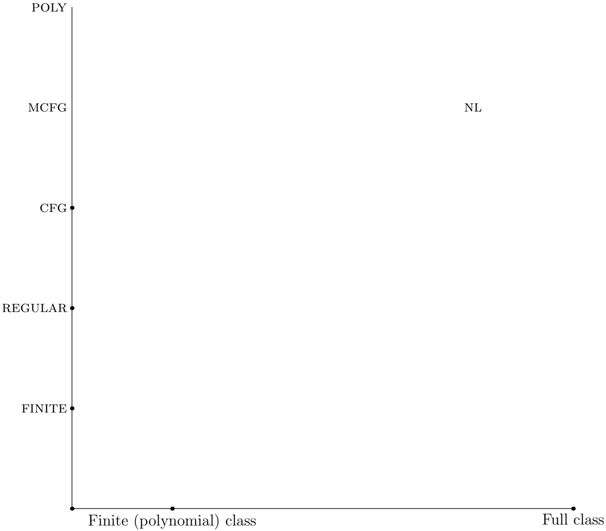
Weak and Strong Learning of Syntax

Alexander Clark

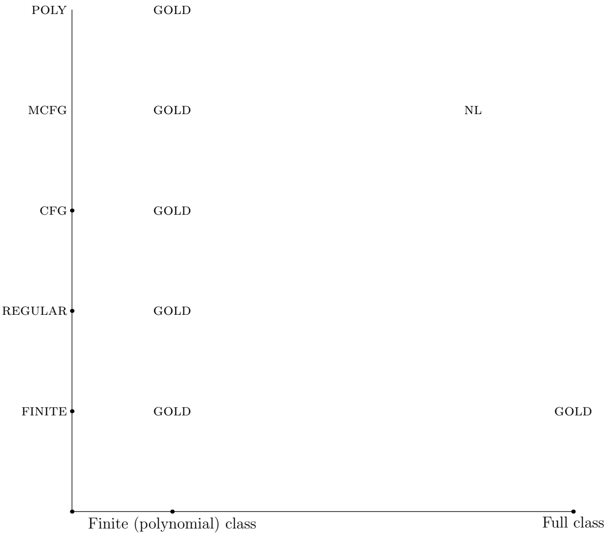
Department of Computer Science
Royal Holloway, University of London
alexc@cs.rhul.ac.uk

ISA 2012
Portland, Oregon

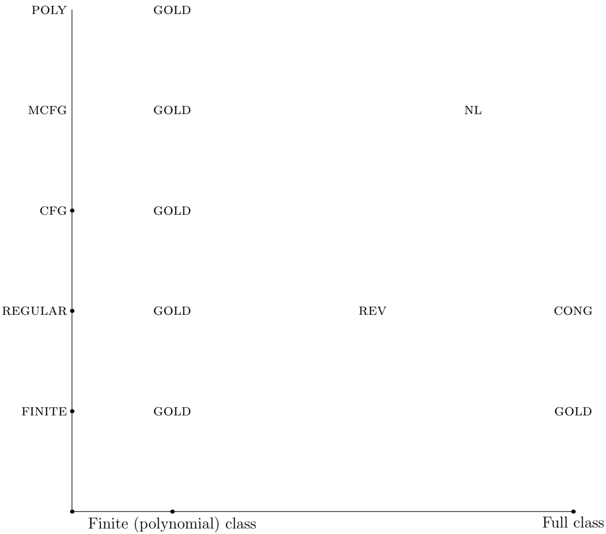
The class of natural languages



1967

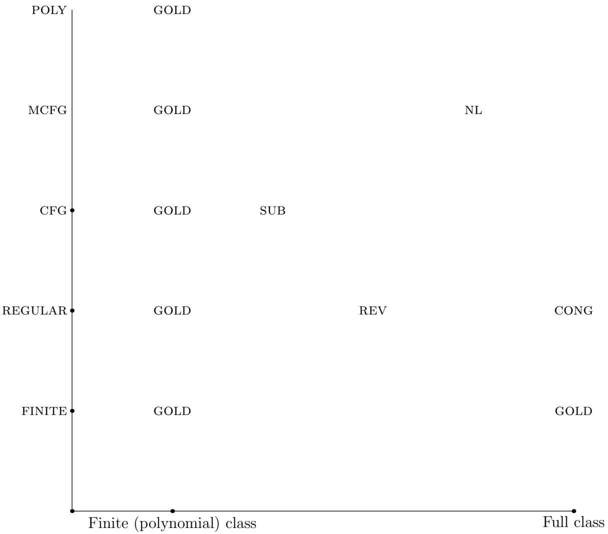

POLY
GOLD
MCFG
GOLD
NL
CFG
GOLD
REGULAR
GOLD
FINITE
GOLD
GOLD
Finite (polynomial) class
Full class

1987

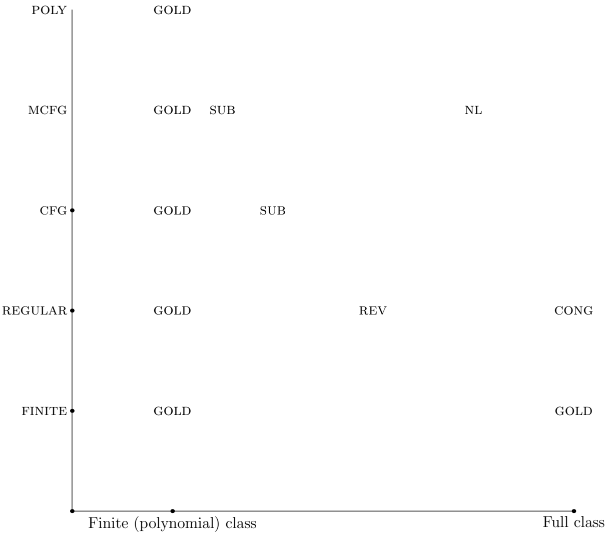


Royal Holloway University of London

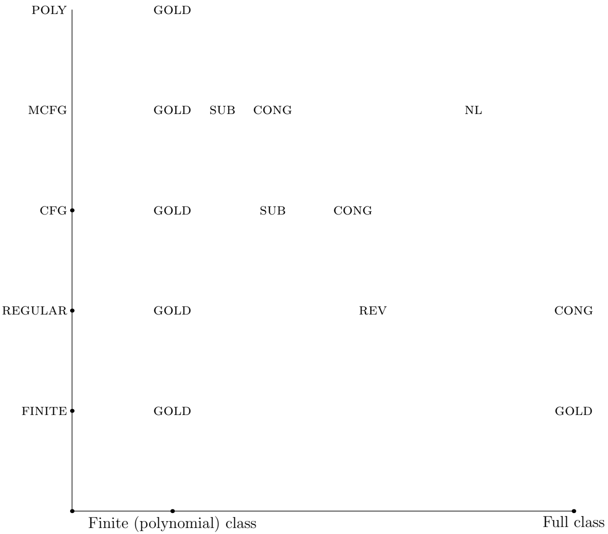
2005


POLY
MCFG
CFG
REGULAR
FINITE
GOLD
GOLD
GOLD
GOLD
GOLD
NL
SUB
REV
CONG
GOLD
Finite (polynomial) class
Full class

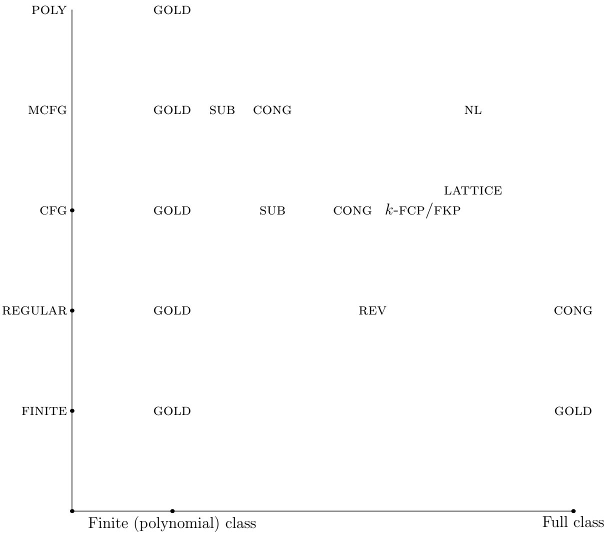
2008


POLY
GOLD
MCFG
GOLD
SUB
NL
CFG
GOLD
SUB
REGULAR
GOLD
REV
CONG
FINITE
GOLD
GOLD
Finite (polynomial) class
Full class

2010


POLY
GOLD
MCFG
GOLD SUB CONG
NL
CFG
GOLD SUB CONG
REGULAR
GOLD REV CONG
FINITE
GOLD GOLD
Finite (polynomial) class
Full class

2011


POLY
GOLD
MCFG
GOLD
SUB
CONG
NL
LATTICE
CFG
GOLD
SUB
CONG
k-FCP/FKP
REGULAR
GOLD
REV
CONG
FINITE
GOLD
GOLD
Finite (polynomial) class
Full class

Weaknesses of these approaches

- Classes of languages are (mostly) too small
- Learning models are (mostly) too easy and/or too idealised
- They (mostly) lack an appropriate “feature calculus”
- These are (all) just weak learnability results.

Weak and strong learning

Berwick et al. 2011

Put another way, language acquisition is not merely a matter of acquiring a capacity to associate word strings with interpretations. Much less is it a mere process of acquiring a (weak generative) capacity to produce just the valid word strings of a language. Idealizing, one can say that each child acquires a procedure that generates boundlessly many meaningful expressions, and that a single string of words can correspond to more than one expression.
. . . Explanations should . . . Yield the correct structures, for purposes of interpretation.

- **Weak** learning: learn the set of strings
- **Strong** learning: learn the right set of bracketed structures

Topics

Questions

Are these weak results irrelevant?

Topics

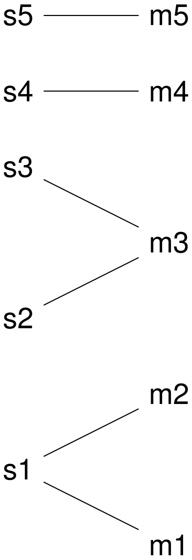
Questions

Are these weak results irrelevant?

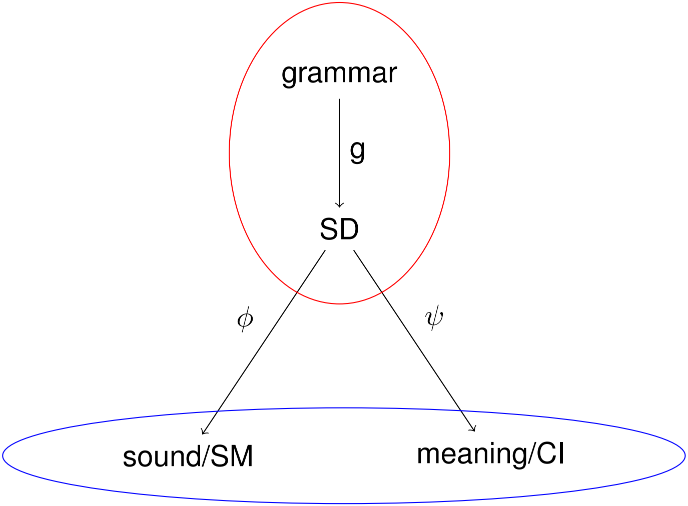
Propositions

- "Strong" learning (convergence of structural descriptions) is irrelevant
- Weak learning is not so far from strong learning once we take account of semantic well formedness
- We can extract appropriate structural descriptions from some of these weak learning algorithms.

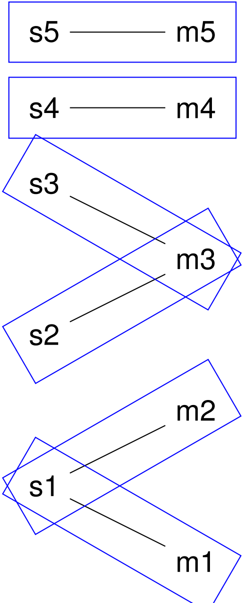
Sound/meaning pairs



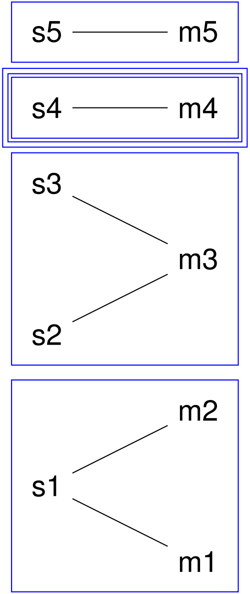
Structural descriptions



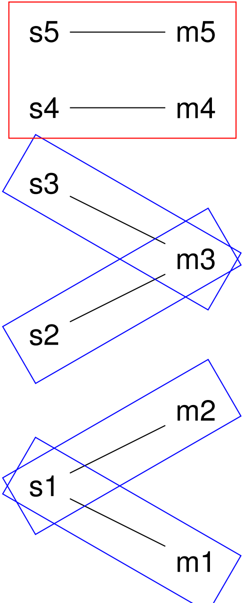
Standard model



Alternative model



Invalid structures



Weak learning

Two reasonable conceptions

Weak learning of strings

Tractable but ignores the role of semantics

Weak learning of sound/meaning pairs

Tractable (Yoshinaka and Kanazawa, 2011)
Assumes that learners have complete access to meanings

- Implausible even for adults
- Language acquisition starts before children plausibly have the cognitive resources to do the relevant inferences.

We observe convergence in these senses

Strong learning I

Strong learning

Inputs are strings, output generates “correct structures”
Intractable but irrelevant

We do not observe strong learning

- Different speakers who have converged to identical sets of sound/meaning pairs might assign slightly different structures.
- For example, memorize different chunks

Psycholinguistic data

Experimental data

Other than from the set of sound/meaning pairs

- Click Perception (Fodor and Bever, 1965)
- Structural Priming (Bock, 1986)
- Neuroimaging (Tettamanti et al. 2002)

Evidence for some types of hierarchically structured representations but don't provide evidence on inter-speaker convergence.

Strong learning II

But we do need to have some structural descriptions:

- Hierarchically structured representations that support semantic interpretation
  - Ambiguity
  - Displaced constituents – “movement”
- Learnable with limited or no information about the semantics
- Unrealistic to expect to be able to distinguish spurious ambiguity from semantic ambiguity
- Don’t require exact convergence

Inputs for weak learning

Assumption

We have some well defined set of strings of phonemes.

$L \subseteq \Sigma^*$

Assume:

- L is the set of syntactically well formed utterances
- Semantic constraints are handled by another component

Inputs revisited

- My aunt is pregnant
- My toothbrush is covered with toothpaste

Inputs revisited

- My aunt is pregnant
- My toothbrush is covered with toothpaste

- # My aunt is covered with toothpaste
- # My toothbrush is pregnant

The input to the child consists largely of semantically well formed utterances.

English example

- Mary kicked the ball and Jill ate the cake

English example

- Mary kicked the ball and Jill ate the cake
- # Mary fed the ball and Jill ate the cake
- # Mary kicked the cake and Jill ate the cake
- # Mary fed the ball and Jill ate the cat
- Mary fed the cat and Jill ate the cake

Semantic

Proposition

The right dependencies can be inferred if the child gets semantically well formed utterances

- Weak learning of the set of semantically well formed utterances gives you strong learning implicitly.
- But how can we make these dependencies explicit?

Structural descriptions from Distributional Learning

Congruential approach

Equivalence classes of strings that are distributionally identical

Lattice based approach

Sets of strings that have some shared distribution
A hierarchy of sets of strings:

- Large sets of strings that have a small set of contexts in common
- Smaller sets of strings that are very similar or identical

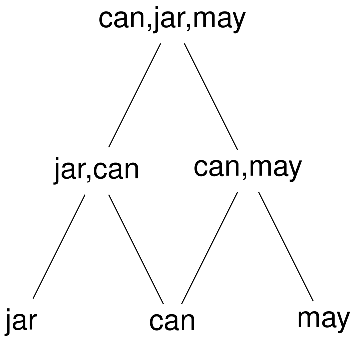
Ambiguity in congruential models

Examples

- Can I have a can of beans?
- May I have a jar of beans?

- “can” and “may” are different distributionally
- “can” and “jar” are different distributionally
- The structural descriptions are thus completely distinct

Ambiguity in lattice models



Lattice labeled trees

Structural descriptions that are ordered trees:

- Each node is labeled with a concept (set of strings) that contains the yield.
- The concept of each node must *properly* contain the concatenation of the concepts of the children of that node.
- The root node must be the concept that consist of the language.
- Now we have a choice of labels within a fixed tree.
- **We want labels to be maximal subject to the above constraints.**

Classic Example

- John is eager to please
- John is easy to please

Classic Example

- John is eager to please
- John is easy to please

- John is ready to eat
- the soup is ready to eat
- the chicken is ready to eat

Classic Example

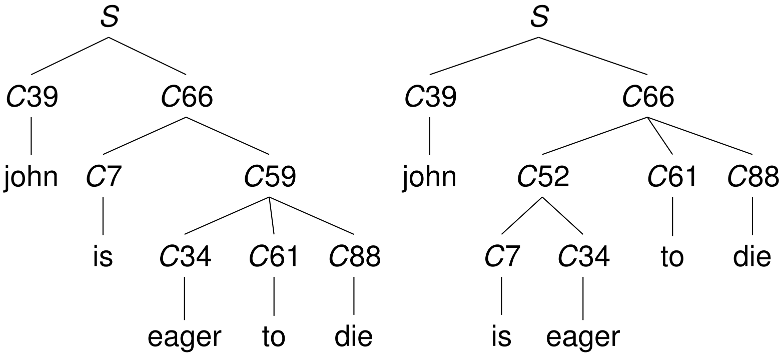
- John is eager to please
- John is easy to please

- John is ready to eat
- the soup is ready to eat
- the chicken is ready to eat

Methodology Toy grammar; learn CFG from examples; derived set of all valid trees.

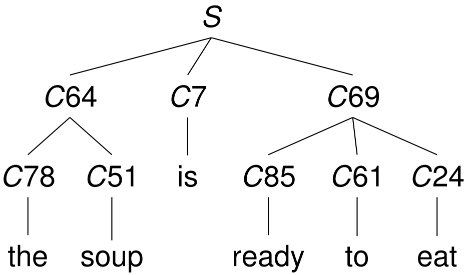
Trees

John is eager to die



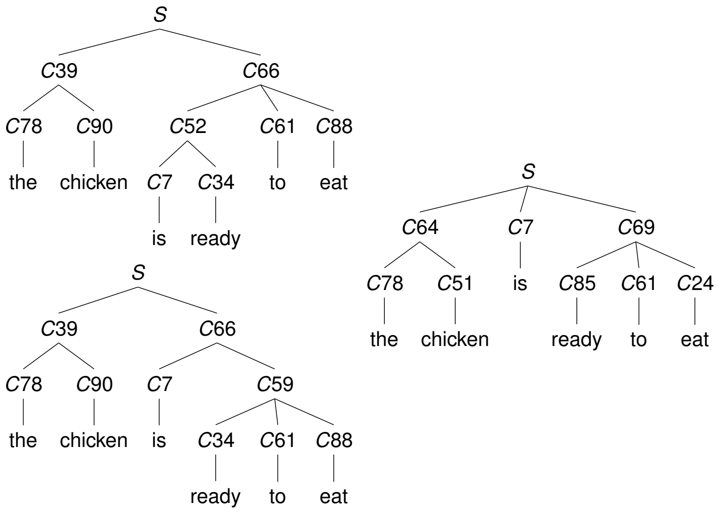
Trees

the soup is ready to eat



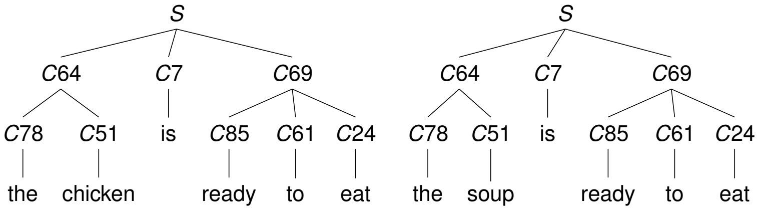
Trees

the chicken is ready to eat



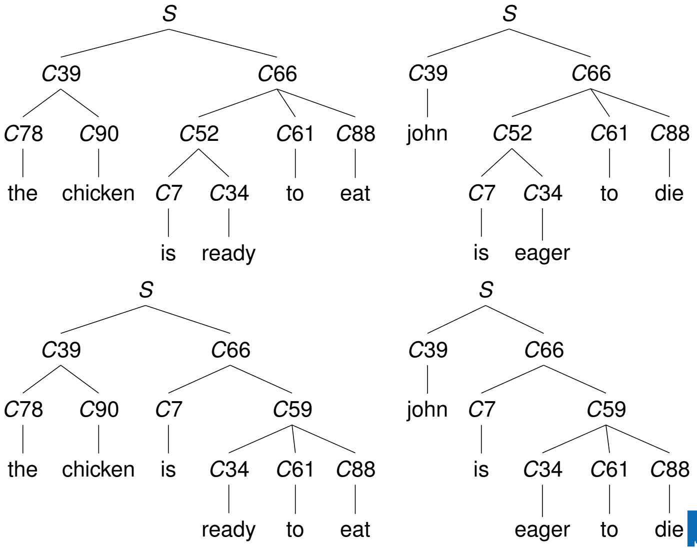
Trees

Comparison 1



Trees

Comparison 2



Multiple context free grammars
Yoshinaka, 2009,2010

"this is the book that I told you to read"

- "this is _ that I told you to _ "
- (the book, read)

Equivalence of MCFGs and Minimalist Grammars

MGs are weakly and strongly equivalent to MCFGs
Derivation trees of MGs can be learned.
Derived trees can be deterministically generated from the derivation trees
This gives a natural treatment of "movement"

A motivational quote

Chomsky, 1957

Precisely constructed models for linguistic structure can play an important role, both negative and positive, in the process of discovery itself. By pushing a precise but inadequate formulation to an unacceptable conclusion, we can often expose the exact source of this inadequacy and, consequently, gain a deeper understanding of the linguistic data.

Weak and strong learning

- Weak learning has made rapid progress in recent years
- Strong learning is largely irrelevant
- Semantic dependencies in the data allow the inference of syntactic dependencies
- Lattice based approaches seem to give structures that are appropriate for supporting semantic interpretation
- Hard to differentiate spurious ambiguity from "real" ambiguity without semantic information.