



LACompLing2018. Symposium on Logic and Algorithms in Computational Linguistics

Stockholm, 28 –31 August 2018

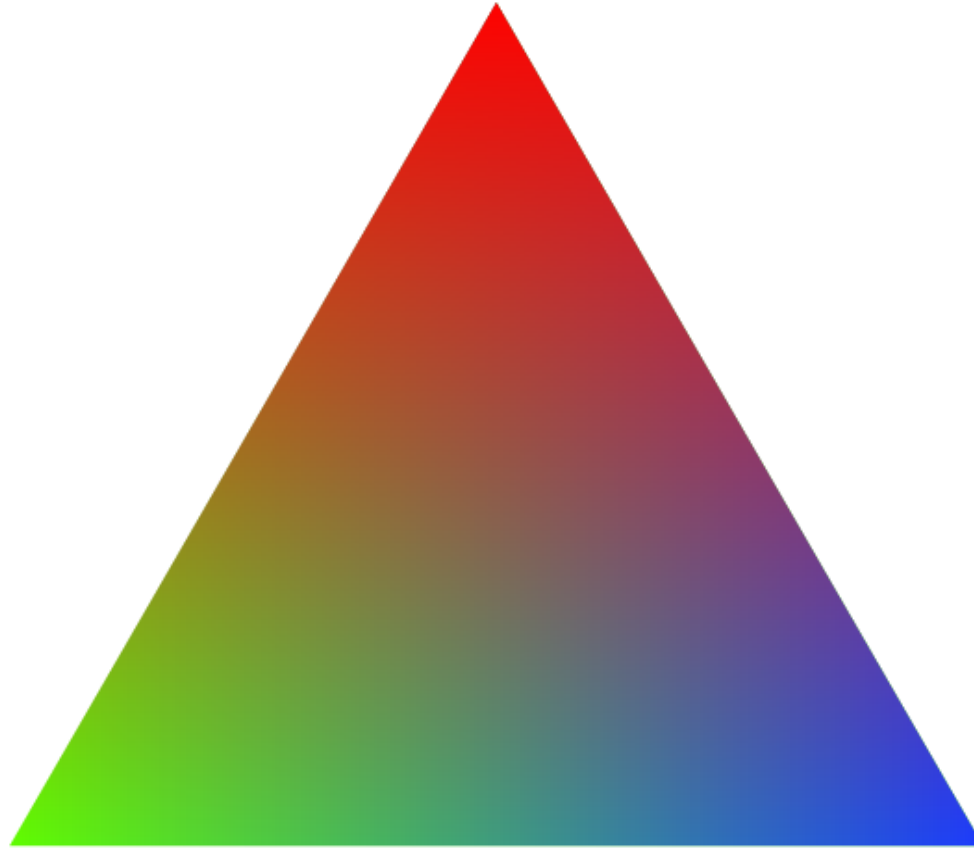
COMPLEXITY, NATURAL LANGUAGE AND MACHINE LEARNING

M. DOLORES JIMÉNEZ-LÓPEZ

GRLMC- RESEARCH GROUP ON MATHEMATICAL LINGUISTICS

UNIVERSITAT ROVIRA I VIRGILI, TARRAGONA

Complexity



**Natural
Language**

**Machine
Learning**



Complexity



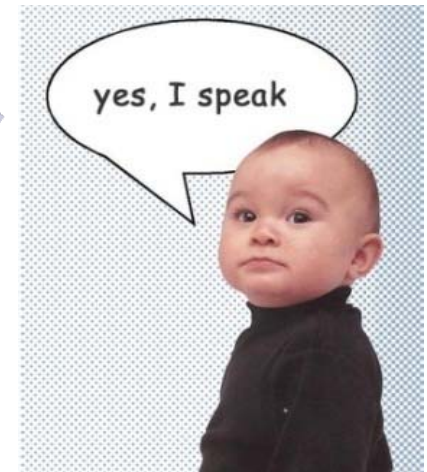
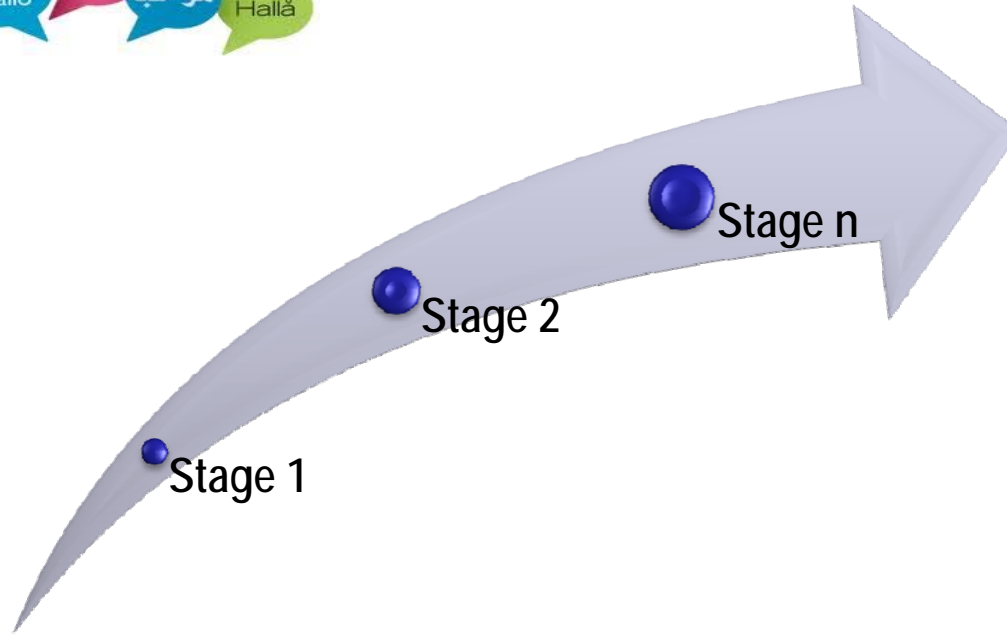
Natural Language Complexity



Machine Learning



LANGUAGE ACQUISITION: learning a first language is something every child does successfully. In every society, in every language, in every child independently of the type of education and intelligence level.



All children acquire language in the same way, regardless of the language they learn.

Children progress through distinct stages in language acquisition.

STAGES/PHASES IN FIRST LANGUAGE ACQUISITION ARE THE SAME REGARDLESS THE LANGUAGE.



Nobody has trouble speaking their mother tongue.
Nobody finds it difficult to speak their native language




ARE ALL LANGUAGE EQUALLY DIFFICULT? DO ALL LANGUAGES HAVE THE SAME LEVEL OF COMPLEXITY?






ARE ALL LANGUAGE EQUALLY DIFFICULT? DO ALL LANGUAGES HAVE THE SAME LEVEL OF COMPLEXITY?



20th Century Linguistics:
INVARIANCE OF LANGUAGE
COMPLEXITY



Some 21st Century Linguists:
DIFFERENT LEVELS
OF COMPLEXITY



ARE ALL LANGUAGE EQUALLY DIFFICULT? DO ALL LANGUAGES HAVE THE SAME LEVEL OF COMPLEXITY?

20th Century Linguistics: INVARIANCE OF LANGUAGE COMPLEXITY

- **Linguistic complexity is invariant:** All languages have the same level of complexity.
- **There are no simple languages and complex languages:** There is no reason to think that some languages are structurally more complex than others- essentially all languages are identical.



LINGUISTIC EQUI-COMPLEXITY Dogma (Kusters 2003)

ALEC Statement “All Languages are Equally Complex” (Deutscher 2009)



LINGUISTIC EQUI-COMPLEXITY Dogma

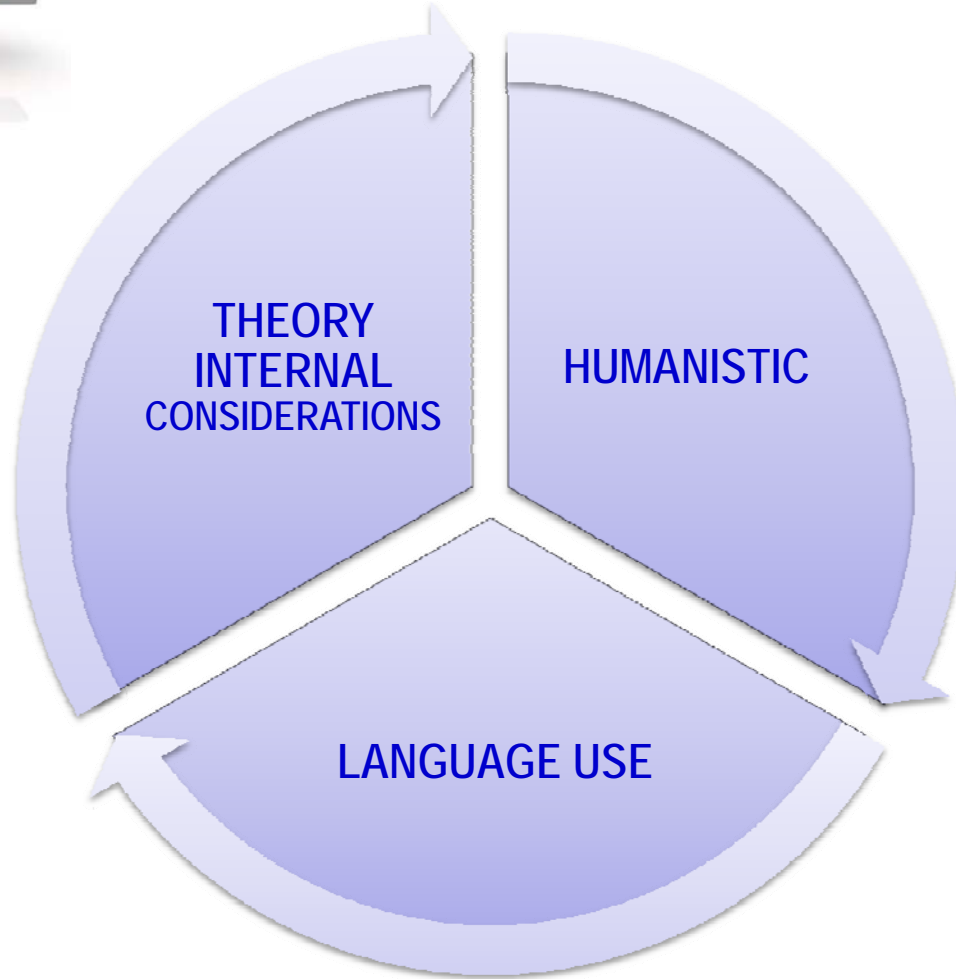
*"Objective measurement is difficult, but impressionistically it would seem that **the total grammatical complexity of any language**, counting both morphology and syntax, **is about the same as that of any other**. This is not surprising, since all languages have about equally complex jobs to do, and what is not done morphologically has to be done syntactically. Fox, with a more complex morphology than English, thus ought to have a somewhat simpler syntax; and this is the case."*

Hockett (1958)

- The **total complexity of a language is fixed** because sub-complexities in linguistic sub-systems trade off.
- **Simplicity in some domain A must be compensated by complexity in domain B, and vice versa**



The nature of universal grammar demands all languages be equally complex



Since all humans groups are in a fundamental sense “equal”, their languages must be “equal” too. Since language is the most central human cognitive faculty, to claim that human languages can differ in complexity is like claiming that human populations can differ in terms of their cognitive abilities.

Complexity in one area will always be “balance out” by simplicity in another area



Equi-complexity Dogma

All languages have the same level of complexity

Regarding complexity, languages are incommensurable

The measurement of linguistic complexity is irrelevant to the knowledge of languages and for functioning

OUTCOME

- ✓ Axiom for 20th Century Linguistics
- ✓ Their validity has rarely been subjected to systematic cross-linguistic investigation.
- ✓ Outcome: Dogmatization and the lack of empirical and theoretical research on language complexity



If languages differ in the complexity of particular subsystems. **Why all languages should be equal in their overall complexity?**

Why complexity in one grammatical area should be compensated by simplicity in another?

What mechanism could cut complexity in one area as soon as another area has become more complex?

What could be the factor responsible for equi-complexity?



ARE ALL LANGUAGE EQUALLY DIFFICULT? DO ALL LANGUAGES HAVE THE SAME LEVEL OF COMPLEXITY?

21st Century: NOT ALL LANGUAGES HAVE THE SAME LEVEL OF COMPLEXITY

There is no objective reason to argue that:

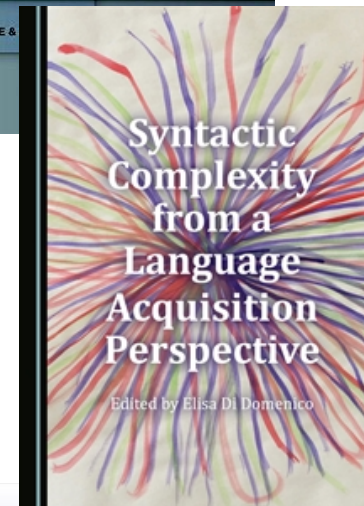
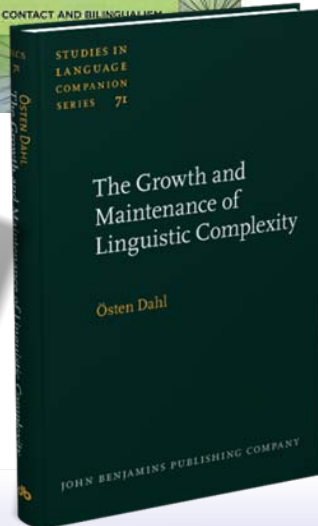
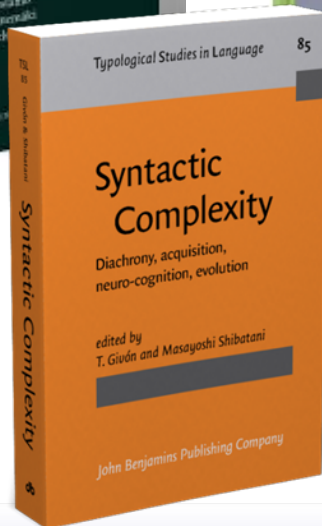
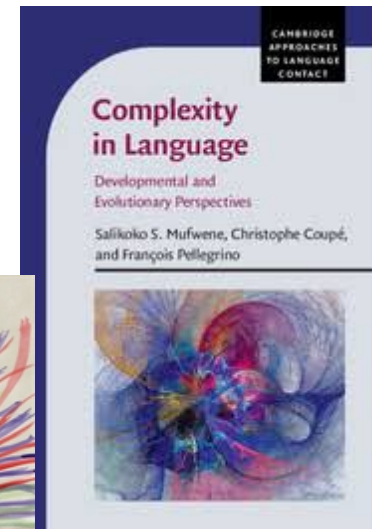
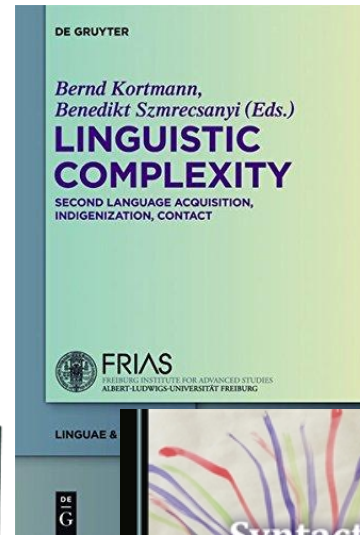
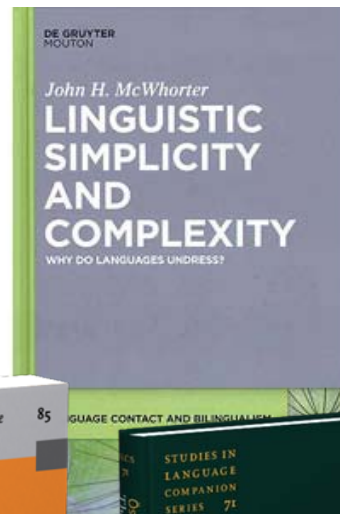
- all languages are equal in their total complexity.
- the complexity in one area is offset with simplicity in another.

*“While it is the case that all languages are roughly equal (that is, no language is six times as complex as any other, and there are no primitive languages), it is by no means the case that they are exactly equal. [...] **There is no doubt that one language may have greater overall grammatical complexity.**”*

(Dixon 1997)



MCWHORTER (2001): Special issue of the journal *Linguistic Typology*: *The world's simplest grammars are creole grammars.*





LINGUISTIC COMPLEXITY

ARE ALL LANGUAGE EQUALLY DIFFICULT?

DO ALL LANGUAGES HAVE THE SAME LEVEL OF COMPLEXITY?

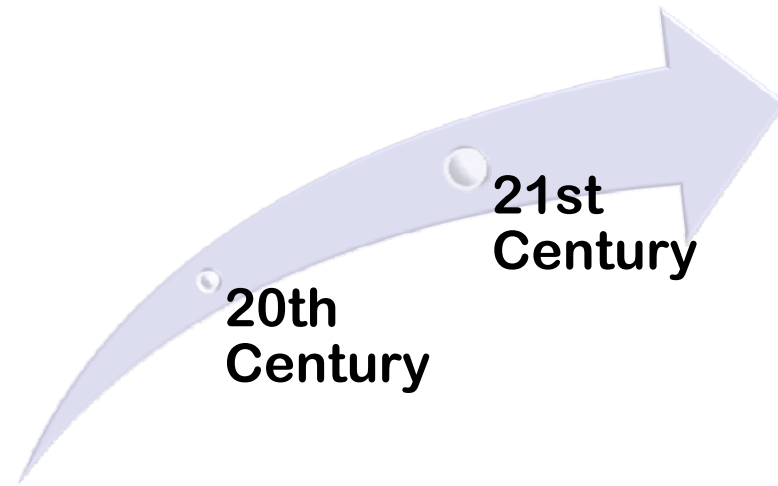
21st Century

**Large number of studies on
linguistic complexity**

20th Century

**To deny the possibility of
calculating the complexity of the
language**

WHY?



Big amount of research on complexity and complex systems in areas such as natural sciences, social sciences, computing ...

Motivated by the lack of systematic research that proves the supposed equi-complexity of languages

There is no objective reason to argue that:

- all languages are equal in their total complexity.
- the complexity in one area is offset with simplicity in another.



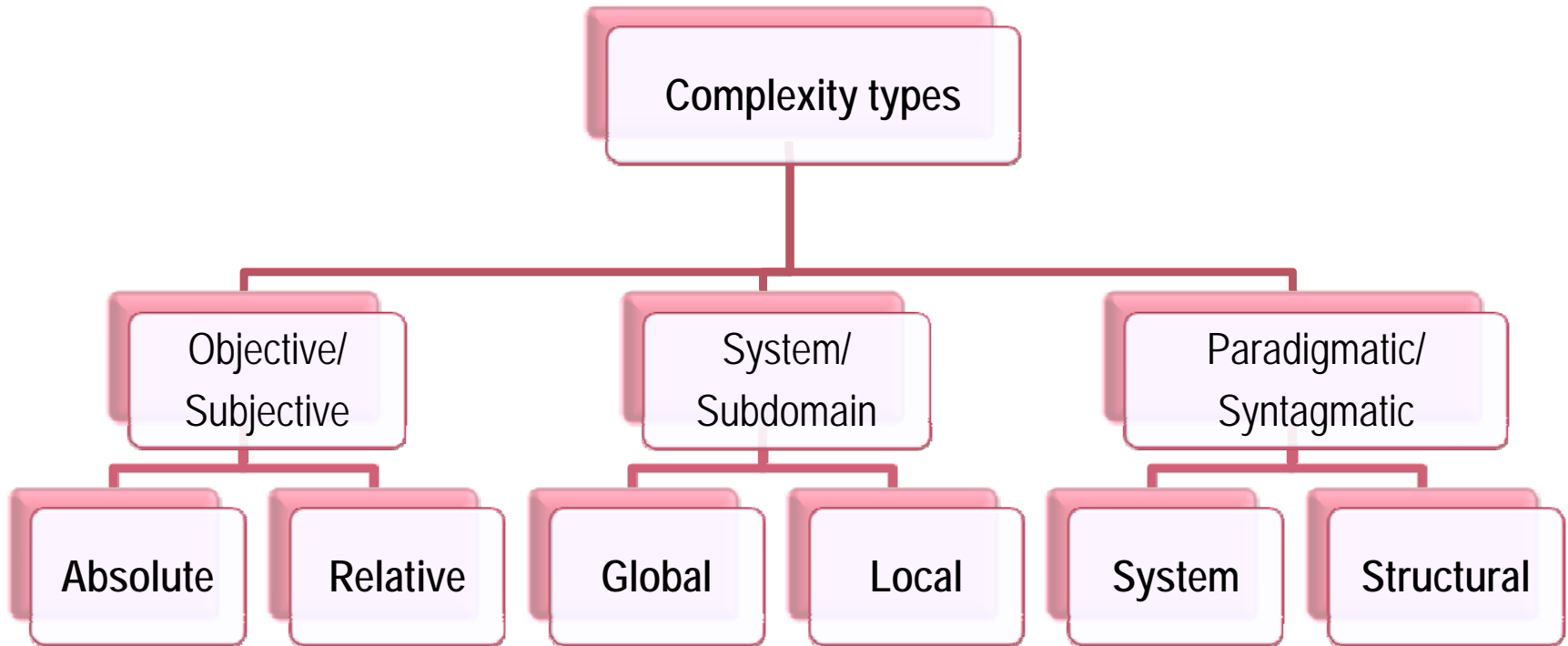


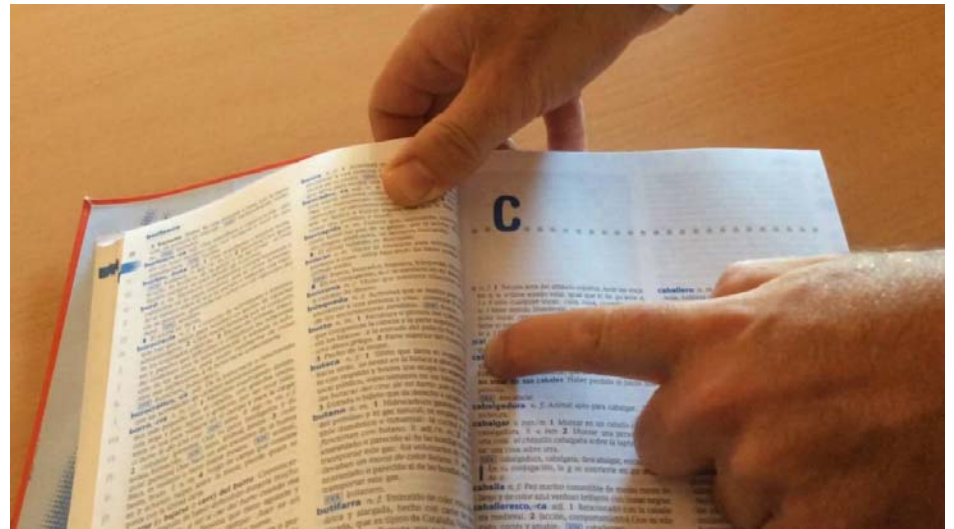
Although, in general, it seems clear **that languages exhibit different levels of complexity, it is not easy to calculate exactly those differences**

Part of that difficulty is due to **different ways of understanding complexity in natural languages.**

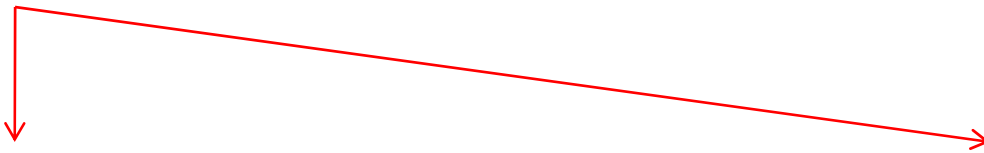


The concept of complexity is difficult to define: This leads directly to an important terminological distinction, which is crucial in discussing complexity

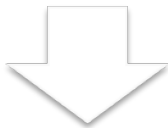




Definition of **COMPLEX**

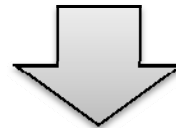


¹Composed of many related parts



ABSOLUTE COMPLEXITY

²Complicated or intricate as to be hard to understand or deal with:



RELATIVE COMPLEXITY



ABSOLUTE COMPLEXITY

Objective: an objective property of an object or a system.

Theory-oriented

Number of parts in a system.

Number of interrelations among parts.

Length of the description of a phenomenon (information-theoretical terms)

Typology

McWhorter (2001), Dahl (2004)

RELATIVE COMPLEXITY

Subjective: It takes into account language users

User-oriented

Difficulty of processing

Difficulty in language learning

Difficulty in language acquisition

Sociolinguistics, Psycholinguistics

Kusters (2003)



Amount of information needed to recreate or specify a system (or the length of the shortest possible complete description of it)



COMPLEXITY

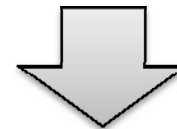


ABSOLUTE COMPLEXITY

Amount of resources that an agent spends in order to achieve some goal



COST



RELATIVE COMPLEXITY

Applies to tasks. Relative to an agent Measured in terms of "risk of failure"



DIFFICULTY

Cost and Difficulty: tasks that demand large expenditure of resources or in particular those that force the agent to or beyond the limits of his or her capacity are experienced as difficult.



GLOBAL COMPLEXITY

The overall complexity of the system.

Complexity of a language

Difficult and ambitious task

Problems:

1. **PROBLEM OF REPRESENTATIVITY:** it is very difficult to account for all aspects of grammar in such detail that one could have a truly representative measure of global complexity.
2. **PROBLEM OF COMPARABILITY:** different criteria used to measure the complexity of a grammar are incommensurable. It is not possible to quantify the complexity of syntax and morphology so that the numbers would be comparable in any useful sense..

LOCAL COMPLEXITY

Complexity of some part of the system

Complexity of a particular domain of grammar

A doable task

Problem when comparing languages:
Is the complexity of a language the sum of the complexity of its subsystems?

(Miestamo 2008, Edmonds 1999)



SYSTEM COMPLEXITY

“How to express that which can be expressed”

Properties of a language

Measures the number of subdistinctions within a category

Content of speakers competence.

Paradigmatic complexity (Moravcsik and Wirth 1986)

STRUCTURAL COMPLEXITY

“Complexity of expressions at some level of descriptions”

Properties of concrete expressions

Amount of structure of a linguistic object

The structure of utterances and expressions

Syntagmatic complexity (Moravcsik and Wirth 1986)



DIFFERENT MEANINGS OF “COMPLEXITY”

STRUCTURAL COMPLEXITY

- A formal property of texts and linguistic systems having to do with the number of their elements and their relational patterns.

COGNITIVE COMPLEXITY

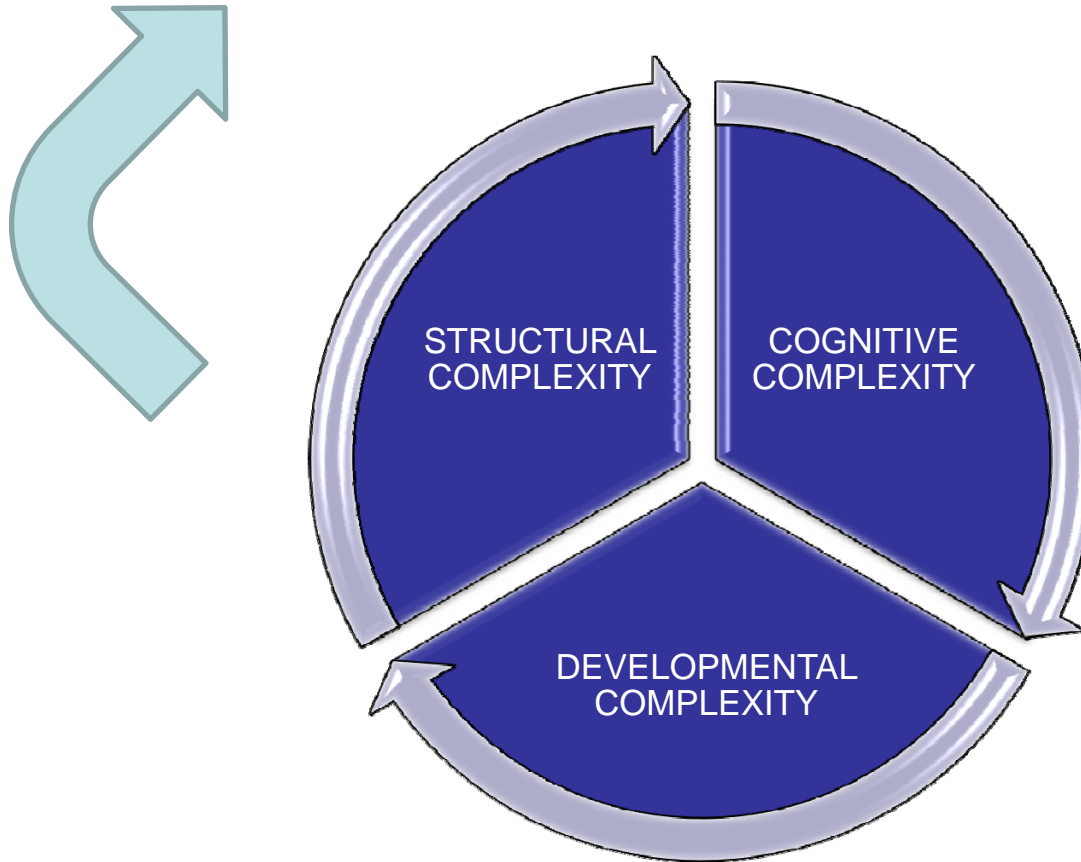
- Having to do with the processing costs associated with linguistic structures

DEVELOPMENTAL COMPLEXITY

- The order in which linguistic structures emerge and are mastered in second (and, possibly, first) language acquisition

Pallotti (2015). A simple view of linguistic complexity. *Second Language Research* 31: 117-134

Crystal: "Complexity refers to both the **INTERNAL STRUCTURING OF LINGUISTIC UNITS** and **PSYCHOLOGICAL DIFFICULTY** in **USING** or **LEARNING** them."



Crystal, D. (1997). *The Cambridge encyclopedia of language*. Cambridge: Cambridge University Press.



Besides the definition issues, we must tackle **the metrics problem**.

There is no conventionally agreed metric for measuring the complexity of natural languages.

The tools, criteria and measures vary **and depend on the specific research interests and on the definition of complexity adopted**

HOW CAN WE MEASURE COMPLEXITY?

1. Many ad hoc complexity measures have been proposed.
2. Information Theory:
 - a. **Shannon entropy**: *"A message is complex if it has a large information content, and a language is complex if sending the message in that language requires much more bandwidth than the information content of the message"*
 - b. **Kolmogorov complexity**: measures the informativeness of a given string as the length of the algorithm required to describe/generate that string. The longer the description of a linguistic structure, the more complex it is. The idea is that the shorter the output of the algorithm, the less complex is the object.
3. Computational Models
4. Theory of Complex Systems





Grammar-based. One measures and compares the degree of complexity of each grammatical component.

Number of categories

Length of description

Ambiguity

Redundancy...

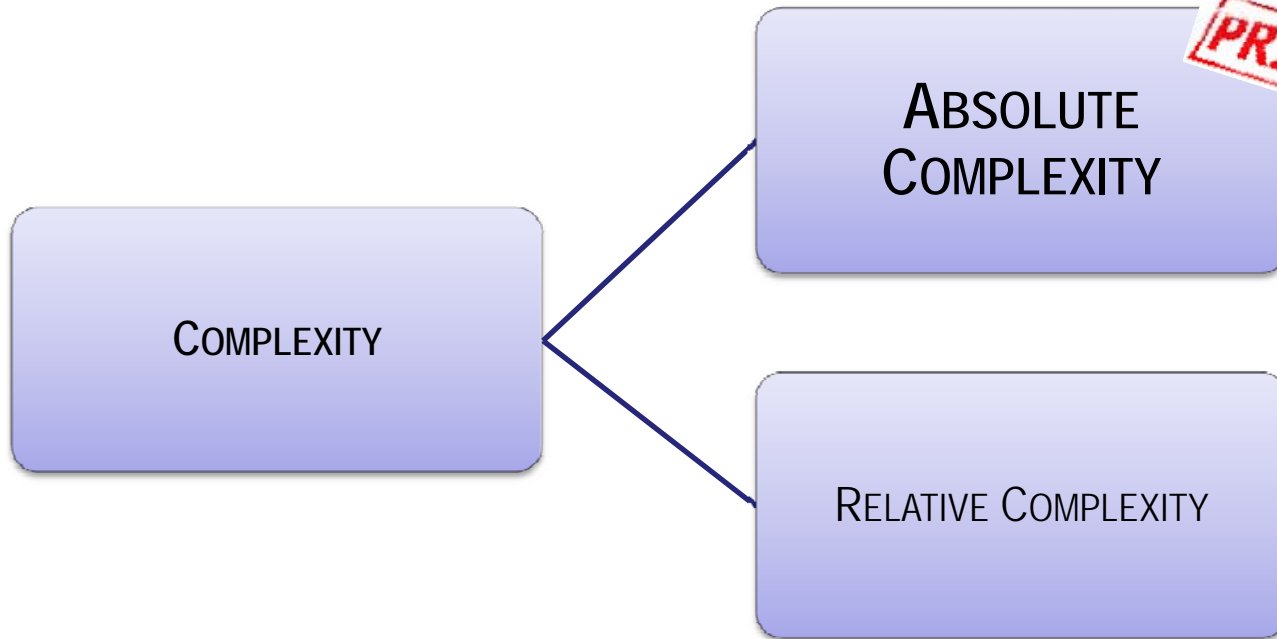


User-based. One measures complexity from the point of view of the language user

First-Language acquisition. Do some grammars take longer for the child to acquire than others?

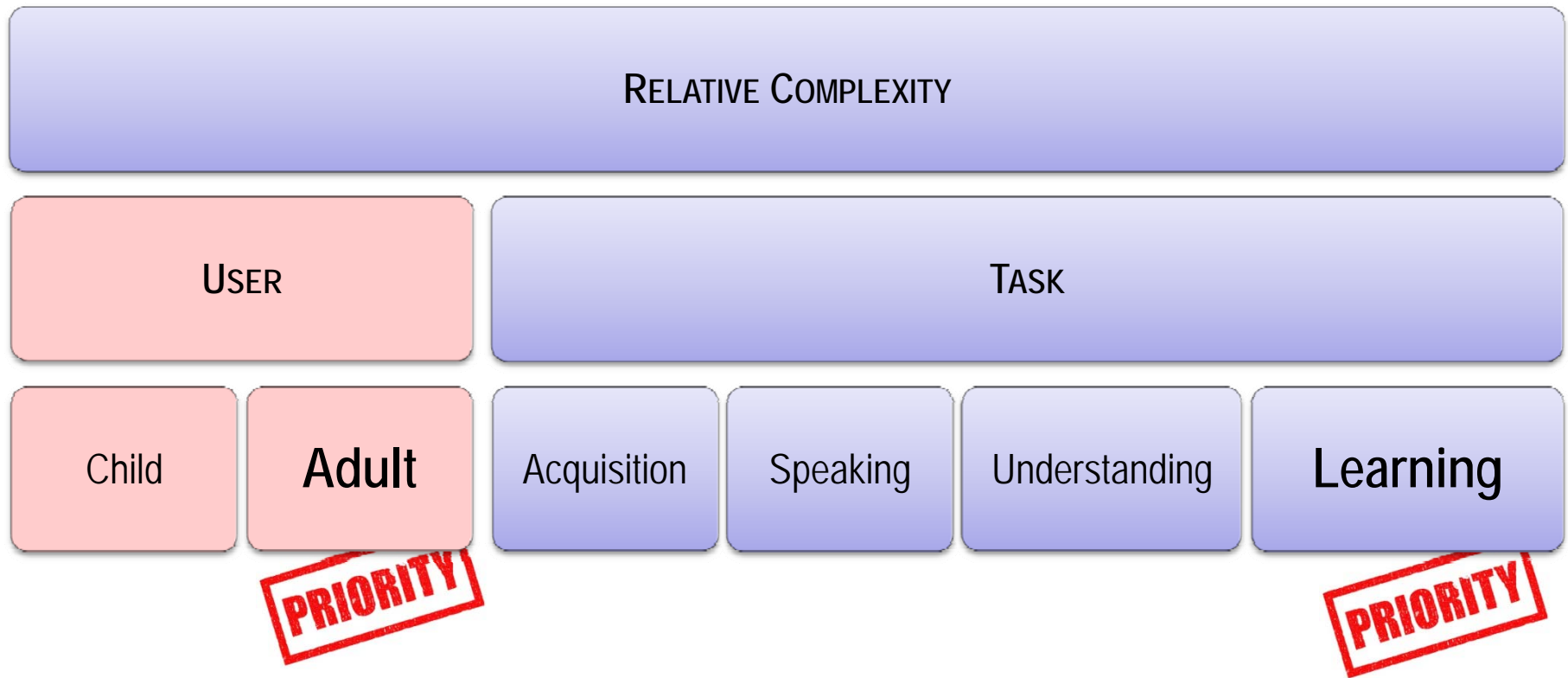
Second-language acquisition. Do some grammars take longer for the adult learner to acquire than others?

Language use. Are some grammars more difficult to use than others?



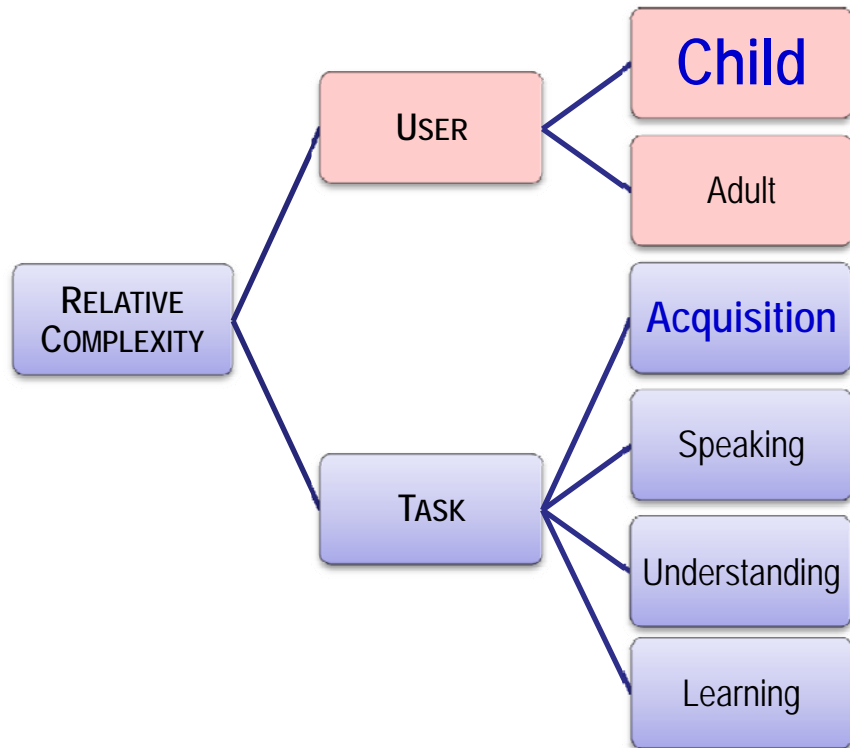
In general, researchers agree that it is **more feasible to approach complexity from an objective viewpoint** than from a subjective point of view.

The relative complexity approach --even though considered as conceptually coherent-- has hardly begun to be developed.

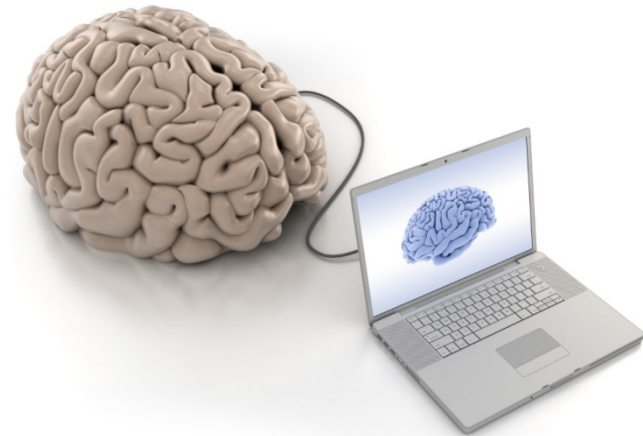


Studies that have adopted a relative complexity approach have showed some **preferences for L2 learners**. To reach a general definition of relative complexity, the primary relevance of L2 learners is not obvious.

Problems that observational and experimental models of acquisition may pose to the study of linguistic complexity

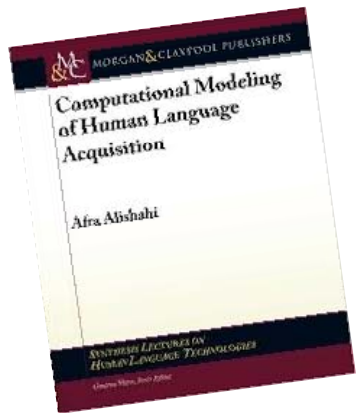
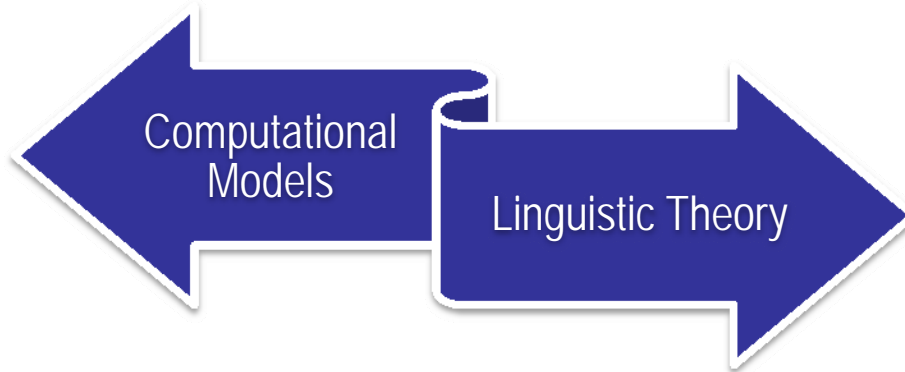


Taking into account the centrality of L1 learners, studies on complexity should consider this process to determine the differences between natural languages



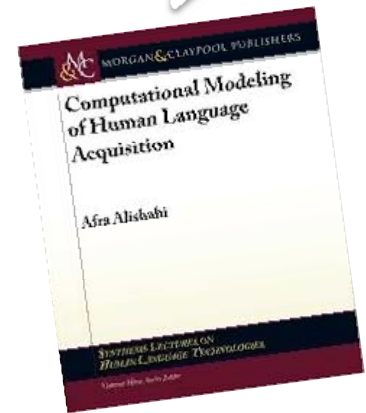
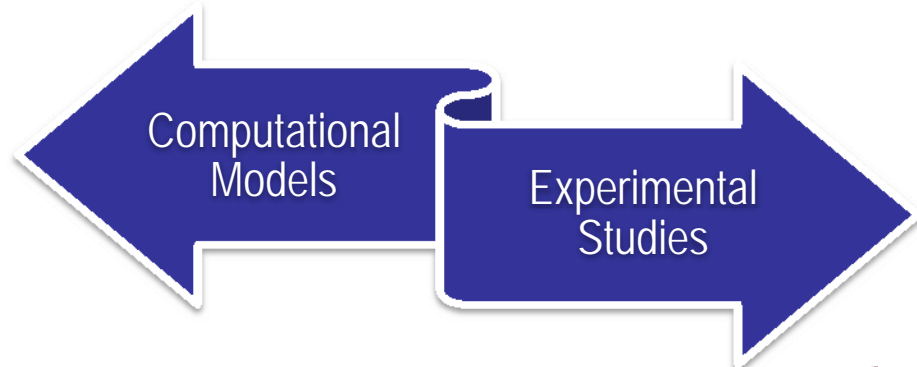
Computational models may be considered as important complementary tools that by avoiding practical problems of analyzing authentic learner productions data will make possible to consider children (or their simulation) as suitable candidates for evaluating the complexity of languages

Using computational tools for studying natural language acquisition offers many
METHODOLOGICAL ADVANTAGES



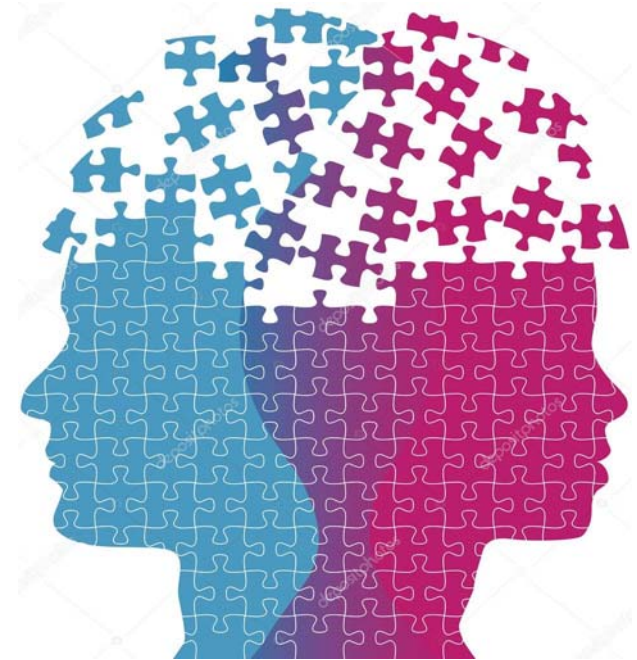
EXPLICIT ASSUMPTIONS. When implementing a computational model, every assumption of the input data and the learning mechanism has to be specified.

Using computational tools for studying natural language acquisition offers many
METHODOLOGICAL ADVANTAGES



COTROLLED INPUT: Computational models offers the possibility to manipulate the language acquisition process and see the results of that manipulation. The researcher has full control over all the input data.

Using computational tools for studying natural language acquisition offers many
METHODOLOGICAL ADVANTAGES

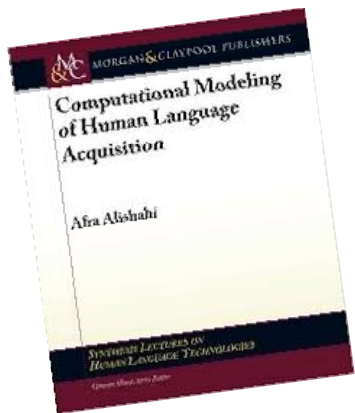


OBSERVABLE BEHAVIOR. The impact of every factor in the input or the learning process can be directly studied in the output of the model. The performance of two different mechanisms on the same data set can be compared against each other.

Using computational tools for studying natural language acquisition offers many
METHODOLOGICAL ADVANTAGES

Computational
Models

Experimental
Studies

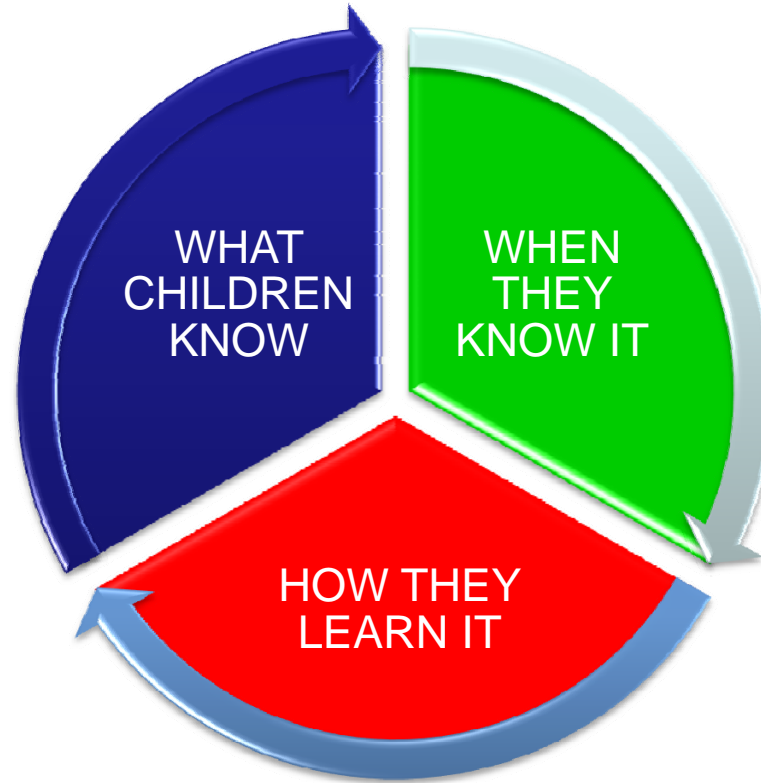


TESTABLE PREDICTIONS. Novel situations or combinations of data can be simulated and their effect on the model can be investigated.

Besides the enumerated advantages, one of the main benefits of computational models of language acquisition for determining relative linguistic complexity is the type of questions these formalisms could answer. According to Pearl (2010), language acquisition research is concerned with **three different questions**:

THEORETICAL RESEARCH

deals with the knowledge that children acquire



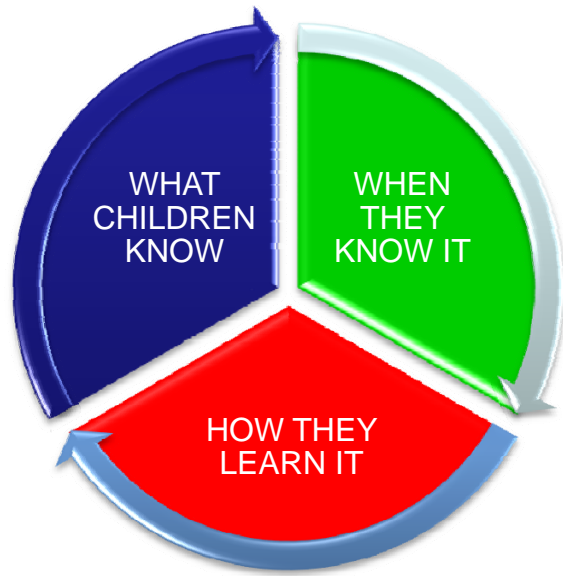
EXPERIMENTAL ANALYSIS

provides information regarding the age at which the child acquires particular linguistic knowledge

COMPUTATIONAL MODELS

can explain how the child learns a language.

Models are meant to be simulations of the child's acquisition mechanism



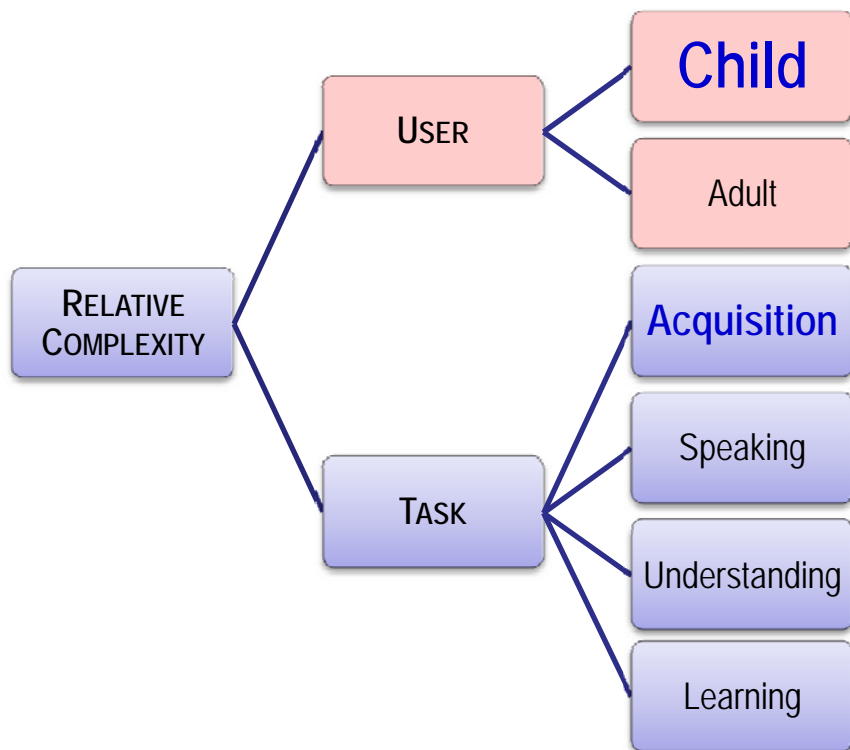
Being tools for explaining the process of natural language acquisition, computational models in general are potential good tools to deal with developmental linguistic complexity.

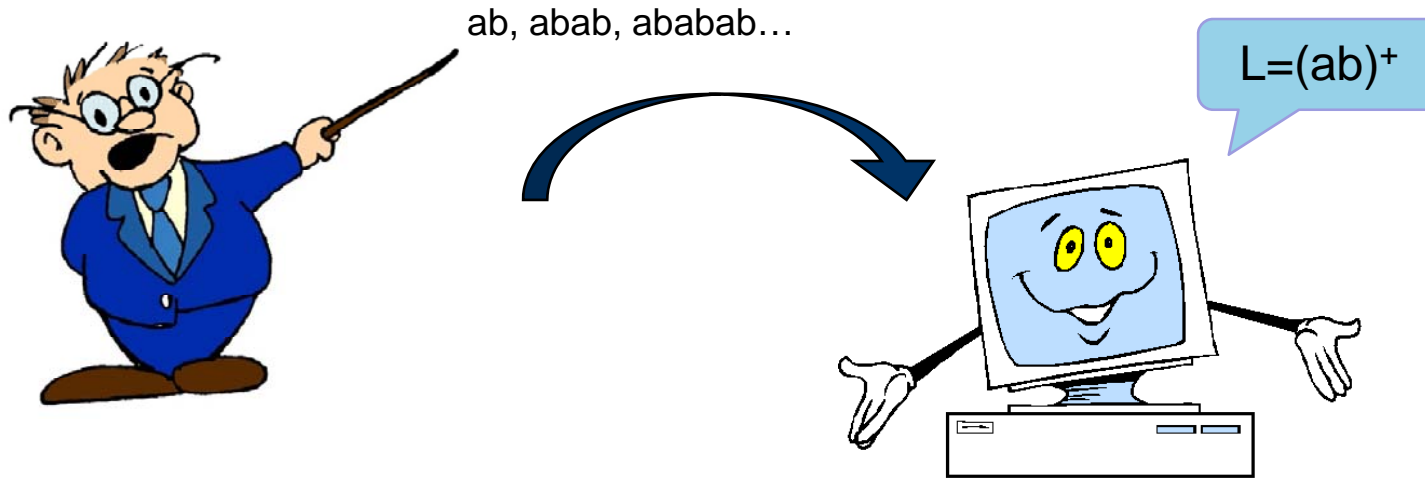
Machine Learning

To calculate the relative complexity, we propose to use a **machine learning model for first language acquisition**.

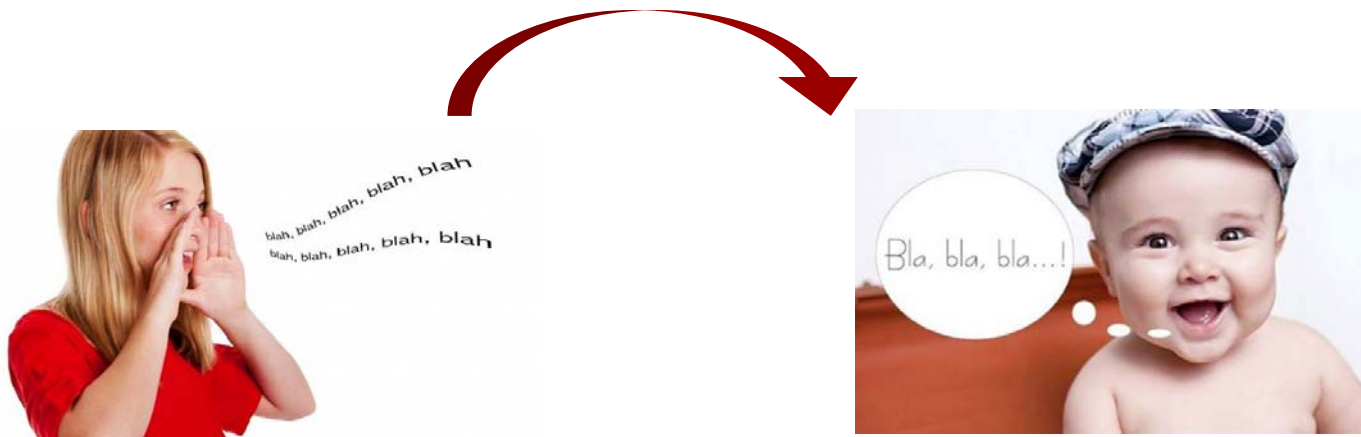
This kind of models deal with **idealized learning procedures for acquiring grammars** on the basis of exposure to evidence about languages

Allows us to reproduce the learning context of first language acquisition





Machine Learning: we provide data to a learner, and a learner (or learning algorithm) must identify the underlying language from this data.



This process have some similarities with the process of language acquisition where children received linguistic data and from them they learn their mother tongue.

Machine Learning

Two approaches

GRAMMATICAL INFERENCE

Angluin and Becerra(2011): An overview of how semantics and corrections can help language learning

GROUNDED LANGUAGE LEARNING

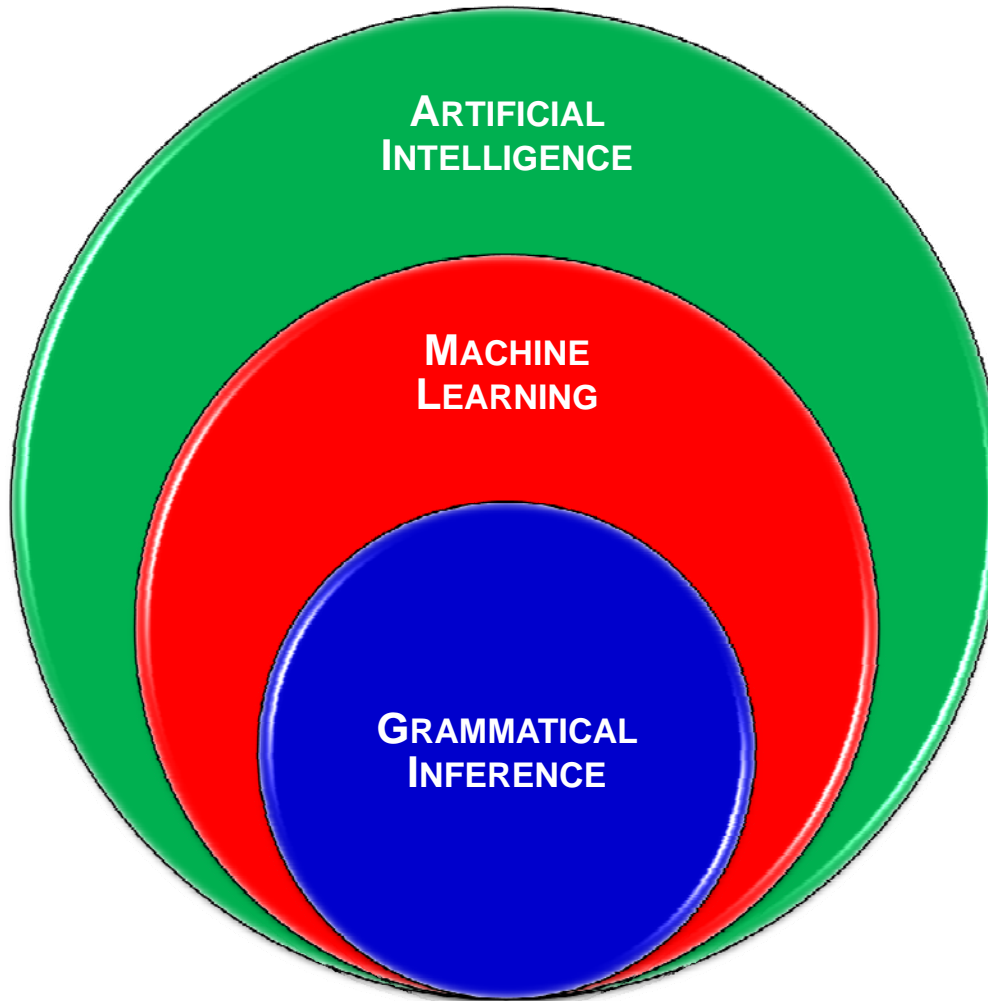
MINIATURE LANGUAGE ACQUISITION TASK

Becerra et al. (2005): A first-order-logic based model for grounded language learning.

Becerra et al. (2016): Learning language models from images with regll.

ABSTRACT SCENES DATASET

Becerra et al. (2016): Relational grounded language learning



How to create computers that are capable of intelligent behavior

Construction and study of algorithms that can learn from data

Subfield that deals with the learning of formal languages



ANGLUIN, D. AND BECERRA-BONACHE, L. (2011): An overview of how semantics and corrections can help language learning

COMPUTATIONAL MODEL FOR LEARNING A LANGUAGE



It provides the algorithm
**POSITIVE DATA AND
CORRECTIONS**

POSITIVE DATA are essential in the process of language acquisition

CORRECTIONS can play a complementary role by providing additional information that may be helpful during the process of language acquisition.

It takes into account
Semantics

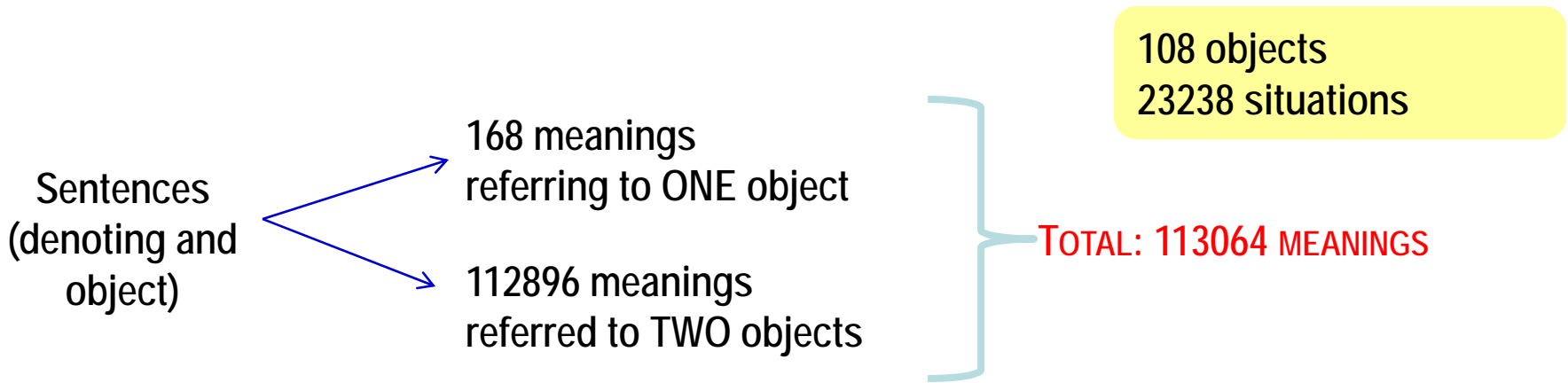
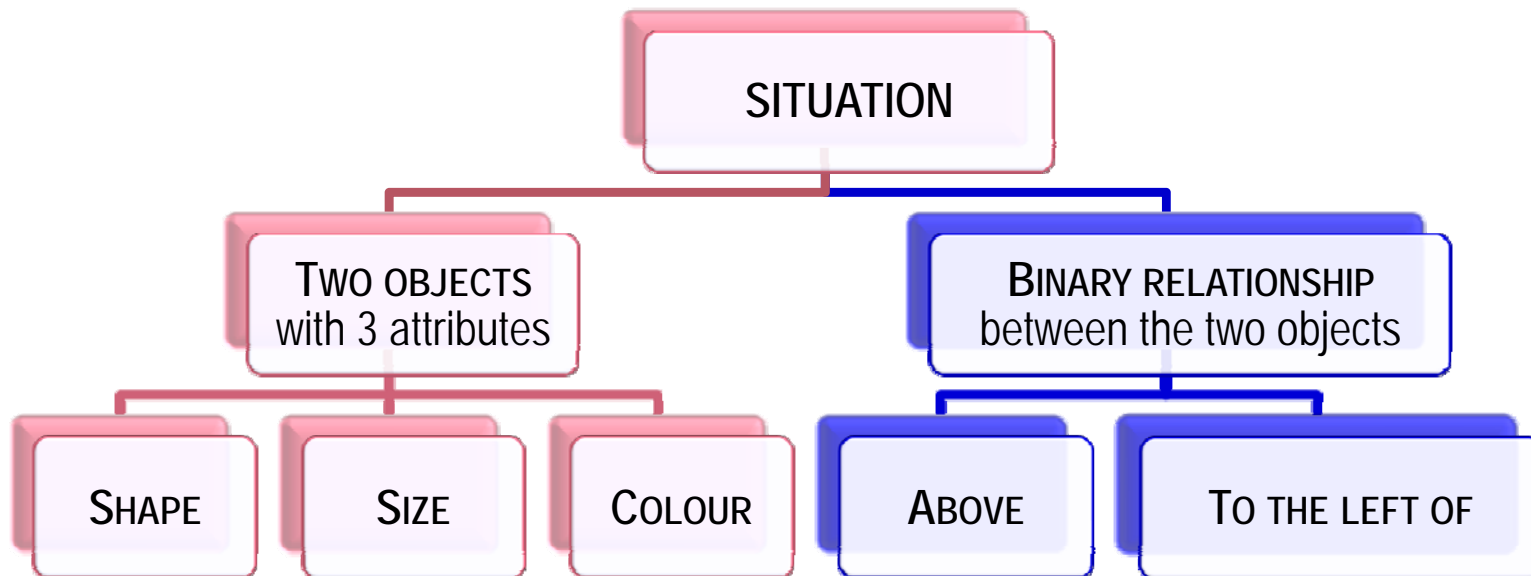
Most models reduce the problem of learning to learn the syntax.



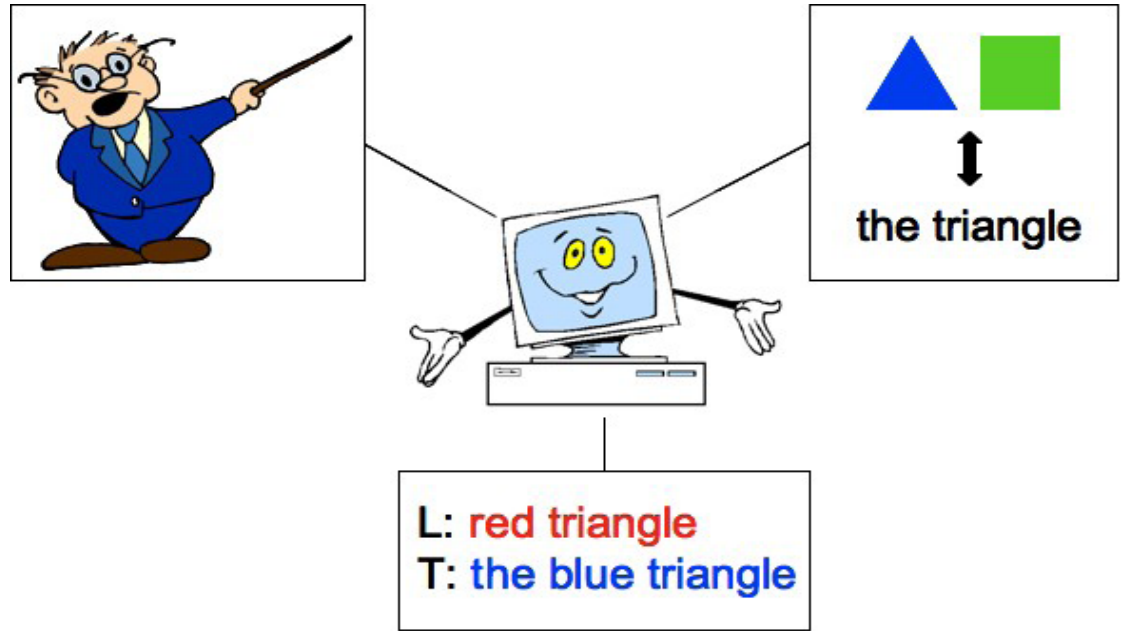
ANGLUIN, D. AND BECERRA-BONACHE, L. (2011): An overview of how semantics and corrections can help language learning

MODEL EVALUATION:

- Simplified version of Feldman Task: *MINIATURE LANGUAGE ACQUISITION TASK* (Stolcke 1994)
- Task: learn a subset of a natural language from pairs of phrases-drawings of geometric figures that have different properties.
- Considered languages: **English, German, Greek, Hebrew, Hungarian, Mandarin, Russian, Spanish, Swedish, Turkish.**

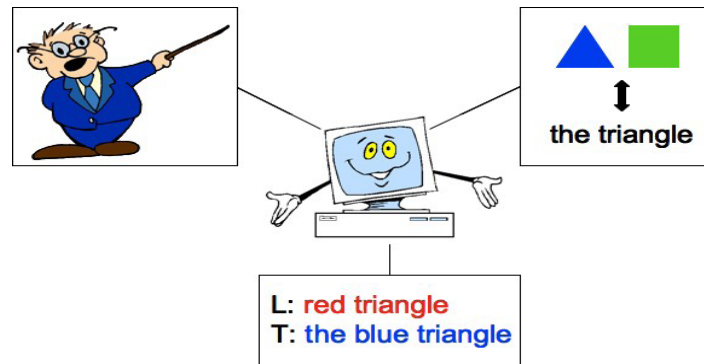


TASK: learning a grammar that allows to produce appropriate sentences in any of the possible contexts



INTERACTION:

1. A situation is randomly generated and it is presented to teacher and learner.
2. The learner tries to produce a sentence to denote one object in this situation.
3. The teacher produces a random sentence that denotes one object in the situation.
4. The learner analyzes the teacher's sentence and updates its current grammar.



MEASURE THE NUMBER OF INTERACTIONS NECESSARY TO ACHIEVE A GOOD LEVEL OF PERFORMANCE

To evaluate the performance of the learner two different measures are used:

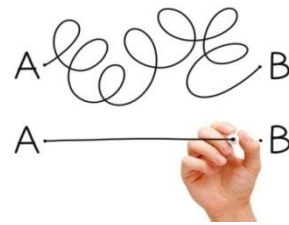
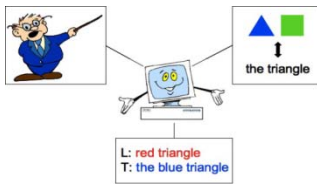
- (i) **CORRECTNESS:** correct sentences that the learner is able to produce.
- (ii) **COMPLETENESS:** learner's ability to produce ALL possible correct sentences

FINAL LEARNER'S GOAL: given a situation, to **ONLY** produce correct sentences and to be able to produce **ALL** the correct sentences.

A learner reaches a level p of performance if both correctness and completeness are at least p . In the experiments, the teacher and learner interact until the learner achieves a level of performance $p = 0.99$.

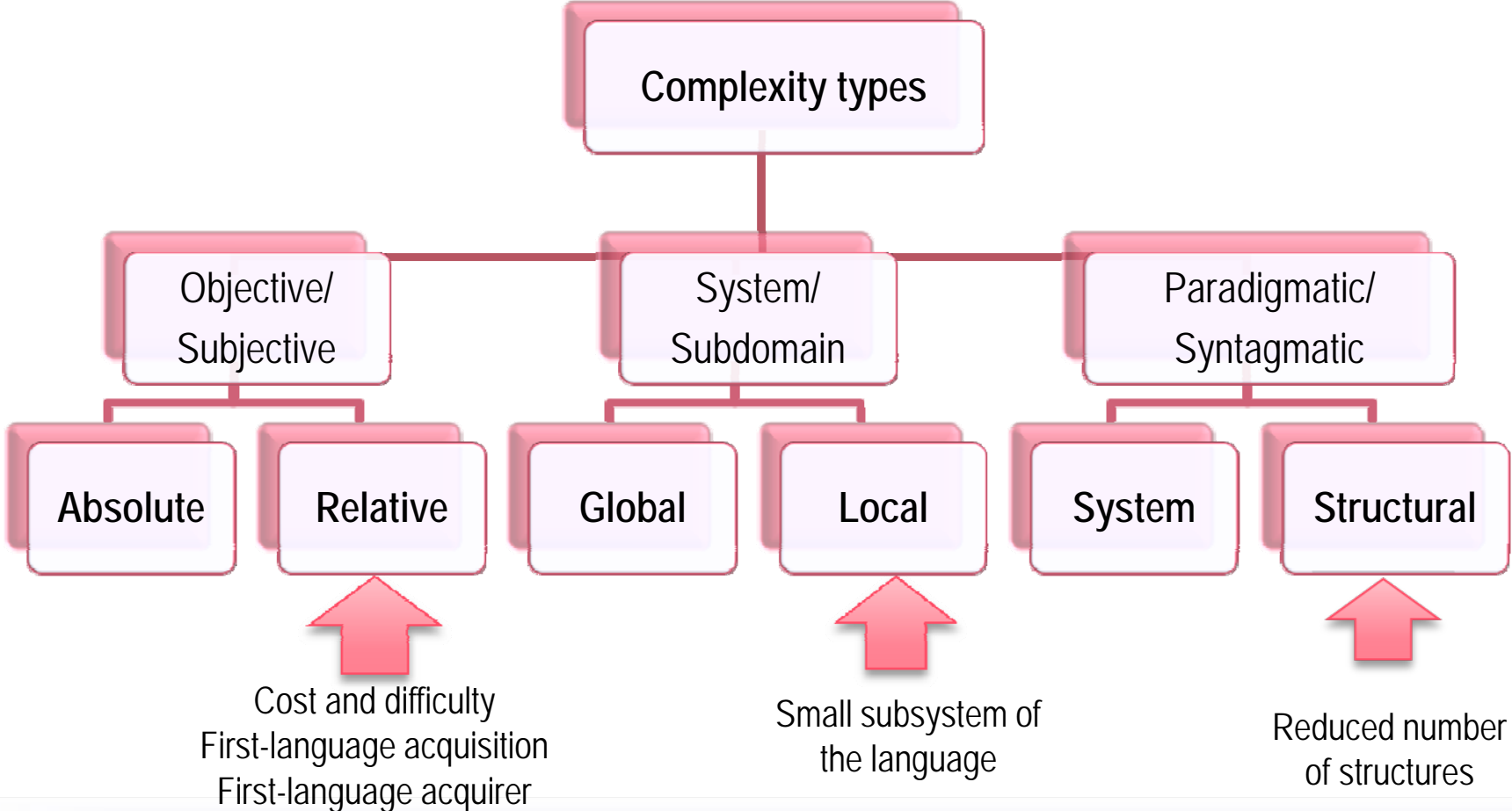
Level	0.60	0.70	0.80	0.90	0.95	0.99
English	200	200	300	400	500	700
German	200	300	300	400	550	800
Greek	400	500	700	1500	2200	3400
Hebrew	200	300	400	500	650	900
Hungarian	200	300	350	450	550	750
Mandarin	200	200	300	400	500	700
Russian	450	500	850	1750	2350	3700
Spanish	200	300	350	500	600	1000
Swedish	200	300	300	400	600	1000
Turkish	200	200	300	400	550	800

Number of interactions necessary for the learner to get a performance level of $p = 0.99$. Extracted from Angluin & Becerra 2010



Calculate the number of interactions to acquire a good level of performance

- Calculate the cost / difficulty to acquire a language.



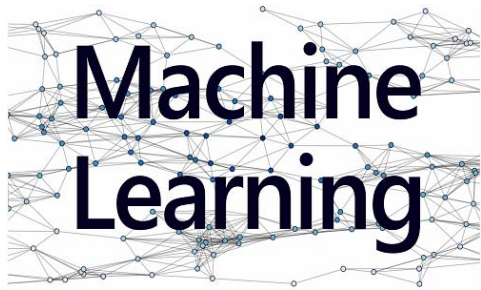
Learning a language is a challenging task that children have to face during the first years of their life.

Children learning their native language **need to map the words they hear to their corresponding meaning in the scene they observe**

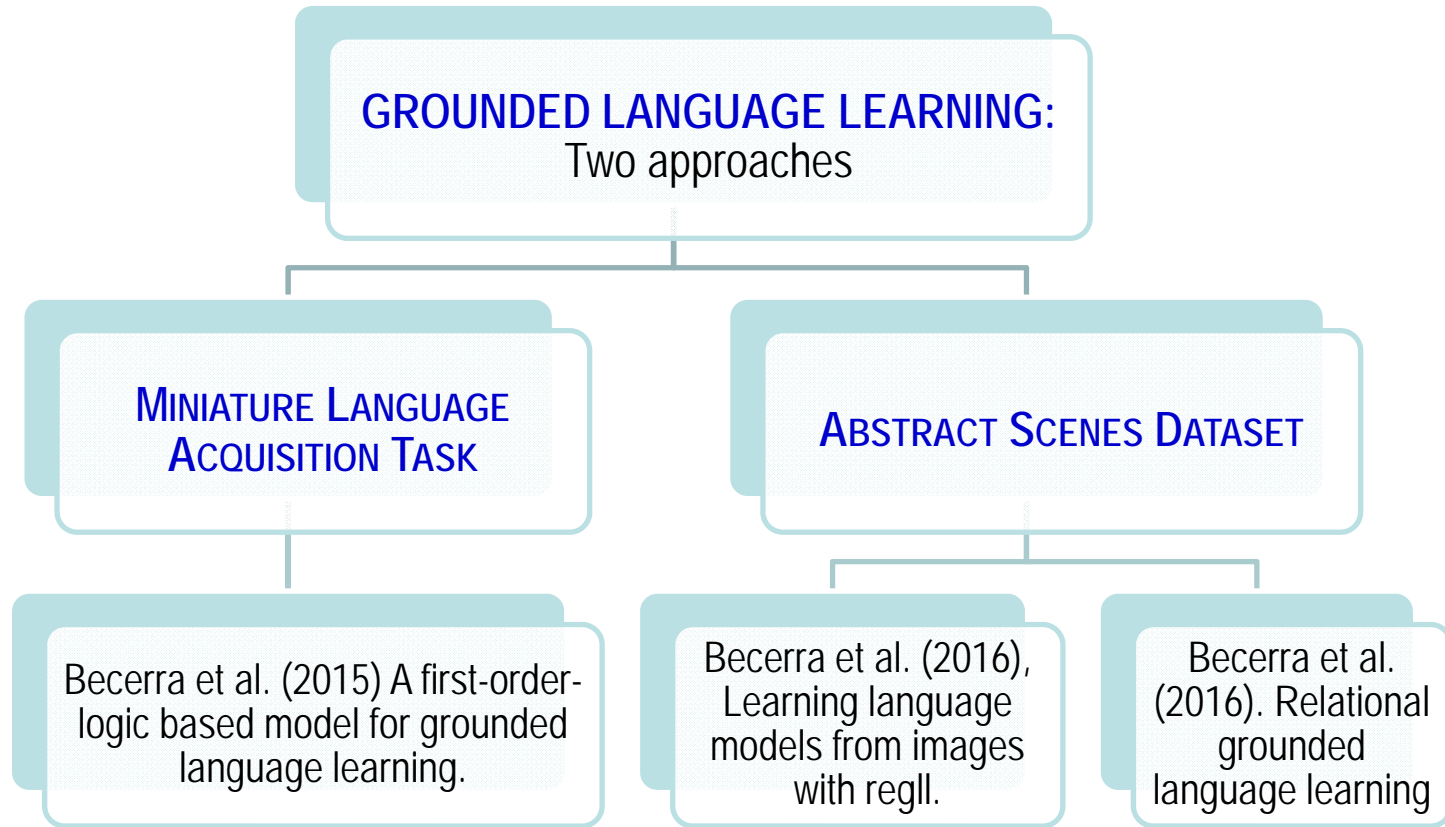
Children have to face, among others, the problems of:

- ✓ **REFERENTIAL UNCERTAINTY** (i.e., they may perceive many aspects of the scene that are not related to the utterance they hear)
- ✓ **ALIGNMENT AMBIGUITY** (to discover which word in the utterance refers to which part of the scene).



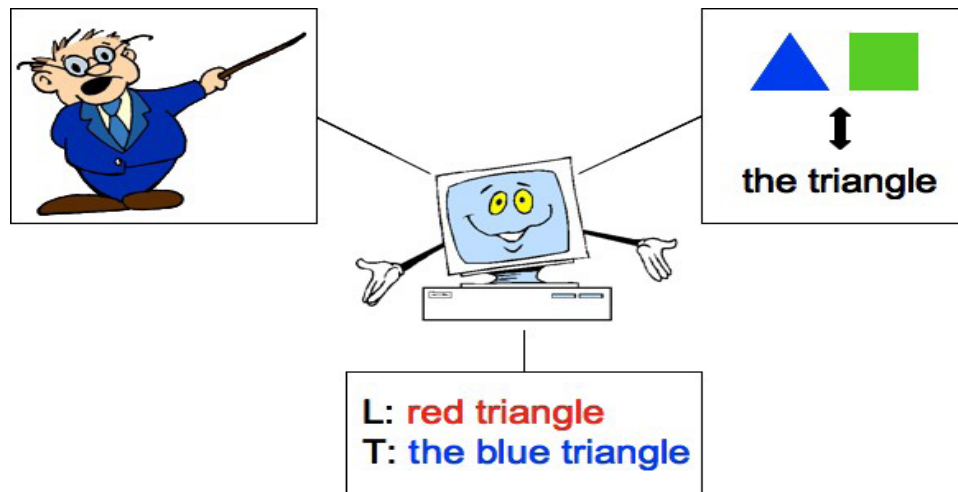


Taking into account all these aspects, Becerra-Bonache et al. developed an **artificial system** that, without any language-specific prior knowledge, is able to learn language models from **pairs consisting of a sentence and the context in which this sentence has been produced**



These systems are inspired by some previous works (Angluin & Becerra 2010,2011,2016)

MINIATURE LANGUAGE ACQUISITION TASK



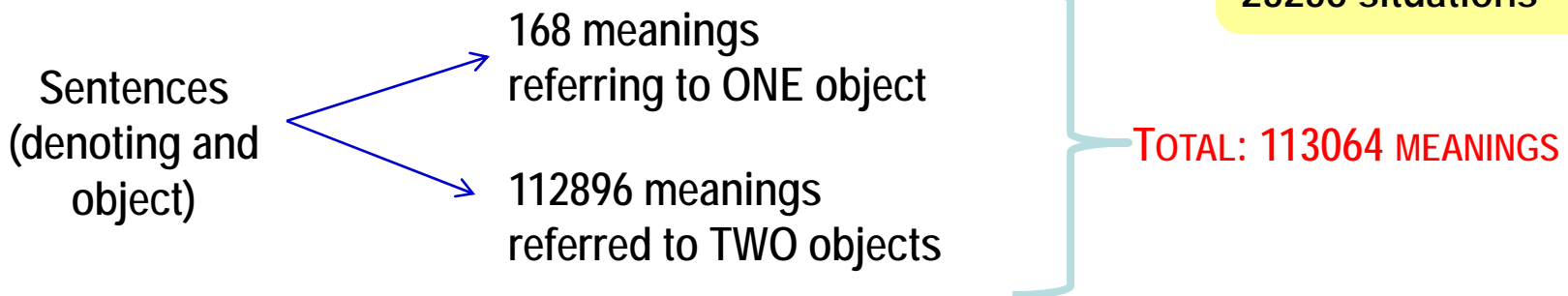
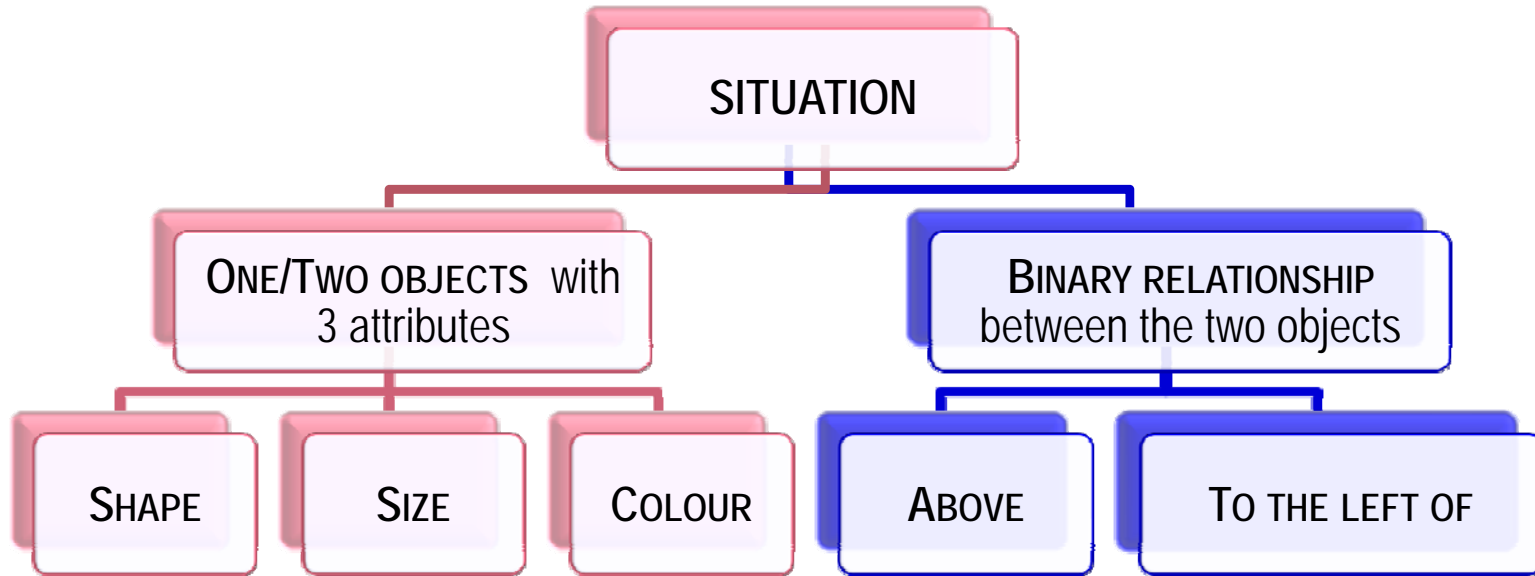
A system based on **INDUCTIVE LOGIC PROGRAMMING TECHNIQUES**

AIM: to learn a **mapping between N-GRAMS** (sequences of words) and **MEANINGS**

The model was tested:

- Simplified version of Feldman Task: **MINIATURE LANGUAGE ACQUISITION TASK** (Stolcke 1994)
- Task: learn a subset of a natural language from sentences-pictures pairs of geometric figures that have different properties

DATA SET: noun phrases that refer to the color, shape, size and position of one or two geometric figures





“a red square to the left of a triangle”



```
{obj(1,clr(1,re),shp(1,sq),sz(1,bg),  
obj(2,clr(2,gr),shp(2,tr),sz(2,bg),  
rp(1,lo,2)}
```

[a, red, square, to, the, left, of, a, triangle]

INPUT TO THE LEARNING ALGORITHM: Pairs made up of **PHASES** and the **CONTEXT** in which this phrase have been produced:

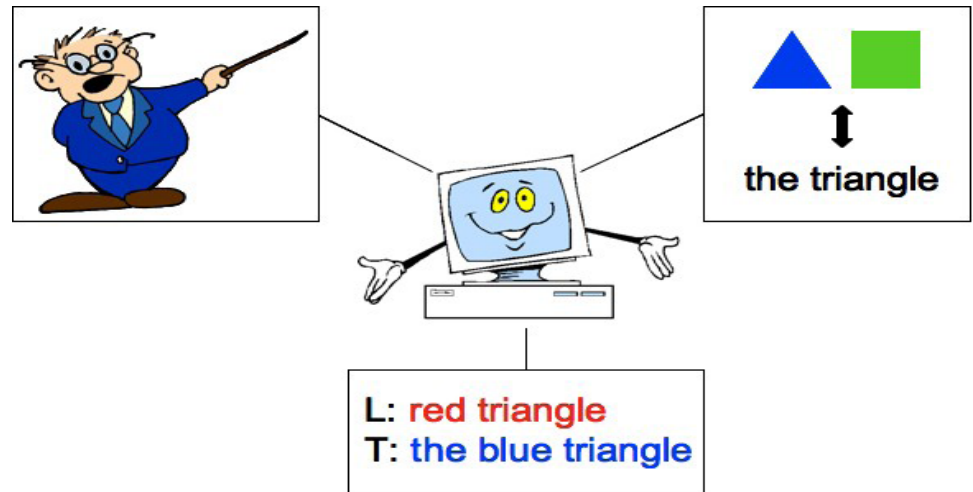
- ✓ **Phrases:** a sequence of words (n-grams)
- ✓ **Context:** a set of ground atoms (first-order logic based representation). Atoms describe properties and relationships between objects.

The **meaning** of phrases **are not provided** in the training set. The learner has to discover the meaning

The **context is just a description of what the learner can perceive** in the world

Phrases cannot refer to something that is not in the context. They can give just a partial description of the context

MINIATURE LANGUAGE ACQUISITION TASK



THE MODEL:

Can explain the gradual learning of simple concepts and language structure

Experiments with 3 languages (English, Dutch, Spanish)

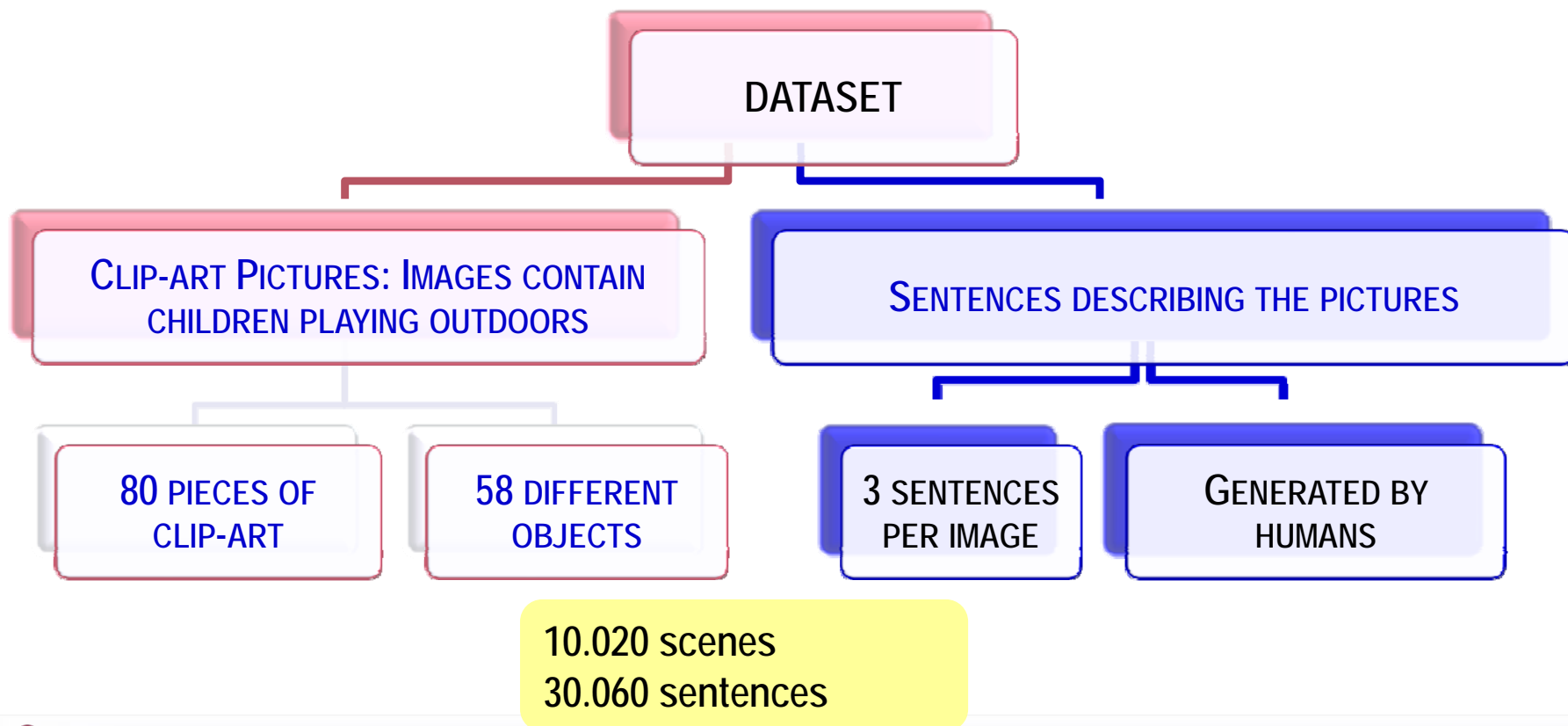
The system learns a language model that can be used to:

- ✓ Understand
- ✓ Generate
- ✓ Translate

ABSTRACT SCENES DATASET



An improvement of the previous model: A system that deals with a more challenging dataset



10.020 scenes
30.060 sentences

DATASET: This dataset was created using Amazon's Mechanical Turk (AMT)



AMT workers were asked **to create scenes from 80 pieces of clip art** depicting a boy and a girl with different poses and facial expressions, and some other objects, such as toys, trees, animals, hats, etc.

Then, a new set of workers were **asked to describe the scenes using one or two sentences description**; the descriptions should use basic words that would appear in a children's book.

In total, the dataset contains 10.020 images and 60.396 sentences

An EXAMPLE OF A SCENE



s_3s.png	0	3	469	31	2	0
p_7s.png	1	7	178	89	2	1
hb0_10s.png	2	10	100	250	1	0
hb1_4s.png	3	4	391	248	1	1
a_4s.png	4	4	205	98	1	0
c_7s.png	5	7	87	181	1	0
t_4s.png	7	4	279	115	1	1

“Mike is wearing a hat with horns”, “Mike kicks a ball to Jenny”, “An owl sits in the tree”

We can see:

- ✓ how **the dataset encodes the objects in the scene**;
- ✓ and some of the **human-written descriptions for that scene**.

It is worth noting that even if we know which objects are present in the image and their position, the **alignment between clip-art images and sentences is not given**, that is, we do not know which actions are depicted in the image (e.g., playing, eating) and which words can be used to describe them (e.g., s3s.png is called sun)

“Mike is kicking
the ball.”



The system learns from PAIRS (S,I) consisting of a **SENTENCE (S)** and an **IMAGE (I)**, where the sentence S (partially) describes the image I.

“Mike is kicking
the ball.”



[\$start, mike, is, kicking,
the, ball, \$stop]



A sentence is represented as a **sequence of words (n-grams)**.

“Mike is kicking
the ball.”



[\$start, mike, is, kicking,
the, ball, \$stop]

[object(o1), sky(o1, sun), color(o1, yellow), size(o1, big), ..., object(o3), human(o3, boy), pose(o3, pose2), expression(o3, happy), object(o4), human(o4, girl), pose(o4, pose3), expression(o4, surprised), ..., object(o6), clothing(o6, glasses), color(o6, violet), object(o7), toy(o7, ball), sport(o7, soccer), act(o3, wear, o6), ...]

For the images, a **basic pre-processing step transforms the information provided by the dataset into a context C** , by using a **FIRST-ORDER LOGIC BASED REPRESENTATION**

Contexts are made up of a **set of ground atoms** that describe **properties and relationships between the objects in the image.**

“Mike is kicking
the ball.”



[\$start, mike, is, kicking,
the, ball, \$stop]

[object(o1), sky(o1, sun), color(o1, yellow), size(o1, big), ..., object(o3), human(o3, boy), pose(o3, pose2), expression(o3, happy), object(o4), human(o4, girl), pose(o4, pose3), expression(o4, surprised), ..., object(o6), clothing(o6, glasses), color(o6, violet), object(o7), toy(o7, ball), sport(o7, soccer), act(o3, wear, o6), ...]

The **MEANING OF AN N-GRAM** is whatever is in common among all the contexts in which it can be used.

It is worth noting that a context describes what the learner can perceive in the world and, in contrast to other approaches, **the meaning is not explicitly represented, the learner has to discover it.**

“Mike is kicking
the ball.”



[\$start, mike, is, kicking,
the, ball, \$stop]

[object(o1), sky(o1, sun), color(o1, yellow), size(o1, big), ..., object(o3), human(o3, boy), pose(o3, pose2), expression(o3, happy), object(o4), human(o4, girl), pose(o4, pose3), expression(o4, surprised), ..., object(o6), clothing(o6, glasses), color(o6, violet), object(o7), toy(o7, ball), sport(o7, soccer), act(o3, wear, o6), ...]

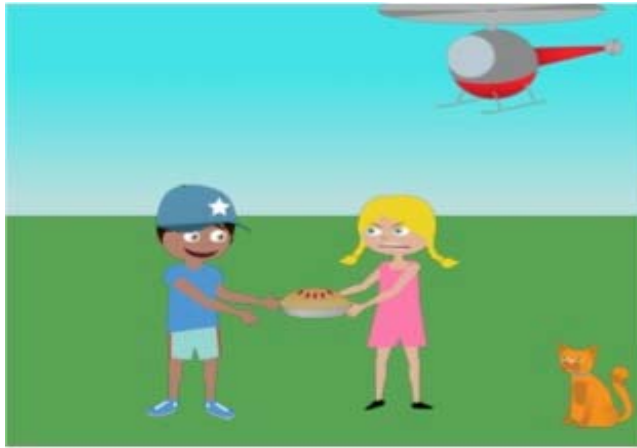


**LANGUAGE
MODEL**

Using **INDUCTIVE LOGIC PROGRAMMING TECHNIQUES**, the **system learns a mapping between n-grams and a semantic representation of their associated meaning.**

Experiments showed that the system was able to learn such a mapping and use it for a variety of purposes. **The system is able:**

- ✓ To learn in noisy environments
- ✓ To learn the meaning of words
- ✓ To generate relevant sentences for a given scene

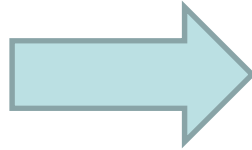


1. the cat is sitting next to jenny
2. jenny is holding a pie
3. mike is wearing a blue cap
4. mike is wearing a blue hat
5. mike is wearing a hat



1. mike is wearing a pirate hat
2. jenny is in the sandbox
3. mike is kicking the soccer ball
4. jenny is angry at mike
5. mike is in the sandbox

“Mike is kicking the ball.”



Linguistic complexity

We propose to use this artificial system to study the complexity of languages from a relative point of view.

The system is linguistically well motivated:

- ✓ the input given to the system has similar properties to those of the input received by children from their learning environment,
- ✓ and the system has no previous knowledge about the language to be learnt.

The system allows to perform cross-linguistic analysis:

- ✓ a unique algorithm is used to learn any language, which could be equivalent to the innate capacity that allows humans to acquire a language.

How to calculate the difficult/cost of learning a language by using this approach?

Counting the number of examples needed for the system to achieve a good level of performance in a given language

- Calculate the cost / difficulty to acquire a language.



To **evaluate the performance of the system**, different measures can be used:

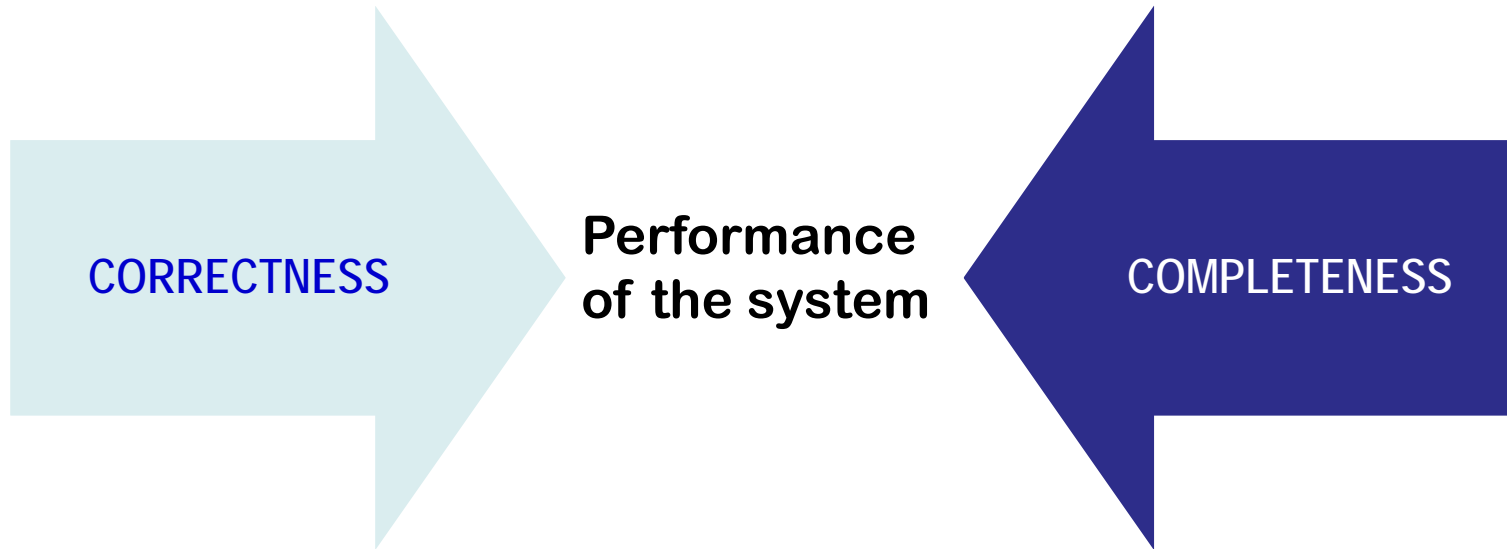
- (i) **PRECISION OF LEARNED REFERENTIAL MEANINGS**
- (ii) **SYSTEM CAPACITY TO PROVIDE CORRECT SENTENCES TO SCENES NOT PREVIOUSLY SEEN = CAPACITY TO GENERATE RELEVANT SENTENCES CONNECTING N-GRAMS**
- (iii) **CORRECTNESS:** Given a set of correct denoting sentences for a given image, the fraction of learner's sentences that are in the correct denoting set.
- (iv) **COMPLETENESS:** Given a set of correct denoting sentences for a given image, is the fraction of the correct denoting sentences that appear in the set of learner's sentences



1. the cat is sitting next to jenny
2. jenny is holding a pie
3. mike is wearing a blue cap
4. mike is wearing a blue hat
5. mike is wearing a hat



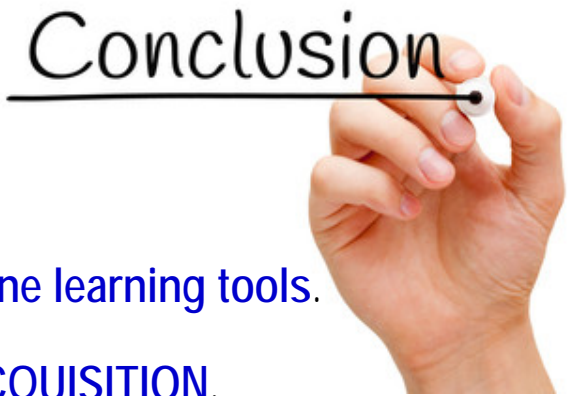
1. mike is wearing a pirate hat
2. jenny is in the sandbox
3. mike is kicking the soccer ball
4. jenny is angry at mike
5. mike is in the sandbox



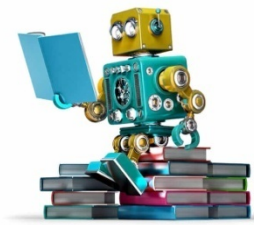
PROBLEM: IT IS NOT TRIVIAL TO SPECIFY WHICH IS THE SET OF CORRECT DENOTING SENTENCES, this is, there is not a GOLD STANDARD TO EVALUATE THE MODEL.

SOLUTION: to define a **LANGUAGE MODEL** to generate the gold standard that will be used to evaluate the performance of the language learning model

Conclusion



- A preliminary approach to study **linguistic complexity with machine learning tools**.
- To quantify the cost/difficulty in **CHILDREN FIRST LANGUAGE ACQUISITION**.
- Advantages of those models for calculating linguistic complexity:
 - They focus on the **LEARNING PROCESS**
 - They do not require **ANY PRIOR LANGUAGE-SPECIFIC KNOWLEDGE**
 - They learn **INCREMENTALLY**
 - They use **REALISTIC DATA**
 - They allow reproducing the **SAME CONTEXT AND THE SAME CONDITIONS FOR THE ACQUISITION OF ANY LANGUAGE**
 - Unlike experiments with children: **AVOID THE PROBLEM OF THE INFLUENCE OF EXTERNAL FACTORS** that can condition the acquisition process



Calculate the number of interactions to acquire a good level of performance

- Calculate the cost / difficulty to acquire a language.

Unique algorithm to learn any language

- Innate capacity that allows humans to acquire a language.

The learner does not have previous knowledge about the language

- It represents the child who has to acquire a language by exposing himself to it.

With the same algorithm not all languages require the same number of interactions

- The difficulty / cost to acquire different languages is not the same. **LANGUAGES DIFFER IN RELATIVE COMPLEXITY**

Conclusion



- ✓ In general, recent work on language complexity takes an absolute perspective of the concept while the **relative complexity approach** –even though considered as conceptually coherent-- **has hardly begun to be developed.**

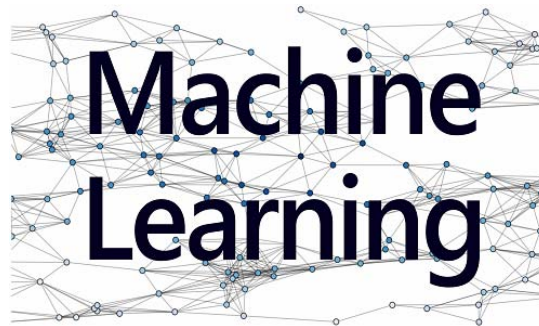
- ✓ **COMPUTATIONAL MODELS OF LANGUAGE ACQUISITION MAY BE A WAY TO REVERT THIS SITUATION.**
 - ✓ Parallelism with the process of language acquisition
 - ✓ Experimental results

- ✓ **STUDIES ON LINGUISTIC COMPLEXITY MAY HAVE IMPORTANT IMPLICATIONS BOTH FROM A THEORETICAL AND FROM A PRACTICAL POINT OF VIEW**



“The issue of linguistic complexity is complex”

Diane Larsen-Freeman, Preface





FFI2015-69978-P, Ministerio de Economía y Competitividad: “**ALGORITMOS DE INFERENCIA GRAMATICAL PARA MEDIR LA COMPLEJIDAD RELATIVA DE LAS LENGUAS NATURALES**”

Thanks



UNIVERSITAT
ROVIRA I VIRGILI



campus
d'excel·lència
internacional
catalunya sud

campus
de excel·lència
internacional
cataluña sur

campus
of international
excellence
southern catalonia

www.urv.cat