In Search of the Computational Primitives of Language

Aniello De Santo
he/him

MeLo Lab
Dept. of Linguistics

aniellodesanto.github.io
aniello.desanto@utah.edu

Computation and Theory Building

[...] this is a confusion of two quite separate issues, ***simulation and explanation****. [...] What we are* ***really*** *interested in [...] is explanation — in developing models that help us* ***understand how it is that people behave*** *that way, not merely demonstrating that we can build an artifact that behaves similarly.*

(Kaplan, 1995)

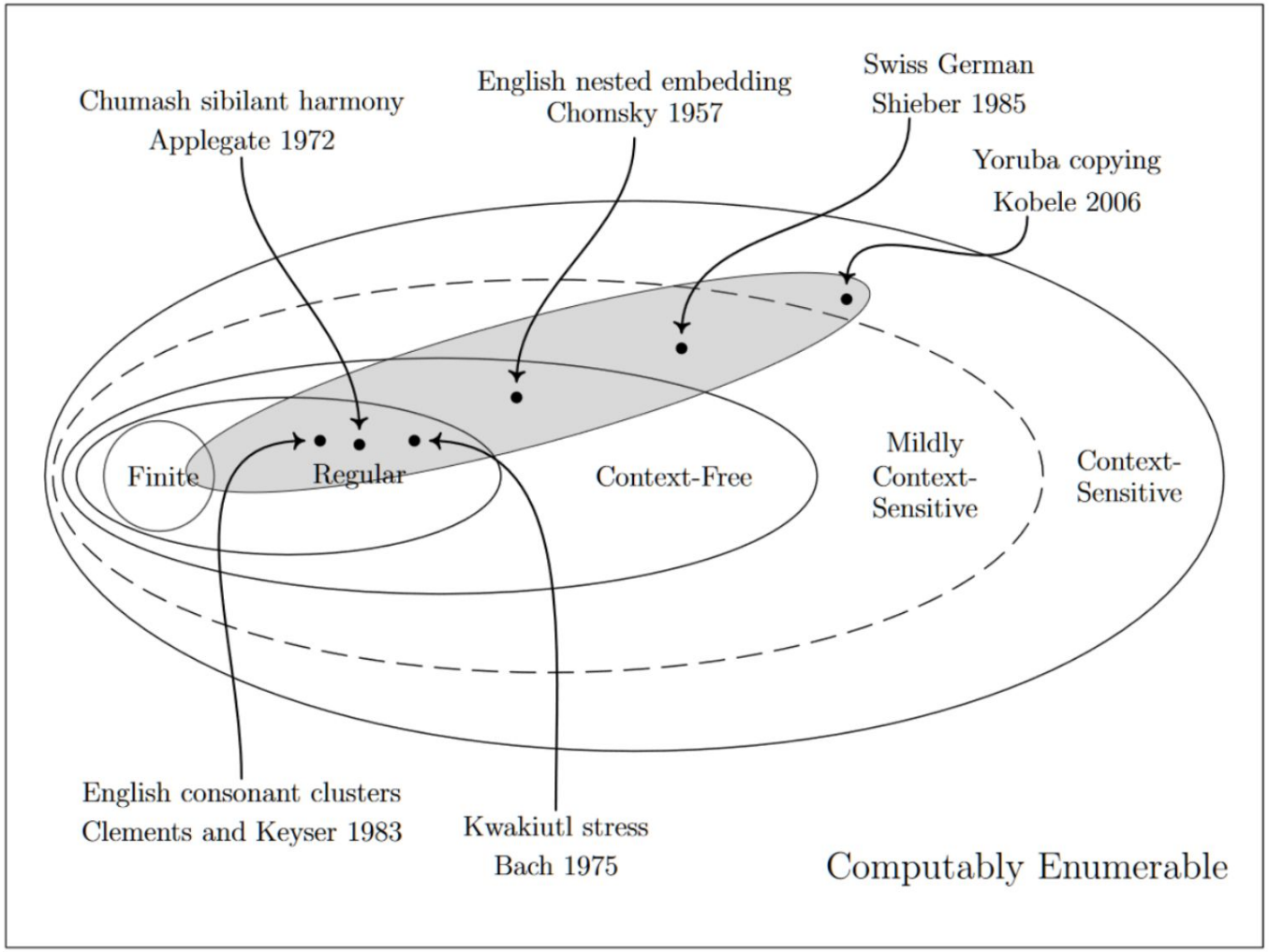
Computation and Theory Building

[...] this is a confusion of two quite separate issues, ***simulation and explanation****. [...] What we are* ***really*** *interested in [...] is explanation — in developing models that help us* ***understand how it is that people behave*** *that way, not merely demonstrating that we can build an artifact that behaves similarly.*

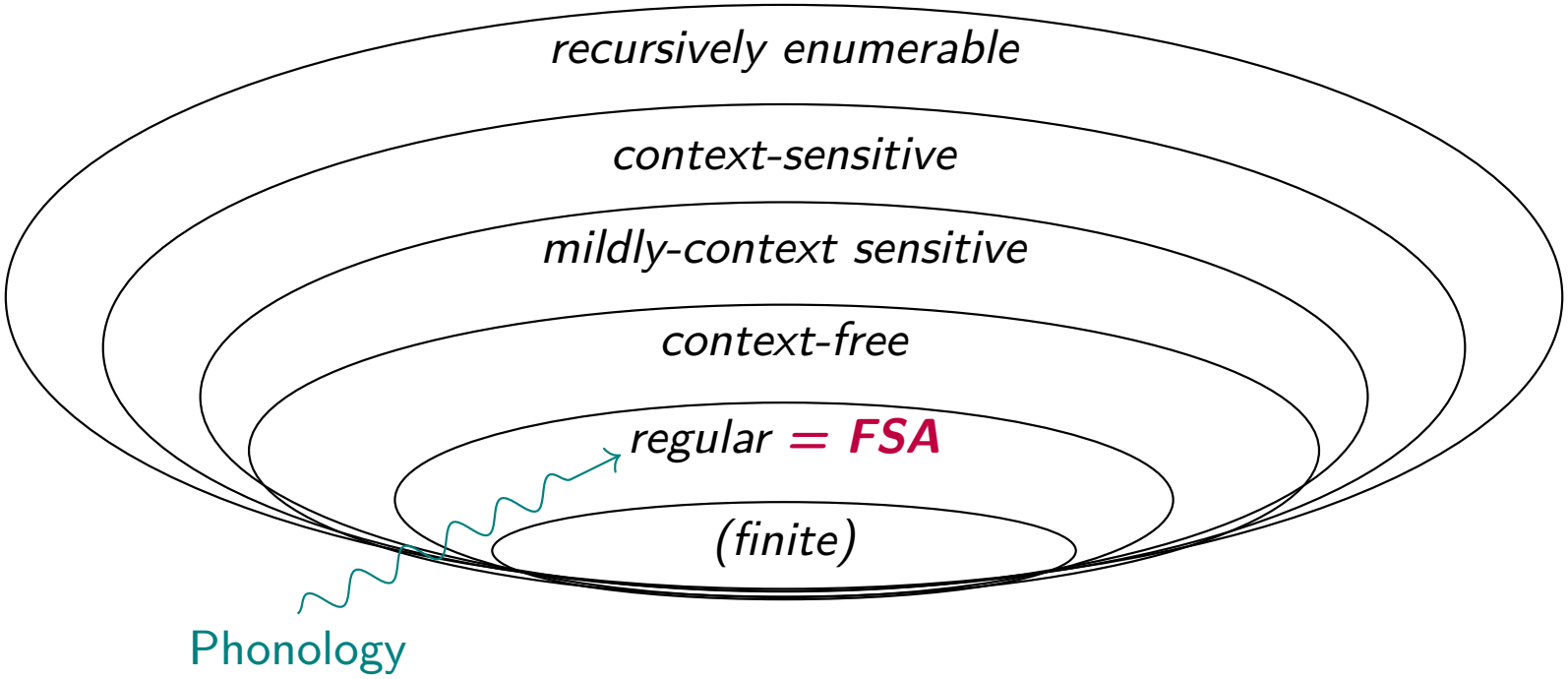
(Kaplan, 1995)

Interpretability for the win!

A Lens: Formal Language Theory



Spoken Languages' Phonology as a Regular System



Kaplan and Kay (1994)

Local Phonotactic Dependencies

1 Intervocalic voicing in Italian

Forbid voiceless segments in between two vowels

(1) a. * /kasa/

b. /kaza/

→ cf. orthography: *"casa"*

Intervocalic voicing is Strictly Local (SL)

- Forbid voiceless segments in-between two vowels: *V[-voice]V
- **Italian**: *ase, *ise, *ese, *isi, . . .

$ k a s a $ $ k a z a $

Local Phonotactic Dependencies

1 Intervocalic voicing in Italian

Forbid voiceless segments in between two vowels

(1) a. * /kasa/

b. /kaza/

→ cf. orthography: *"casa"*

Intervocalic voicing is Strictly Local (SL)

- Forbid voiceless segments in-between two vowels: *V[-voice]V
- **Italian**: *a**s**e, *i**s**e, *e**s**e, *i**s**i, . . .

$ k a **s** a $ $ k a z a $

Local Phonotactic Dependencies

1 Intervocalic voicing in Italian

Forbid voiceless segments in between two vowels

(1) a. * /kasa/

b. /kaza/

→ cf. orthography: *"casa"*

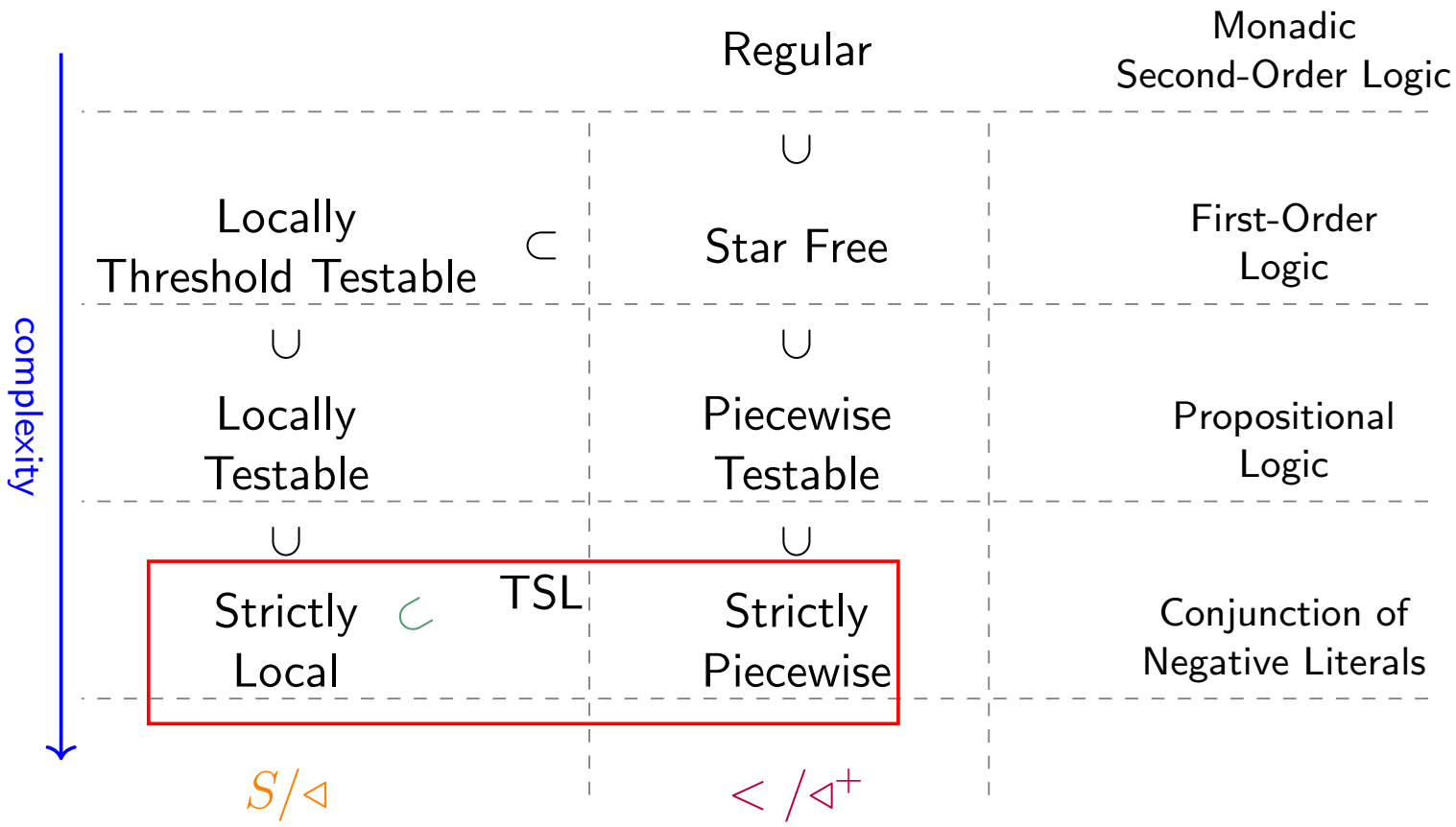
Intervocalic voicing is Strictly Local (SL)

- Forbid voiceless segments in-between two vowels: *V[-voice]V
- **Italian**: *a**s**e, *i**s**e, *e**s**e, *i**s**i, . . .

* $ k a s a $

^{ok} $ k a z a $

Beyond Automata: Subregular Languages[1]



[1] McNaughton & Papert (1976), Heinz (2011), Chandlee & Heinz (2014), De Santo & Graf (2019), De Santo & Rawski (2022), a.o.

FLT, Linguistics, and LLM Expressivity

Saturated Transformers are Constant-Depth Threshold Circuits

William Merrill[*] [†] **Ashish Sabharwal** [*] **Noah A. Smith**[*] [‡]
[*] Allen Institute for AI [†] New York University [‡] University of Washington
{ashishs,noah}@allenai.org

Between Circuits and Chomsky:
Pre-pretraining on Formal Languages Imparts Linguistic Biases

Michael Y. Hu[1] **Jackson Petty**[2] **Chuan Shi**[1] **William Merrill**[1] **Tal Linzen**[1,2]

Neuro-Symbolic Language Modeling with Automaton-augmented Retrieval

Uri Alon [1] **Frank F. Xu** [1] **Junxian He** [1] **Sudipta Sengupta** [2] **Dan Roth** [3] **Graham Neubig** [1]
[1]Language Technologies Institute, Carnegie Mellon University [2]Amazon AWS [3]AWS AI Labs
{ualon,fangzhex,junxianh,gneubig}@cs.cmu.edu {sudipta,drot}@amazon.com

What Makes Instruction Learning Hard?
An Investigation and a New Challenge in a Synthetic Environment

Matthew Finlayson **Kyle Richardson** **Ashish Sabharwal** **Peter Clark**
Allen Institute for AI, Seattle, WA
{matthewf,kyler,ashishs,peterc}@allenai.org

FLaNN

Switching Gears: Benchmarking with Sentence Processing

Not All Structures Are Processed Equally

- Subject VS object relative clause

SRC **The horse** [$_{RC}$ that kicked the wolf] went home.
ORC **The horse** [$_{RC}$ that the wolf kicked] went home.

Switching Gears: Benchmarking with Sentence Processing

Not All Structures Are Processed Equally

- Subject VS object relative clause

SRC **The horse** [$_{RC}$ that kicked the wolf] went home.
ORC **The horse** [$_{RC}$ that the wolf kicked] went home.

Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies

Tal Linzen[1,2] **Emmanuel Dupoux**[1]
LSCP[1] & IJN[2], CNRS,
EHESS and ENS, PSL Research University
{tal.linzen,
emmanuel.dupoux}@ens.fr

Yoav Goldberg
Computer Science Department
Bar Ilan University
yoav.goldberg@gmail.com

Switching Gears: Benchmarking with Sentence Processing

Not All Structures Are Processed Equally

- Subject VS object relative clause

  SRC **The horse** $[_{RC}$ that **t** kicked the wolf] went home.
  ORC **The horse** $[_{RC}$ that the wolf kicked **t**] went home.

- Attachment preferences

  1.a I shot an elephant in my pajamas
  1.b I **[shot an elephant]** **[in my pajamas]**

Switching Gears: Benchmarking with Sentence Processing

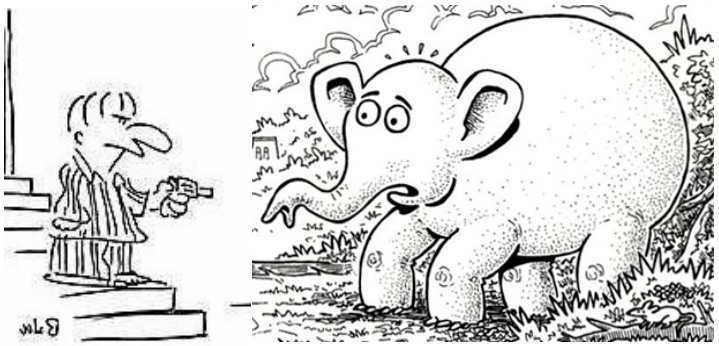
Not All Structures Are Processed Equally

- Subject VS object relative clause

  SRC **The horse** [$_{RC}$ that **t** kicked the wolf] went home.
  ORC **The horse** [$_{RC}$ that the wolf kicked **t**] went home.

- Attachment preferences

  1.a I shot **[an elephant in my pajamas]**
  1.b I **[shot an elephant]** **[in my pajamas]**



Ambiguity All the Way Down

So What?

Ambiguity is ubiquitous in natural language!

For Cognitive Science

- How do humans handle multiple structural representations?
- What principles guide ambiguity resolution cross-linguistically?
- Language specific properties vs. general biases/mechanisms?

For NLP

- How do LLMs handle multiple structural representations?
- What principles guide ambiguity resolution cross-linguistically?
- Language specific properties vs. general biases/mechanisms?

So What?

Ambiguity is ubiquitous in natural language!

For Cognitive Science

- How do humans handle multiple structural representations?
- What principles guide ambiguity resolution cross-linguistically?
- Language specific properties vs. general biases/mechanisms?

For NLP

- How do LLMs handle multiple structural representations?
- What principles guide ambiguity resolution cross-linguistically?
- Language specific properties vs. general biases/mechanisms?

Ambiguity and Relative Clauses Cross-linguistically [2]

- They saw the daughter of the actress that was on the balcony

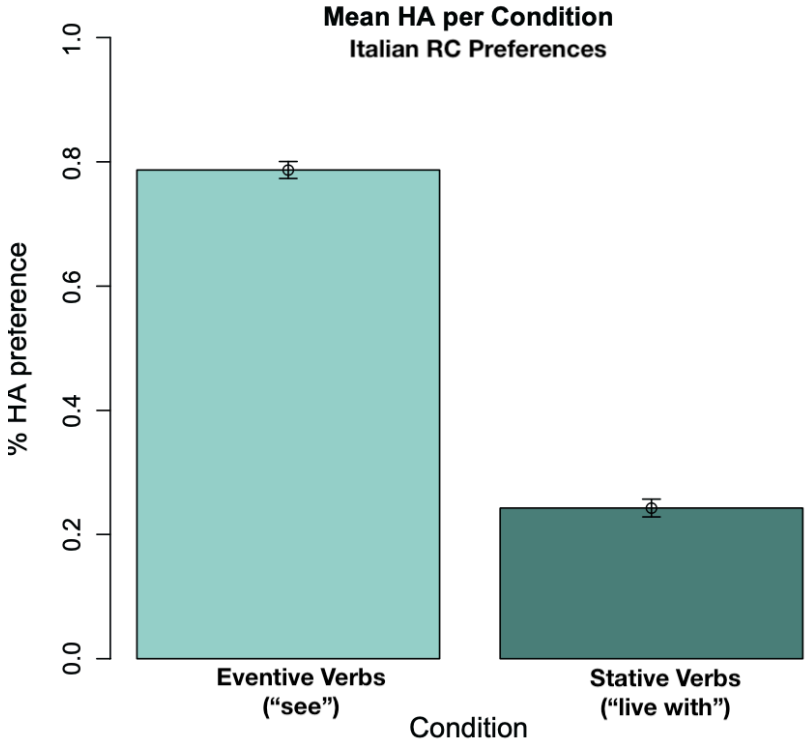
NP_1	**The daughter** was on the balcony	HA
NP_2	**The actress** was on the balcony	LA

[2] Grillo & Costa (2015,) De Santo & Lee (2023), Lee & De Santo (2024)

Ambiguity and Relative Clauses Cross-linguistically [2]

- They saw the daughter of the actress that was on the balcony

NP_1	**The daughter** was on the balcony	HA
NP_2	**The actress** was on the balcony	LA

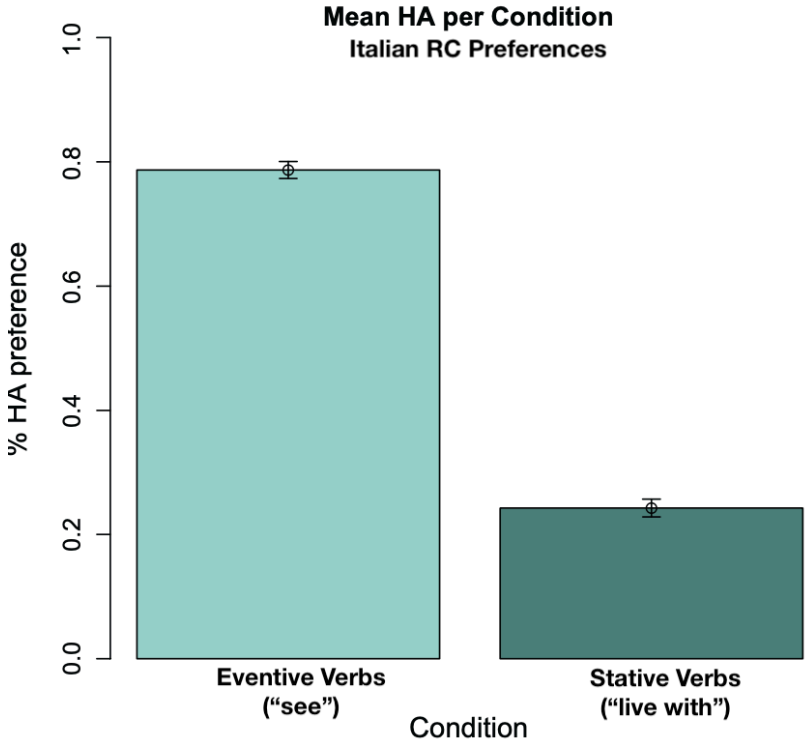


[2] Grillo & Costa (2015,) De Santo & Lee (2023), Lee & De Santo (2024)

Ambiguity and Relative Clauses Cross-linguistically [2]

- They saw the daughter of the actress that was on the balcony

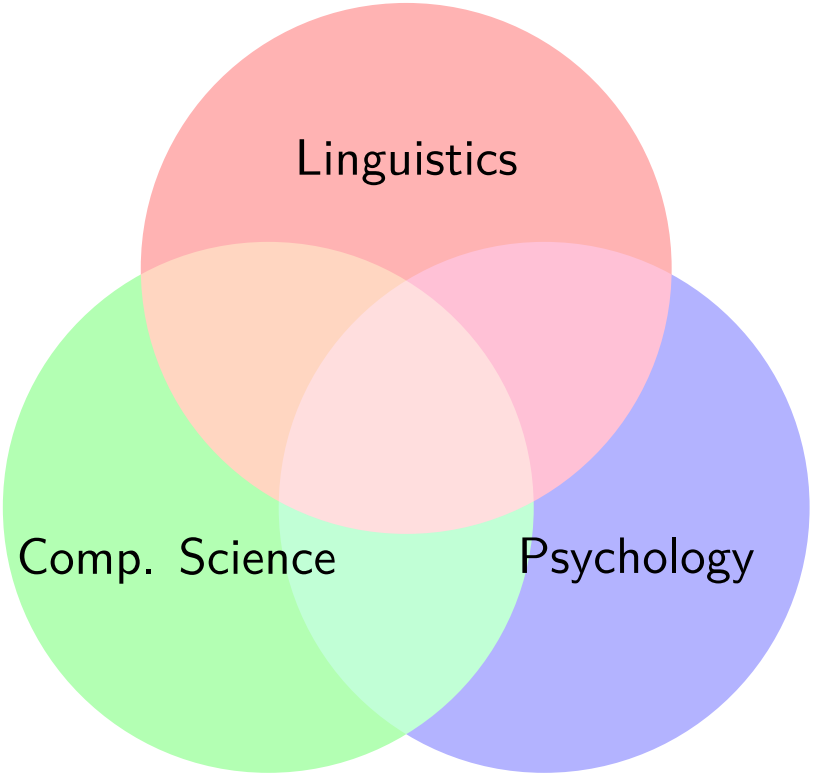
NP_1	**The daughter** was on the balcony	HA
NP_2	**The actress** was on the balcony	LA

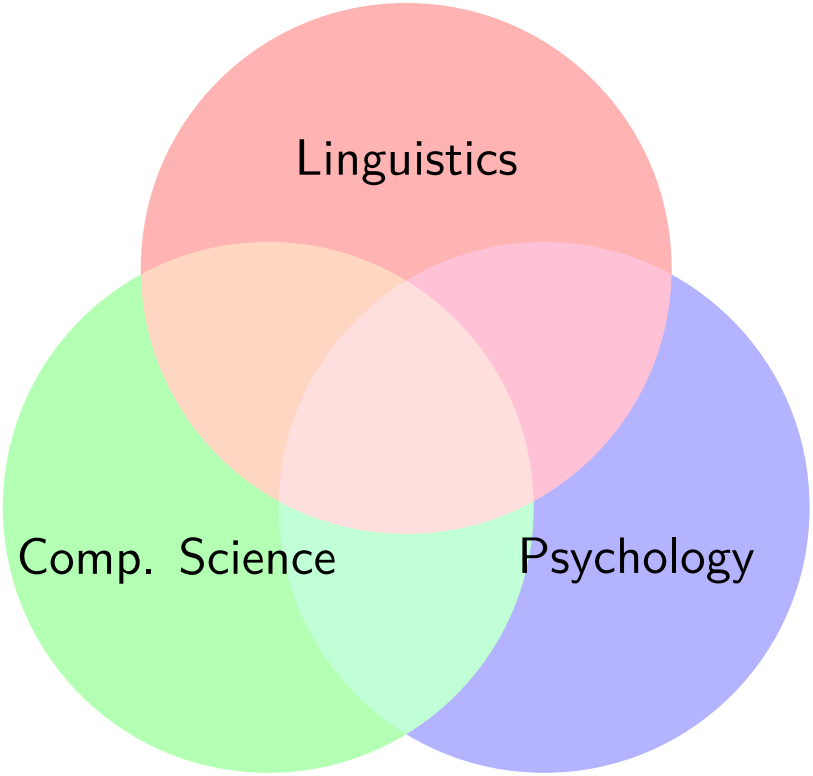


LLMs and Italian RCs?

Check out the poster!

[2] Grillo & Costa (2015,) De Santo & Lee (2023), Lee & De Santo (2024)


Linguistics
Comp. Science
Psychology



Let's chat!

Unbounded Dependencies Are Not SL

- **Samala Sibilant Harmony**
  Sibilants must not disagree in anteriority.
  (?)

  (2) a. * ha**s**xintilawaʃ
  b. * haʃxintilawa**s**
  c. haʃxintilawaʃ

Example: Samala

*$ h a **s** x i n t i l a w a ʃ $

$ h a ʃ x i n t i l a w a ʃ $

Unbounded Dependencies Are Not SL

- **Samala Sibilant Harmony**
  Sibilants must not disagree in anteriority.
  (?)

  (2) a. * ha**s**xintilawaʃ
  b. * haʃxintilawa**s**
  c. haʃxintilawaʃ

Example: Samala

*$ h a **s** x i n t i l a w a ʃ $

$ h a ʃ x i n t i l a w a ʃ $

Unbounded Dependencies Are Not SL

- **Samala Sibilant Harmony**
  Sibilants must not disagree in anteriority.
  (?)

  (2) a. * hasxintilawaʃ
  b. * haʃxintilawas
  c. haʃxintilawaʃ

Example: Samala

*$ h a s x i n t i l a w a ʃ $

$ h a ʃ x i n t i l a w a ʃ $

Unbounded Dependencies Are Not SL

- **Samala Sibilant Harmony**
  Sibilants must not disagree in anteriority.
  (?)

  (2) a. * hasxintilawaʃ
  b. * haʃxintilawas
  c. haʃxintilawaʃ

Example: Samala

*$ h a s x i n t i l a w a ʃ $

$ h a ʃ x i n t i l a w a ʃ $

Unbounded Dependencies Are Not SL

- **Samala Sibilant Harmony**
  Sibilants must not disagree in anteriority.
  (?)

  (2) a. * hasxintilawaʃ
  b. * haʃxintilawas
  c. haʃxintilawaʃ

Example: Samala

*$ h a s x i n t i l a w a ʃ $

$ h a ʃ x i n t i l a w a ʃ $

- **But:** Sibilants can be arbitrarily far away from each other!

*$ s t a j a n o w o n w a ʃ $

Unbounded Dependencies Are Not SL

- **Samala Sibilant Harmony**
  Sibilants must not disagree in anteriority.
  (?)

  (2) a. * hasxintilawaʃ
  b. * haʃxintilawas
  c. haʃxintilawaʃ

Example: Samala

*$ h a s x i n t i l a w a ʃ $

$ h a ʃ x i n t i l a w a ʃ $

- **But:** Sibilants can be arbitrarily far away from each other!

*$ s t a j a n o w o n w a ʃ $

Unbounded Dependencies are TSL

- Let's revisit Samala Sibilant Harmony

(3) a. * ha**s**xintilawaʃ

b. * haʃxintilawa**s**

c. haʃxintilawaʃ

- What do we need to project? [+strident]
- What do we need to ban? *[+ant][−ant],*[−ant][+ant]

I.E. *sʃ, *sʒ, *zʃ, *zʒ, *ʃs, *ʒs, *ʃz, *ʒz

Example: TSL Samala

s ʃ	ʃ ʃ
* $ h a s x i n t i l a w ʃ $	ok $ h a ʃ x i n t i l a w ʃ $

Unbounded Dependencies are TSL

- Let's revisit Samala Sibilant Harmony

(3) a. * ha**s**xintilawaʃ

b. * haʃxintilawa**s**

c. haʃxintilawaʃ

- What do we need to project? [+strident]
- What do we need to ban? *[+ant][−ant],*[−ant][+ant]

I.E. *sʃ, *sʒ, *zʃ, *zʒ, *ʃs, *ʒs, *ʃz, *ʒz

Example: TSL Samala