

# What kind of grammar do LSTMs learn?

Thesis Project Proposal - Year 3

---

Ludovica Pannitto

Thu, Mar 26th 2021

34th cycle PhD student - CIMeC - Center for Mind/Brain Sciences

# Do RNNs NLMs learn grammar?

A popular question, relating to **productivity** and **compositionality**<sup>1</sup>.  
Can machines master these fundamental traits of natural language?

How come such a simple architecture, fed with unrealistic input, with no access to perceptual information or hard-coded syntax can learn such a fundamental part of language?<sup>2</sup>

---

<sup>1</sup>“Linguistic generalization and compositionality in modern artificial neural networks” (Baroni 2020)

<sup>2</sup>“Colorless green recurrent networks dream hierarchically” (Gulordava et al. 2018),  
“The Emergence of Number and Syntax Units in LSTM Language Models” (Lakretz et al. 2019)

## Our Setup

---

How much language ( $L$ ) can be learnt from a certain level of computational complexity ( $C$ ) with a certain type of data ( $I$ )?

$$C \times I \xrightarrow{f} L \quad (1)$$

- we fix the level of computational complexity to a vanilla LSTM (character-based)
- we explore different sources of input in a specific range  $\{I_i\}$  selected based on their complexity level
- we want to explore the features of the produced language  $\ell \in L$

$$(\text{LSTM}, \{I_i\}) \xrightarrow{f} \ell \quad (2)$$

# Research Questions and Hypotheses

Our questions are the following:

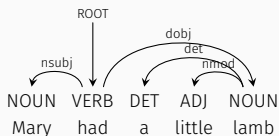
1. *How much grammar is learned overall by the system?*
2. *What is the influence of the ~~complexity~~ shape of the input on the learning process?*

If the network is able to *abstract* some grammatical knowledge from raw data, then:

- A **incrementality**: the learning process must be incremental and hierarchical
- B **categories**: the structures learned can be described through data-driven categories

## Definition of *Catena*:

“a word, or a combination of words which is continuous with respect to dominance”



- Mary had lamb
- had a lamb
- little lamb
- Mary had NOUN
- nsubj VERB dobj

**Figure 1:** Dependency representation for the sentence: *Mary had a little lamb*

The number and composition of *catenae* depends on **how elements are arranged** in the structure of the dependency tree.

1A (i) The **quantity** of learned structures grows with training, (ii) the **quality** of learned structures changes with training

$$(i) |G(\ell_{s_1})| \leq |G(\ell_{s_2})| \leq \dots \leq |G(\ell_{s_i})| \leq \dots \leq |G(\ell_{s_n})|$$

$$(ii) |G_L(\ell_{s_1})| \geq |G_L(\ell_{s_2})| \geq \dots \geq |G_L(\ell_{s_i})| \geq \dots \geq |G_L(\ell_{s_n})| \text{ and} \\ |G_C(\ell_{s_1})| \leq |G_C(\ell_{s_2})| \leq \dots \leq |G_C(\ell_{s_i})| \leq \dots \leq |G_C(\ell_{s_n})|$$

1B The **distributional properties** of structures at timestep  $t$  help explaining the distribution at timestep  $t + j$ .

$$\phi_i(x_l, x_c) \leq \phi_j(x_l, x_c)$$

2A Abstraction is faster if the input is given with progressive levels of complexities

$$G(\ell_i) \subseteq G(\ell_j) \iff c(\ell_i) \leq c(\ell_j)$$

# Recurrent Babbling

---



- vanilla **char-LSTM** trained on a limited amount of **child-motivated language** - **LSTMs can be seen as domain-general attention and memory mechanisms**, without any explicitly hard-coded grammatical knowledge.
- introduce a methodology to evaluate the **distribution of grammatical items**, focusing on the network's generated output - its *babbling*
- explore the **interaction** between meaning representations and the abstraction abilities of the network

The study is conducted on **English**.

# Child-motivated input

We've collected a portion of existing corpora, with specific attention at developmental language.

**CHILDES** - Child-directed utterances of the NA and UK portions of the CHILDES database.

**Gutenberg** - Books and newspapers from 18 children-related bookshelves of Project Gutenberg (incl. literature, instructional books and others).

**Opensubtitles** - Movie and TV series subtitles from the OpenSubtitle corpus, filtered on the content-rating label (G for movies and TV-Y, TV-Y7. TV-G for tv series).

**Simplewikipedia** - 2019 dump of Simple English Wikipedia, written in basic and learning English.

# Pipeline

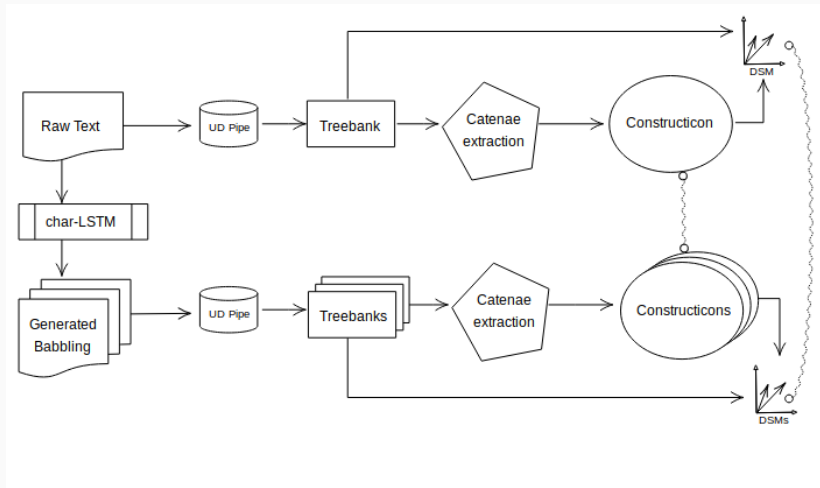


Figure 2: A summary of the work pipeline

# Catenaes extraction

catena	frequency	mi
<b>largest mi</b>		
@nsubj @root	294.59K	633.93K
_DET _NOUN	189.97K	552.32K
_VERB @obj	190.72K	520.82K
_PRON _VERB	271.44K	503.17K
@nsubj _AUX @root	129.60K	478.86K
<b>smallest mi</b>		
_PRON @nsubj	17.50K	-35.54K
@root @nsubj	27.61K	-34.89K
@nsubj _PRON	11.63K	-30.47K
_VERB @nsubj	12.79K	-26.82K
_AUX _PRON	15.75K	-26.67K

**Table 1:** Examples of catenaes extracted from CHILDES. Largest and smallest mutual information are reported, in top and bottom tier of the table respectively.

Part of Speech are prefixed by “\_” and syntactic relations are prefixed by “@”

**Q1:** To what extent is the network able to generate **new** language?

- We expect the network to reproduce the **statistical regularities** of the input, we further investigate what kind of regularities are acquired and how do the language models differ.

**Q2:** On what conditions is the network able to generalize its *grammatical* knowledge?

- We can state that the network has learned some grammar once it is able to use an acquired pattern in a **productive** and **creative** way.
- We expect this generalization ability to evolve during training and the **distributional properties** of patterns to be in relation with the grammatical abilities of the network at various stages of learning.

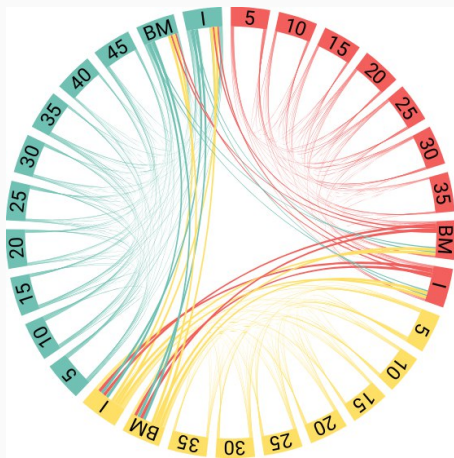
## Q1: What do ANNs approximate?

We evaluated **Spearman**  $\rho$  among the top 10K catenae extracted from the input and from each *babbling* stage produced by the LSTM.

Our analysis shows that the network has acquired statistical **regularities at the level of grammatical patterns**, and is able to use them productively to generate **novel** language fragments that **adhere to the same distribution as the input**.

Catenae extracted from babblings almost perfectly correlate with those extracted from the same input, but correlation values are quite **loose for out-of-domain pairs**.

# Q1: What do ANNs approximate?

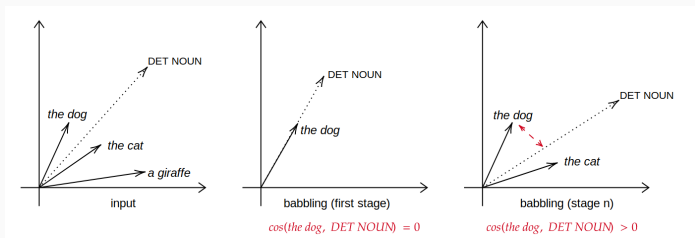


**Figure 3:** The thickness of the connections is **inversely** proportional to correlation. OpenSubtitles is shown in green on the left of the plot, CHILDES in red in the top right and Simple Wikipedia in yellow at the bottom.

## Q2: Meaning and abstraction

### The case of [SBJ V OBJ OBJ2]<sup>3</sup>

The meaning of the ditransitive pattern emerges from its strong association with **give** in child-directed speech: part of the meaning of *give* remains attached to the construction.



**Figure 4:** The network is supposed to capture stereotypical instances at early stages of learning and the productivity of the pattern will increase during training

<sup>3</sup>Constructions at work: The nature of generalization in language (Goldberg 2006)



## Q2: Meaning and abstraction

<i>cat</i> <sub>1</sub>	<i>cat</i> <sub>2</sub>	input	5	10	...	30	35	shift
a minute	a _NOUN	0.28	0.71	0.51	...	0.37	0.34	0.37
a minute	a @root	0.13	0.49	0.37	...	0.22	0.20	0.30
you _VERB it	_PRON @root @expl	0.10	0.46	0.28	...	0.17	0.21	0.25
you _VERB you	you _VERB @iobj	0.28	0.68	0.56	...	0.42	0.43	0.25
we can _VERB	_PRON can @root	0.51	0.79	0.74	...	0.61	0.57	0.22

**Table 2:** Pairs of catenae (*cat*<sub>1</sub>, *cat*<sub>2</sub>), their cosine similarity in the space obtained from CHILDES and in the spaces obtained from intermediate *babbling* stages.

The last column shows the difference between cosine similarity at epoch 5 and cosine similarity at epoch 35.

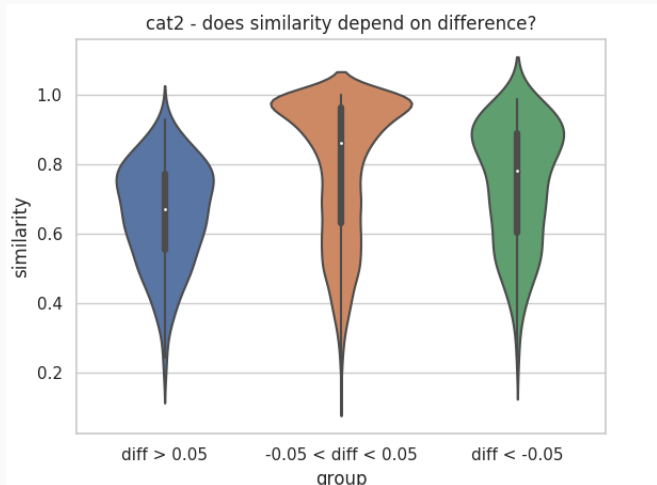
## Q2: Meaning and abstraction

Hypotheses:

- pairs with very **high input similarity** are unlikely to exhibit abstraction: the *catena* that is part of the *Construction* is the least abstract one, and there is **no need** for the more abstract category - i.e., non productive idioms like *talk through your hat* vs. *talk through your N*
- **low similarity** pairs, on the other hand, may simply contain **unrelated catenae** - i.e., too generic associations, like *the dog* vs *DET NOUN*

Instead, given pairs ( $cat_1, cat_2$ ) with  $cat_1$  being a less abstract instance of  $cat_2$ , we expect the highest shifts to happen at **intermediate levels of similarities in the input distributional space.**

## Q2: Meaning and abstraction



**Figure 5:** Distribution of average cosine similarities for the three groups of  $cat_2$ , showing low, intermediate and high average shifts respectively.

ANNs approximate the **distribution of constructions** at a quite refined level, even when trained over a bare 3M words from the CHILDES corpus.

We can follow paths of abstraction by putting our **grammar formalism** in a vector space.

- no sharp distinction between **lexicon** and **grammar** → different items can therefore be compared, irrespective of their lexical nature
- no assumption about the **stability** of the construction → what is relevant for productivity at the earliest stages of learning might become superfluous later on
- all items are **form-meaning** pairs → i.e., constructions
- **distributional semantics** is used both as a quantitative tool and as a usage-based cognitive hypothesis<sup>4</sup> → in line with the view of constructions as “*invitations to form categories*”<sup>5</sup>

---

<sup>4</sup>“Distributional semantics in linguistic and cognitive research” (Lenci 2008)

<sup>5</sup>*Explain me this: Creativity, competition, and the partial productivity of constructions* (Goldberg 2019)

Next steps

---

# Tentative thesis structure

## (i) Literature review:

- Neural Language Modeling in the usage-based framework
- Assumptions widely made in evaluating NLMs performances

## (ii) Acquiring constructions:

- Do NLMs acquire construction-like linguistic knowledge?
- What do NLMs approximate?
- Relation between meaning and the abstraction process

## (iii) Exploring the core:

- Are there constructions that are more core than others?

## (iv) Limits of what we can expect:

- What's the relation between the competences we found in NLMs and the shape and features of the input? Is there a "better" input?
- We - humans - learn some language almost whatever is the input we are exposed to. There are however differences in our competences and probably in the grammars we conceptualize. What do NLMs help us say about the boundaries of this variability?

# Tentative thesis structure

## (i) Literature review:

- Neural Language Modeling in the usage-based framework
- Assumptions widely made in evaluating NLMs performances

## (ii) Acquiring constructions:

- Do NLMs acquire construction-like linguistic knowledge?
- What do NLMs approximate?
- Relation between meaning and the abstraction process

## (iii) Exploring the core:

- Are there constructions that are more core than others?

## (iv) Limits of what we can expect:

- What's the relation between the competences we found in NLMs and the shape and features of the input? Is there a "better" input?
- We - humans - learn some language almost whatever is the input we are exposed to. There are however differences in our competences and probably in the grammars we conceptualize. What do NLMs help us say about the boundaries of this variability?

Do NLMs acquire  
construction-like linguistic  
knowledge?

---



Usage-based approaches rely on the idea that language is a **network** of relations among constructions.

**Constructionalization** (Cxzn) refers to the creation of new nodes (i.e., Cxns) in the network:

*the development through which certain structural patterns acquire their own meanings, so that they add meaning to the lexical elements occurring in them*<sup>6</sup>

In a framework for diachronic construction grammar, Traugott and Truesdale propose constructionalization as:

*establishment of a **new symbolic association** of form and meaning which has been replicated across a network of language users*<sup>7</sup>

---

<sup>6</sup>“Diachronic construction grammar and grammaticalization theory” (Noël 2007)

<sup>7</sup>based on *Constructionalization and constructional changes* (Traugott and Trousdale 2013)

Changes are not sudden and happen during a process

→ **Constructional Changes**

*Modulations of contextual uses prior to and following cxzn*<sup>8</sup>

Therefore we can identify:

1. Pre-cxzn changes
- ↓
2. Constructionalization
- ↓
3. Post-cxzn changes

---

<sup>8</sup>based on *Constructionalization and constructional changes* (Traugott and Trousdale 2013)

Can we show pre-cxzn and post-cxzn effects in NLMs' learning trajectories?

## Pre-cxzn

- loss of compositionality within a Cxn
- replication of semantic content or syntactic contexts that are connected with the emerging new Cxn, and increase in frequencies of these

## Post-cxzn

- Collocational expansion
- change in token productivity of the new Cxn
- loss of compositionality
- loss of analyzability within a Cxn
- incorporation into a more abstract, schematic type-node
- expansion of the schema

## loss of compositionality within a Cxn

Is the meaning of the (emerging) construction shifting from the meaning composed by summing its parts?

## replication of semantic/syntactic contexts

Is the distribution of context skewed towards some specific contexts before the emergence of the new construction?

## Collocational expansion

Is the construction being produced in new contexts?

## Change in token productivity

Is the frequency of constructs sanctioned by the construction increasing?

## Loss of analyzability<sup>9</sup>

The sub-constructions become less distinct and accessible, either because they get removed from the network or because the form of the new cxn changes

## Incorporation into a schema, Expansion of the schema

Shift to a new neighborhood of constructions

---

<sup>9</sup>Probably hard to see in our setting

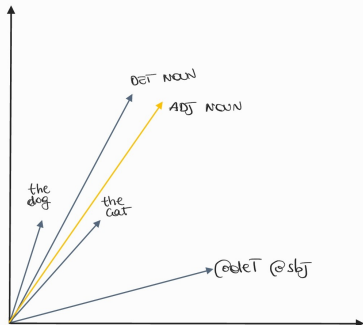
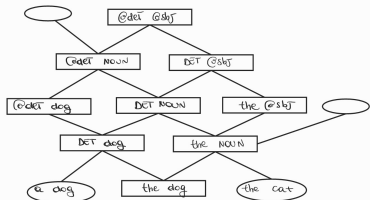
## Exploring the core (and the limits)

---

Within the network, constructions vary in terms of **schematicity**, **productivity** and **compositionality**.

Some constructions are usually regarded as more *core* than others, but it's often not clear what does this *core-ness* refer to.

- The core is what speakers agree more on
- The core is made up of highly schematic, highly productive and highly compositional cxns





*Schemas are abstractions across sets of Cxns which are (unconsciously) perceived by language users to be closely related to each other in the neural network*

The cxn  $c_i$  is a schema (to a greater extent) if its descending neighbors (less abstract neighbors  $N(c_i)$ ) in the network are also its distributional neighbors:

$$S(c_i) = \frac{1}{|N(c_i)|} \sum_{a_j \in N(c_i)} \cos(\vec{a}_j, \vec{c}_i) \quad (3)$$

*The degree to which a schema is “extensible”, i.e. the extent to which:*

- *a schema sanctions subschemas*
- *a schema is constrained*

The cxn  $c_i$  is productive depending on the number of (lexicalized) constructs ( $a_j \in L(c_i)$ ) it instantiates and their frequency ( $f(a_j)$ ):

$$P(c_i) = \sum_{a_j \in L(c_i)} f(a_j) \quad (4)$$

*The extent to which the link between meaning and form is transparent, and speakers can recognize the contribution that each component makes to the whole.*

The cxn  $c_i$  is compositional depending on the average distance of its vector ( $\vec{c}_i$ ) from the vectors obtained by composing its subparts

$P(c_i) = (p, q)$

$$C(c_i) = \frac{1}{|P(c_i)|} \sum_{(p,q) \in P(c_i)} \cos(\vec{c}_i, \overrightarrow{p+q}) \quad (5)$$













## The core is what speakers agree more on:

Do core cxns show similar distributional properties among speakers?

“Speakers” are NLMs trained on different inputs:

- Each of them has a different network of cxns
- Each of them has a different distributional space associated to cxns

We measure how cxns at different levels of schematicity, productivity and compositionality behave across speakers

	SPEAKER 1	SPEAKER 2	SPEAKER 3
Sentence 1			
Sentence 2			
Sentence 3			
Sentence 4			

# How to create the speakers?

*Core-ness* → speakers are created by randomly selecting subsets of text from the same collection

- Data: generic corpora composed of different genres (e.g., wikipedia, subtitles, books...).
- Sentences: variable presence of schematic, productive and compositional constructions

*Limits* → speakers are created trained by selecting subsets of text based on socioeconomic variables

- Data: ?
- Sentences: ?

## How to measure agreement?

For each sentence, for each speaker, we compute the set of constructions that the speaker uses to interpret the sentence

Let's consider the sentences where speaker  $S_1$  recognized construction  $c_i$ :

- what set of constructions  $d_1^i, \dots, d_k^i$  is elicited from other speakers  $S_2, \dots, S_n$ ?
- Are the distributional properties of the set  $c_i, d_1^i, \dots, d_k^i$  different depending on the degree of schematicity, productivity and compositionality of  $c_i$ ?

Thank you!



# References

---

- Baroni, Marco (2020). “Linguistic generalization and compositionality in modern artificial neural networks”. In: *Philosophical Transactions of the Royal Society B* 375.1791, p. 20190307.
- Goldberg, Adele E (2006). *Constructions at work: The nature of generalization in language*. Oxford University Press on Demand.
- (2019). *Explain me this: Creativity, competition, and the partial productivity of constructions*. Princeton University Press.
- Gulordava, Kristina et al. (2018). “Colorless green recurrent networks dream hierarchically”. In: *Proceedings of NAACL-HLT*, pp. 1195–1205.
- Lakretz, Yair et al. (2019). “The Emergence of Number and Syntax Units in LSTM Language Models”. In: *Proceedings of NAACL-HLT*, pp. 11–20.
- Lenci, Alessandro (2008). “Distributional semantics in linguistic and cognitive research”. In: *Italian journal of linguistics* 20.1, pp. 1–31.
- Noël, Dirk (2007). “Diachronic construction grammar and grammaticalization theory”. In: *Functions of language* 14.2, pp. 177–202.
- Traugott, Elizabeth Closs and Graeme Trousdale (2013). *Constructionalization and constructional changes*. Vol. 6. Oxford University Press.