CALaMo: a Constructionist Assessment of Language Models

Ludovica Pannitto
CIMeC
University of Trento
ludovica.pannitto@unitn.it

Aurélie Herbelot
CIMeC/DISI
University of Trento
aurelie.herbelot@unitn.it

Abstract

This paper presents a novel framework for evaluating Neural Language Models' linguistic abilities using a constructionist approach. Not only is the usage-based model in line with the underlying stochastic philosophy of neural architectures, but it also allows the linguist to keep meaning as a determinant factor in the analysis. We outline the framework and present two possible scenarios for its application.

1 Introduction

Over the years, linguists have given a lot of thought to what language *is*, and how it can be best formally described. Different approaches with sometimes contradictory aims have produced an extremely rich array of conceptual tools to describe linguistic phenomena. Such tools play diverse roles in explaining the phylogenetic, ontogenetic or historical-cultural facets of language and are often heavily interlaced with one another. In this research landscape, computational modelling has largely been used to simulate and investigate speaker behaviour at various levels of granularity. A specific area within the computational community is known as *(Neural) Language Modelling*, which aims at reproducing linguistic surface structure by means of (pseudo)-probabilistic models. Neural architectures have played a special role in this subfield of research, due to their flexibility.

The extreme complexity of theoretical tools found in linguistics gets cut down by order of magnitudes when it comes to the analysis of language processing using computational modelling. For instance, when the term *language* is mentioned in relation to Artificial Neural Networks, it seems that the word is often used as a mere synonym of *grammar*: while it is clear from a broader theoretical perspective that the two objects do not overlap, the distinction gets blurred in many computational studies. That is, assumptions which would be clearly stated in theoretical linguistics (e.g. how grammatical abstraction fits into the concept of *language*), are not explicitly discussed by computational studies: it is often the case that a specific set of choices concerning the description of language are taken as default. Most current work also seems to implicitly make a number of assumptions about what kind of grammar is supposed to emerge from neural language models (henceforth, NLMs), and this underlying choice is often echoed in the most common evaluation settings and in the conclusions that are being drawn from such experiments. Most of these default assumptions are inherited from the nativist Chomskian tradition and the Universal Grammar (UG) framework (Chomsky, 1986; Smith and Allott, 2016), which has pervaded a lot of the computational work on grammar, and continues to do so in the recent literature on neural models.

Ironically, the nativist assumptions that permeate the mainstream computational methodology are at odds with the very nature of the models created by the field. Neural models are essentially based on pattern learning and are completely agnostic about the nature of the data they are made to process. The idea that language can be abstracted from a general purpose statistical mechanism is more akin to usage-based (henceforth, UB) approaches (Barlow and Kemmer, 2000; Goldberg, 2003; Tomasello, 2003), and NLMs would provide a much more natural testbed for that theoretical strand. In the cognitive and UB accounts, the exploitation of predictability during language development (and again we refer to development at all the three tiers of philogeny, ontogeny and cultural evolution) is the root of a number of fundamental mechanisms such as schematization, entrenchment and distributional analysis (Lewkowicz et al., 2018). In the light of these processes, language, seen as a structured inventory of constructions, gets build through generations (Cornish et al., 2017) and throughout a speaker's lifetime: shared linguistic

Proceedings of the First International Workshop on Construction Grammars and NLP (CxGs+NLP, GURT/SyntaxFest 2023), pages 21 - 30

material among utterances, such as morphological markers for instance, enable the identification of particular patterns or constructions as units bearing meaning (Croft, 2001).

The perceived gap between the nativist and non-nativist traditions with respect to computational modelling probably stems from historical factors. The Chomskian school and its formal approach offered a definition of language that, in the past, could easily be interpreted and implemented by emergent computational approaches. But there is no reason for this bias to perdure. In this paper, we argue in favour of a usage-based framework to analyse language acquisition in ANNs. We first point out the aspects of nativist theories that have so far influenced the evaluation of NLMs (§2). We then introduce a framework for a quantitative and qualitative analysis of NLMs linguistic abilities within the constructionist perspective (§3). We finally show some preliminary analyses performed with the proposed formalization (§4).

2 Nativist vs. non-nativist approaches to language acquisition

All theories of language use and development recognize that at the root of human linguistic ability is the capacity to handle symbolic structures. But they disagree on the specific content of speakers' linguistic knowledge, the mode of acquisition of such content, and the extent to which linguistic productivity is affected by this stored knowledge (Bannard et al., 2009a). Theories diverge with respect to three aspects: input, stability and systematicity. The perspective taken on each of these aspects has consequences for the conclusions drawn from NLMs' responses to the evaluation setting. In the following, we consider each aspect in turn and specifically highlight how the evaluation of computational models becomes biased due to a lack of explicitness in relating experimental and theoretical aspects of the research question.

Input. One of the main arguments introduced by nativist frameworks is the *poverty of the stimulus*: the input children are exposed to is underdetermined and does not explain acquisitional generalizations observed in learners (Crain and Pietroski, 2001). Such theories assume that children navigate a hypotheses space defined by innate constraints (Eisenbeiß, 2009). Constructionist approaches, instead, posit that language emerges from the input through domain-general mechanisms: this implies that the input is shaped and skewed in a specific way in order to enhance learnability (Boyd and Goldberg, 2009). A well established line of research has shown how children are proficient statistical learners (Gómez and Gerken, 2000; Romberg and Saffran, 2010; Christiansen, 2019). The emergence of language-like structure from purely linear signal has also been shown in recent experiments such as (Cornish et al., 2017), which demonstrated how important aspects of the sequential structure of language may derive from adaptations to the cognitive limitations of human learners and users (Christiansen and Chater, 2016b). The crucial difference between the nativist and the non-nativist approach here is how strict the relation between the received input and the acquired linguistic structure is: if we commit to a view in which the input only serves as a trigger of an almost pre-determined cognitive structure, we are naturally driving our attention far from the features of the input and primarily to the features of the structure. On the other hand, deriving the linguistic structure from the input structure itself requires investigating the two aspects together. So far, most studies on NLMs have disregarded the effect of the input on experimental results (Pannitto and Herbelot, 2022).

Stability. The *continuity assumption* was first introduced by Pinker (1984) in order to reconcile aspects of developmental language with the generative framework. It posits that the differences between adult and children linguistic structures is negligible and merely due to performance factors. In contrast, what we can refer to as the *developmental hypothesis* claims that the mechanisms underlying acquisition remain the same throughout a life-long acquisition process, but the structures and abstractions they generate evolve over time. UB models also put emphasis on the linear and time-dependent nature of the linguistic signal (Christiansen and Chater, 2016b; Cornish et al., 2017). According to the UB account, generalizations appear gradually, as productivity emerges from item-specific knowledge (Bannard et al., 2009b).

Another aspect of stability is inter-speaker differences. UG posits that all speakers eventually converge to the same grammar (Lidz and Williams, 2009; Crain et al., 2009). Individual differences have however been found in almost every area of grammar, depending on a variety of factors including environmental ones (Street and Dąbrowska, 2010). The 'sameness' assumption pervades the

computational linguistics literature, where evaluation is performed according to a single 'gold standard' per task. For traditional tasks such as sentiment analysis or word similarity ratings, the annotations of human subjects are averaged, and the system is evaluated against the average. For language modeling, model perplexity is computed with respect to the statistical features of a large corpus, which aggregates the writing styles and linguistic habits of thousands of speakers. While this state of affairs has started to be criticised by various researchers, it remains for now the status quo. When considering language development as a speaker-dependent process, strongly affected by the nature of the input, an evaluation based on an 'average speaker' becomes truly unsatisfactory. We cannot assume the existence of a ground truth, and must rely on softer evaluation measures: it is clear that the linguistic behaviours of different speakers must overlap sufficiently to allow for communication, but that we also want to observe in the output of the network the kind of variability that is seen in humans.

Systematicity. The ability to understand and generate an unbounded number of novel sentences, using finite means, is considered one of the hallmarks of our language faculty. The boundaries of this systematicity remain however largely unclear: provided that we agree on what the finite means at our disposal are, not all the possibilities are actually realised by speakers and not all realised possibilities share the same cognitive or linguistic status.

One way to look at systematicity is that of compositionality, for which the most widely known version is probably due to Katz and Fodor (1963), that port Chomsky's innateness theory to semantics: a set of rules or constraints is needed in order to systematically build the meaning of sentences by integrating meaning of words. Even the Montagovian formal approach to compositionality (Montague, 1970) relies on Chomskian-derived ideas of a stable lexicon that stores meanings, and the existence of a set of precise interpretation rules that allow for those meanings to be mixed and modulated *through* the filter of syntax. The core of both visions is still very much syntax-centered (to which semantics has to be isomorphic) and very little space is left for indeterminacy, negotiation between speakers and other aspects related to the interactive and communicative nature of language (different individuals can retain in fact quite different concepts associated to the same lexical label for instance, Labov, 1973). In a nutshell, if we see systematicity from the standpoint of compositionality, the quasi-regularities of linguistic structure represents a major hurdle to surpass.

Quasi-regularity is instead the engine of productivity, as in the ability of speakers to use all the available linguistic means to cue the intended meaning. Just like compositionality, productivity deals with the domain in which a grammatical pattern can be employed in a linguistic context without losing interpretability, and it deals with what is actually possible in the language and where to draw the boundaries of acceptability. The shift has not just been syntactic: in the formal representation of these two aspects of systematicity in semantics, for instance, composition-oriented (Katz and Fodor, 1963) or productivity-oriented (Fillmore, 1976) theories have conceptualized the idea of selectional constraints differently.

Knowledge on systematicity is in both cases considered as implicit knowledge that the speaker has about their language. Nativist approaches have however primarily dealt with compositionality, and so are NLMs often evaluated: given grammar rules and lexicon, what are the computational mechanisms that allow them to combine? UB theories, on the other hand, have primarily been dealing with productivity: how far can meaning boundaries be forced? What are the mechanisms that allow for linguistic creativity? This of course entails, in the UB community, a relation to surface properties of the input as well: Croft and Alan Cruse (2004), for instance, note how the maximally schematic constructions, such as `sbj verb obj`, are also the most productive ones, and that this has a relation to their frequency too, both as a type and for each of their instantiations.

3 CALaMo

In our proposed methodology, CALaMo (Constructionist Assessment of Language Models), we incorporate the UB perspective across all three aspects: input, stability and systematicity.

As far as *input* is concerned, CALaMo differs from standard approaches by considering input data an important factor in determining the shape of the learner's grammatical knowledge. In traditional scenarios, the input only serves as a triggering factor and its features play little to no role in the analysis. From a UB perspective, instead, the relation

between the abstract grammatical structure of the input and the acquired grammar, which then constrains the production of the learner, is strict.

Regarding *stability*, depending on the view that is taken on the continuity hypothesis, we can see NLM's grammatical competence either as a binary or as a gradient property. In the first case, we test whether the network is able or not to handle some linguistic phenomenon, while in the second case, as advocated by CALaMo, we are interested in seeing how and why some linguistic aspect becomes more and more salient to the network during training.

The *compositionality vs. productivity* perspectives, finally, entail a different organization of linguistic knowledge: the mainstream compositionality perspective tends to set meaning aside, and treat the lexicon as an organized repository of meanings (it makes sense therefore to test NLM's capabilities on semantically nonsensical sentences or to extend the known rules to completely unknown lexical items). In the productivity perspective, instead, meaning is intrinsically part of the process and is treated as a systematic aspect of grammar, too.

3.1 Acquiring language

When talking about NLMs and their linguistic capabilities, the issue of language acquisition (A) is often formalized as how much language Λ can be learned by the (artificial) speaker, given a certain level of computational complexity C by being exposed to a certain type of data I:

$$A : C \times I \mapsto \Lambda \tag{1}$$

All the components of the equation have been central to the linguistic debate. However, starting from this basic formalization, we identify two major focus points that we specifically address in our framework. Firstly, the above formula describes acquisition as instantaneous, but it is actually better described as a process $A = (a_0, a_1, \cdots a_N)$ (§3.2). From a cognitive perspective the process is fully continuous, while in the artificial scenario, input data is often fed in 'batches'. We can however imagine that, if we had the ability to increase the number of steps at will (i.e., make N larger while keeping constant the amount of data), we could formalize steps small enough to make the two processes comparable.

Secondly, language is often seen as something that the learner has acquired and gained knowledge of. We want to bring back in the framework the role of the linguist-observer, that builds an abstraction over the linguistic behavior of the speaker (§3.3). As the actual knowledge acquired by the speaker is undetectable and only explainable metalinguistically, in a way that is not viable with neural networks (i.e., we cannot ask NLMs what they know about linguistic regularities), we must take into account the fact that we are always analyzing both the linguistic input received by the speaker and the output produced as an effect of the acquisition process through analytical categories that are created and used by the linguist-observer. In other words, Λ is not a property of the speaker, but rather a function operated by the linguist-observer. It does not evolve per se during the acquisition process, but rather it helps us detect and characterize the evolution of the speaker's abilities.

3.2 The process of acquisition

All the elements of Equation 1 ideally change throughout time as the acquisition process unfolds.

The input I to which the learner is exposed, in a real-life scenario, changes continuously. We can therefore define $I = (\iota_0, \iota_1, \cdots, \iota_N)$, where ι_i is the collection of input data to which the learner has been exposed to in-between a_i and a_{i+1}. Again ideally, with N large enough, each ι_i could even correspond to a single sentence. The computational complexity also co-evolves with the acquisition function, as linguistic knowledge gets incorporated into it. In the human case, the initial state is unobservable and in the artificial scenario it is often not interesting as initialization of neural models is random. At step i, instead, the computational mechanism that has incorporated knowledge up to step $i-1$ is exposed to ι_i. For these reasons, we define $C = (c_\emptyset, c_0, \cdots, c_{N-1})$. As an effect, Λ identifies different subsets $\lambda_0, \lambda_1, \cdots, \lambda_N$ throughout the acquisition process, namely $\Lambda = \bigcup_{i=0}^{N} \lambda_i$

Each step of the broader process A can be therefore defined as:

$$\begin{cases} a_0 : \iota_0 \times c_\emptyset \mapsto \lambda_0 \\ a_i : \iota_i \times c_{i-1} \mapsto \lambda_i \end{cases} \tag{2}$$

3.3 How do we observe *learned* language?

The notion of language that we introduced incorporates that of grammar, namely the analytical categories that we superimpose on the linguistic stream in order to analyze it and its unfolding over time. We do not test language as a cognitive state of the

speaker: we intend it instead a set of categories that the observer (i.e., the linguist) considers relevant to the description of the linguistic stream produced by the (artificial) speaker. There exists, therefore, a striking difference between the linguistic stream (either the input perceived or the output produced by the speaker) and its representation through the lens provided by *language*.

If we wanted to be more precise with the notation, we should acknowledge the fact that language, i.e. Λ, as we mean it is actually a function by itself, that takes as input some linguistic stream (some observable data) and returns a representation of it. We could therefore rewrite the definition of a_i as $a_i : \iota_i \times c_{i-1} \mapsto \Lambda(o_i)$ where o_i is the linguistic stream produced by the speaker as a result of acquisition step a_i.

As we are interested in the categories that are acquired by the speaker and deployed during language comprehension and production, defining $\lambda_{o_i} = \Lambda(o_i)$ allows us to apply the same transformation on the input ι_i to which the speaker is exposed, thus obtaining λ_{ι_i} that is comparable to λ_{o_i} in terms of linguistic categories.

Sticking to the constructionist perspective while trying to make the fewest possible assumptions on the actual content of linguistic knowledge, we hypothesize language as made up of a network of structures that are supposed to approximate constructions. As constructions are form-meaning pairs, the notion of grammar incorporates that of a meaning space spanning beyond the lexical level. This can be easily implemented by extending the notion of vector space models that has been extensively explored and used in distributional semantics (Lenci, 2008; Erk, 2012; Lenci, 2018). This represents a major difference with nativist approaches and the standard evaluation framework: meaning cannot be factored out of grammar effects and the acquisitional framework must account for its role in the process. If we had to formalize the content of any λ_i, therefore, we could expand it as $\lambda_i = \{(\kappa, \vec{\kappa}) \mid \kappa$ is a construction wrt. some linguistic stream$\}$

Unpacking this, we are saying that each obtained constructicon λ_i is a network of structures. These can be more or less lexicalized, with their abstractness being a proxy for linking the structures in the network as we will explain in the next paragraph. Each construction is associated with a distributional vector (Figure 1), which represent its meaning.

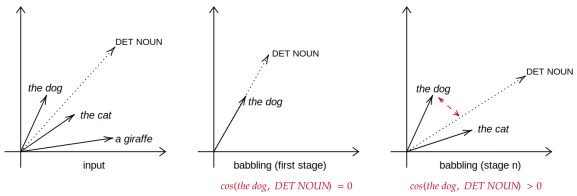

Figure 1: Let's assume that Λ contains both contructions *DET NOUN* and *the dog*, with the latter being a lexicalized instance of the former. At different steps during acquisition, the two constructions can assume different meanings and be therefore associated with different distributional vectors. A distributional vector condenses in fact information about co-occurrences between linguistic items in a given piece of text. In the figure, we see that a cluster of vectors gather around *DET NOUN* in the constructicon built from the input data (leftmost panel). This means that a variety of lexicalized instances exist for the construction *DET NOUN*. During learning, the constructicons built from generated output show different distributions for the construction *DET NOUN*. In the central panel, the cosine distance between *DET NOUN* and *the dog* is 0, meaning that their distributional contexts (i.e., their co-occurrences) perfectly overlap. In the rightmost panel instead, the distance between the two vectors has increased as another lexicalized instance (i.e., *the cat*) is being produced. In this scenario, the contexts where *DET NOUN* appears do not perfectly overlap with those where *the dog* appears.

3.4 Desiderata: the structure induced by Λ

We defined Λ as a function that takes as input a linguistic stream τ and returns a *constructicon* λ_τ: a structured repository of form-meaning pairs. In order to define and explore the internal structure of the constructicon, we introduce a few auxiliary functions and definitions:

(i): having meaning defined as a distributional space allows for distance computation $d(\kappa_i, \kappa_j)$ with $d : \Lambda \times \Lambda \mapsto [0, 1]$. $d(\cdot, \cdot)$ is a metric function that computes the distance between two meaning vectors. Usually, $d(\kappa_i, \kappa_j) = 1 - cos(\vec{\kappa_i}, \vec{\kappa_j})$, where $cos(\vec{\kappa_i}, \vec{\kappa_j})$ is the cosine similarity between the two vectors associated to κ_i and κ_j;

(ii): constructions bearing different abstraction levels are linked in the network. In order to navigate the network we introduce the function $c(\kappa_i, \kappa_j)$ with $c : \Lambda \times \Lambda \mapsto \{0, 1\}$ being a boolean function that computes whether two constructions constitute an *abstraction chain*. For instance, $\kappa_i =$ nsubj, GIVE, iobj, dobj and $\kappa_j =$ nsubj, root, iobj, dobj form a chain with κ_i being a

partially lexicalized (hence, less abstract) instance of κ_j.

3.5 Use scenarios

3.5.1 Individual acquisition over time

The framework can be used to observe how the acquisition process unfolds over time. We can in fact set a number of steps n and observe: (i) how the shape of grammar changes over the course of learning, comparing the various steps, as in: $\Lambda(o_1) \sim \Lambda_(o_2) \sim \cdots \sim \Lambda_(o_n)$, (ii) how the grammar of the input can be compared to that acquired by the speaker, as in: $\Lambda(I) \sim \Lambda(o_n)$. Given a subset $K \subseteq \Lambda(I)$[1] of interesting constructions, we can observe their behaviour over the learning process.

A popular constructionist hypothesis (Goldberg, 2006), for example, states that the meaning of a construction (e.g., the ditransitive pattern *Subj V Obj Obj2*), and therefore its productivity, emerges from the association with specific lexical items in the input received by the learner (e.g., *give* in the case of the ditransitive): part of the lexical meaning remains attached to the meaning of the syntactic pattern, and therefore its distributional properties with it. Let's assume that the speaker has acquired some construction κ (e.g., the ditransitive construction). Once they're able to use it in a productive and creative way (i.e., in a more varied contexts than the *give* contexts the construction is strongly associated with in the input), we can use the proposed framework to check whether the distributional meaning of two constructions $\kappa_i, \kappa_j \in \Lambda(I)$ with $c(\kappa_i, \kappa_j) = 1$ (i.e., with κ_i being a less abstract instance of κ_j) influences the learnability of κ_j as an independent construction.

The notion of *abstraction chain* introduced before helps us testing this hypothesis as we can check the behaviour of the chain (κ_i, κ_j) at each timestep. We can denote $\kappa_i^{\lambda_k}$ the construction $\kappa_i \in \lambda_k$ and similarly $\kappa_j^{\lambda_k}$ the construction $\kappa_j \in \lambda_k$, through distributional analysis we can capture how the contexts in which κ_i and κ_j vary, and whether this variation is associated with grammatical generalization. We expect, in fact, $d(\kappa_i, \kappa_j)$ to increase during acquisition:

$$d(\kappa_i^{\lambda_a}, \kappa_j^{\lambda_a}) \leq d(\kappa_i^{\lambda_b}, \kappa_j^{\lambda_b}) \;\; \forall\, a, b \mid a \leq b \quad (3)$$

If κ_j is produced in contexts that do not perfectly overlap with those where κ_i is produced, this indicates that the speaker has gained a productive use of construction κ_j, which is recognized as an independent construction from κ_i. If conversely their distance decreases during acquisition, we might deduce that the speaker has recognized κ_j as unnecessary by restricting its application cases to those of κ_i.

[1] Actually, we have to make sure that $K \subseteq \Lambda(I) \cap \lambda_0 \cap \lambda_1 \cap \cdots \cap \lambda_n$

3.5.2 Language as the expression of a population of speakers

We are often interested in defining grammar in terms of what can be considered shared linguistic knowledge among the speakers. A core aspect of construction grammar is in fact conceiving language primarily as a social and external phenomenon, as opposed to nativist theories that focus on its inner nature. By means of the framework, we can analyze grammar as an abstraction over the linguistic productions of a population of P speakers $\Pi = (\sigma_1, \sigma_2, \cdots, \sigma_P)$. We can define the grammatical conventions deployed by the community Π as $\Lambda_\Pi = (\lambda_{\sigma_1}, \lambda_{\sigma_2}, \cdots, \lambda_{\sigma_P})$. This allows for modelling variation between the acquisition process of the different speakers. Speaker σ_i might be exposed to a unique series of input material $\iota_0^{\sigma_i}, \cdots, \iota_N^{\sigma_i}$ that does not necessarily coincide with that of speaker σ_j.

In this setting, we can for instance investigate what is learned *no-matter-the-input*, and what is instead specific or idiosyncratic for each speaker. We can define:

$$G_{\geq p} = \left\{ \kappa \mid \sum_{i=0}^{P} X(\kappa, \sigma_i) \geq p \right\} \quad (4)$$

as the set of constructions that we can observe in the linguistic productions of p or more speakers. With:

$$X(\kappa_i, \sigma_j) = \begin{cases} 1 & \text{if } \kappa \in \Lambda^{\sigma_j} \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

being an auxiliary function that evaluates to 1 if the construction κ appears in the production of speaker σ_j and 0 otherwise (this just helps us count how many speakers use construction κ productively). G_P would for instance be the set of constructions shared by all speakers in a population, and could be therefore identified as the set of *core* constructions in Λ^Π. When, instead, $p \ll P$, we are observing constructions that are not shared by a significant amount of speakers in the population, and their use can therefore depend on specific input

instances or tendencies in subgroups of speakers. Following the same logic we can of course also just define $G_{(\sigma_i,\sigma_j)}$ as the constructions that are common to the two speakers σ_i and σ_j. By means of G, we can define $\widetilde{\Lambda_G}$ as an approximation of the function Λ, which only uses the categories that are retained in G. $\widetilde{\Lambda_{G_{\geq P}}}$ would for instance be a function that considers only linguistic knowledge shared by the entire population Π, while $\widetilde{\Lambda_{\sigma_i}}$ would be restricted to the constructicon λ_{σ_i}. Considering speakers σ_i and σ_j, with their respective produced linguistic outputs O_{σ_i} and O_{σ_j}, we can produce and compare $\widetilde{\Lambda_{G_{\sigma_i}}}(O_{\sigma_j})$ and $\widetilde{\Lambda_{G_{\sigma_j}}}(O_{\sigma_i})$: respectively, what speaker σ_i is able to retrieve from O_{σ_j} and what speaker σ_j is able to retrieve from O_{σ_i}.

The fact that speakers use the same constructions κ to build their linguistic productions does not of course ensure that the corresponding meanings $\vec{\kappa}$ coincide.[2] Different speakers, depending on the input they have been exposed to, and to the partial randomness attributed to computational mechanisms, could associate different meaning spaces to the same construction. Given two speakers σ_i and σ_j, and a sentence s, we could therefore compare the portions of λ_{σ_i} and λ_{σ_j} meaning spaces that are activated to linguistically (de)compose the sentence s.

4 Exploratory experiments

In order to explore the potential applications of the framework described in §3, we built a simple instance using the CHILDES corpus (MacWhinney, 2000) as input data I and a vanilla character-based LSTM (Hochreiter and Schmidhuber, 1997) as computational mechanisms C. With this simple setting, we explored two aspects: (i) we tested whether distributional similarities in λ_I would influence the acquisition of constructicons $\lambda_1, \cdots, \lambda_n$, and (ii) we tried to describe grammar as it emerges from a population of speakers. Constructions were approximated through *catenae* (Osborne et al., 2012): subtrees extracted from a dependency parsing syntactic representation (see Figure 2).

4.1 Abstracting grammar over training

We first replicate an analysis presented in Pannitto and Herbelot (2020), where a character-based LSTM was trained on CHILDES corpus. The authors fixed 7 steps during the LSTM's acquisition process, each after 5 epochs of training. In our formalization, this equates to 7 constructicons λ_1 to λ_7. The distributional space for each λ_i is obtained by counting co-occurrences between constructions within the same sentence. We can then consider abstraction chains (κ_i, κ_j) in I (i.e., in Λ(CHILDES) and computed $d(\kappa_i^{\lambda_7}, \kappa_j^{\lambda_7}) - d(\kappa_i^{\lambda_1}, \kappa_j^{\lambda_1})$ for each abstraction chain, namely the difference in cosine similarity between step 7 and step 1. Grouping all chains by κ_i and κ_j, it is possible to compute the average distributional shift as shown in Table 1 (i.e., for each κ_i to its more abstract instances and for each κ_j to its more concrete instances).

Three bins are considered, based on average distributional shift: the hypothesis is that constructions that underwent the highest shifts during training are those showing intermediate levels of similarities in the input distributional space. Indeed, chains with very high input similarities are unlikely to exhibit abstraction: according to constructionist intuition, their distributional similarity means that the construction that is part of the *Constructicon* is the least schematic one, and there is no need for the more schematic (and therefore, *abstract*) category to be created. Low similarity pairs, on the other hand, may simply contain unrelated constructions.

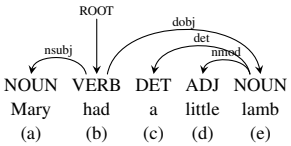

Figure 2: The dependency representation of the sentence *Mary had a little lamb*, annotated with morpho-syntactic and syntactic information. In this structure, we can identify the following *catenae*: a, b, c, d, e, ab, abce, abde, abcde, abe, bce, bde, be, ce, de, cde. Other possibilities would have been *strings* (e.g., a, ab, abc, ... b, bc, ...e) or *constituents* (i.e., a, abcde, c, d, cde).

κ_i	**shift**	**cosine**	κ_j	**shift**	**cosine**
@nsubj @root so	0.18	0.43	more @root	0.2	0.21
@nsubj only @root	0.18	0.41	_AUX know @obj	0.19	0.66
what @root @obj	0.18	0.39	@advmod tell	0.17	0.64
what @advmod _VERB	0.16	0.19	@aux know @obj	0.16	0.71
only @root	0.16	0.38	@advmod can _VERB	0.15	0.76
more @root	0.16	0.23	know @obj	0.15	0.62
@root it @xcomp	0.15	0.61	a _NOUN	0.13	0.52
@det minute	0.15	0.25	might @root	0.13	0.70
_PRON only @root	0.15	0.53	_PRON @root n't	0.12	0.53
_VERB _DET minute	0.15	0.33	@root that _VERB	0.12	0.65
_PRON @root so	0.14	0.54	_VERB 'll @ccomp	0.12	0.71
_DET minute	0.134	0.33	_VERB me @obl	0.12	0.76

Table 1: Constructions with highest average shifts.

[2] This makes sure that G_{σ_i} does not coincide with λ_{σ_i}

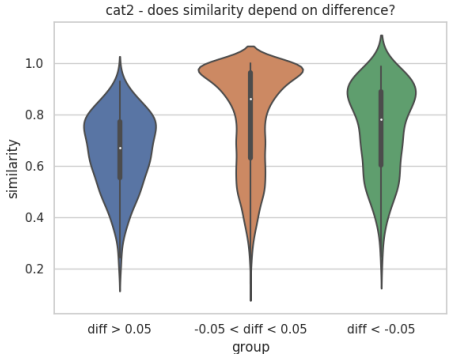

Figure 3: Distribution of average cosine similarities for the three groups of $kappa_j$, showing low, intermediate and high average shifts respectively.

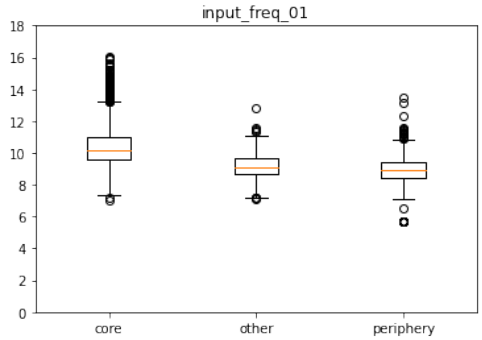

Figure 4: Difference in input frequency between the three groups of constructions: *core* as the ones shared by all speakers, *periphery* as the ones shared by half of the speakers or less, and *other* as the remaining ones.

The three groups show different distributions[3] as shown in Figure 3.

4.2 A population of artificial speakers

Following on Pannitto and Herbelot (2020) experiments, we consider a population of 10 speakers modeled with 10 vanilla character-based LSTMs trained on random samples of the CHILDES corpus (each containing 1 million words).

With this setting, we try to identify the locus of variation among different speakers, under the assumption that some 'core' constructions *must* be shared by all individuals, while others are less important to successful communication.

We restrict the analysis to the constructions to which all 10 speakers have been exposed to through their input (11051 constructions) and create G_{10} as the set of *core constructions* and $G_{\leq 5}$ as the set of *periphery constructions*, i.e. the ones shared by half of the speakers or less. Being trained on random samples taken from the same distribution, the speakers share most of the constructions (9086 out of 11051). However, we expect these numbers to change significantly when the input language varies along more refined sociolinguistic axes.

We also checked, for all speakers, whether the constructions of the *core* group and the constructions of the *periphery* group had significantly different frequencies in the input given to each speaker. As shown in Figure 4, the difference between the three groups are significant despite not appearing as striking as one would expect.

Lastly, we explored the input through $\widetilde{\Lambda_{G_{10}}}$ and $\widetilde{\Lambda_{G_{\leq 5}}}$, as shown in Table 2: both representations (the one obtained through *core* constructions and the one obtained through *periphery* constructions) highlight meaningful patterns in the sentence, but only the former can be considered a grammatical representation shared by the population.

corpus	Core	Periphery
does	AUX	-
n't	n't	-
that	that	PRON
seem	@root	VERB
kind	-	ADV
of	-	-
silly	ADV	ADV

Table 2: A sentence (left column) as it would appear if we restricted to only *core* (middle column) or *periphery* (right column) constructions.

5 Concluding remarks

The nature of linguistic representations is a core issue in linguistic theories of language development. We feel this aspect has been overlooked in the NLMs literature and propose an approach that brings back theoretical insights into the picture. We commit here to the UB constructionist framework, not as an ideal model of human language acquisition, but rather as a set of tools and categories that suffice to explain NLMs' generated language.

Since learning a language largely overlaps with learning to process the input, there must be a relation between processing biases relating to certain types of constructions and the distribution of those constructions in the linguistic input (Christiansen and Chater, 2016a). As experience grounds linguistic knowledge, distributional properties become a key aspect to determine the content of linguistic representations. In this framework, language is not considered as an autonomous cognitive system. Rather, the acquisition of grammar is regarded as any other conceptualization process and knowledge

[3] A Kruskall-Wallis one-way ANOVA was performed and resulted in significant values.

of language emerges from use.

To conclude, the observation of NLMs linguistic abilities would benefit from a constructionist approach. The evaluation can take place at multiple levels and includes properties of the situation described by the linguistic signal, but also properties of the linguistic signal itself. The UB framework may in fact provide useful categories to analyze the statistical properties of artificial language learners, and most importantly allows us to examine the semantic and the syntactic layers in parallel, both in the input received by the learner and in the stochastic output it generates.

References

Colin Bannard, Elena Lieven, and Michael Tomasello. 2009a. Modeling children's early grammatical knowledge. *Proceedings of the National Academy of Sciences*, 106(41):17284–17289.

Colin Bannard, Elena Lieven, and Michael Tomasello. 2009b. Modeling children's early grammatical knowledge. *Proceedings of the National Academy of Sciences of the United States of America*, 106(41):17284–17289.

Michael Barlow and Suzanne Kemmer. 2000. Usage-based models. *Language. Stanford*.

Jeremy K Boyd and Adele E Goldberg. 2009. Input effects within a constructionist framework. *Modern Language Journal*, 93(3):418–429.

Noam Chomsky. 1986. *Knowledge of language: Its nature, origin, and use*. Greenwood Publishing Group.

Morten H Christiansen. 2019. Implicit statistical learning: A tale of two literatures. *Topics in Cognitive Science*, 11(3):468–481.

Morten H Christiansen and Nick Chater. 2016a. *Creating language: Integrating evolution, acquisition, and processing*. MIT Press.

Morten H Christiansen and Nick Chater. 2016b. The Now-or-Never bottleneck: A fundamental constraint on language. *The Behavioral and brain sciences*, 39:e62.

Hannah Cornish, Rick Dale, Simon Kirby, and Morten H Christiansen. 2017. Sequence memory constraints give rise to language-like structure through iterated learning. *PloS one*, 12(1).

Stephen Crain and Paul Pietroski. 2001. Nature, nurture and universal grammar. *Linguistics and philosophy*, 24(2):139–186.

Stephen Crain, Rosalind Thornton, and Keiko Murasugi. 2009. Capturing the evasive passive. *Language acquisition*, 16(2):123–133.

William Croft. 2001. *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford University Press.

William Croft and D Alan Cruse. 2004. *Cognitive Linguistics*. Cambridge University Press.

Sonja Eisenbeiß. 2009. Generative approaches to language learning. 47(2):273–310.

Katrin Erk. 2012. Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, 6(10):635–653.

Fillmore. 1976. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences*.

Adele E Goldberg. 2003. Constructions: A new theoretical approach to language. *Trends in cognitive sciences*, 7(5):219–224.

Adele E Goldberg. 2006. *Constructions at work: The nature of generalization in language*. Oxford University Press on Demand.

Rebecca L Gómez and LouAnn Gerken. 2000. Infant artificial language learning and language acquisition. *Trends in cognitive sciences*, 4(5):178–186.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Jerrold J Katz and Jerry A Fodor. 1963. The structure of a semantic theory. *Language*, 39(2):170–210.

William Labov. 1973. The boundaries of words and their meanings. *New ways of analyzing variation in English*.

Alessandro Lenci. 2008. Distributional semantics in linguistic and cognitive research. *Italian journal of linguistics*, 20(1):1–31.

Alessandro Lenci. 2018. Distributional models of word meaning. *Annual review of Linguistics*, 4:151–171.

David J Lewkowicz, Mark A Schmuckler, and Diane M J Mangalindan. 2018. Learning of hierarchical serial patterns emerges in infancy. *Developmental psychobiology*, 60(3):243–255.

Jeffrey Lidz and Alexander Williams. 2009. Constructions on holiday. 20(1):177–189.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for analyzing talk. Third Edition.* Lawrence Erlbaum Associates.

Montague. 1970. English as a formal language. *Linguaggi nella Societa e nella Tecnica*.

Timothy Osborne, Michael Putnam, and Thomas Groß. 2012. Catenae: Introducing a novel unit of syntactic analysis. *Syntax*, 15(4):354–396.

Ludovica Pannitto and Aurélie Herbelot. 2020. Recurrent babbling: evaluating the acquisition of grammar from limited input data. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 165–176.

Ludovica Pannitto and Aurelie Herbelot. 2022. Can recurrent neural networks validate usage-based theories of grammar acquisition? *Frontiers in Psychology*, 13.

Steven Pinker. 1984. *Language learnability and language development*. Harvard University Press, Cambridge, MA.

Alexa R Romberg and Jenny R Saffran. 2010. Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(6):906–914.

Neil Smith and Nicholas Allott. 2016. *Chomsky: Ideas and ideals*. Cambridge University Press.

James A Street and Ewa Dąbrowska. 2010. More individual differences in language attainment: How much do adult native speakers of english know about passives and quantifiers? *Lingua. International review of general linguistics. Revue internationale de linguistique generale*, 120(8):2080–2094.

Michael Tomasello. 2003. *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press.