Written research report - Year 1

Ludovica Pannitto
CIMeC - Center for Mind/Brain Sciences
University of Trento

August 24, 2019

Theoretical landscape

Linguistic creativity, which we simplistically define as the ability to reuse existing, small linguistic bits to build up new, unseen blocks, is one of the most peculiar traits that distinguish human language from animal communicating systems, and, more strikingly, it has also been recognized as a skill that speakers acquire overtime (Bannard et al., 2009): the progress to linguistic productivity is in fact shown gradually by children, whose competence builds up on knowledge about specific items and on restricted abstractions before, if ever, getting to general categories and rules (Goldberg, 2006; Tomasello, 2003).

All theories of language development and use recognize that at the root of human linguistic ability is their capacity to handle symbolic structures: what theories do not agree on is the content of people's linguistic knowledge, on how this content is acquired and to what extent linguistic creativity is affected by this stored knowledge (Bannard et al., 2009). Even in recent formulations of the universal grammar (UG) framework (Hauser et al., 2002), the child's linguistic knowledge is described in terms of abstract rules and categories: many studies have questioned this assumption, showing how the empirical input to which children are exposed is enough to explain much of their linguistic development, provided that the child is equipped with the right tools to decode it.

Usage-based theories have argued against the two main tenets of generative models, namely the *poverty of the stimulus* (Chomksy, 1959; Chomsky, 1968) and the *continuity assumption* (Pinker, 1984), showing that language is probably a rich-enough signal for learners to pick up on, and also that children dispose of mechanisms of attention and memory that allow to explain and constrain many phenomena in language learning (Gómez and Gerken, 2000; Saffran et al., 2006; Romberg and Saffran, 2010)

One central aspect that distinguishes the usage-based approaches and the generative ones is the emphasis that former pose on the linear and time-dependent nature of the linguistic signal (Elman, 1990). While certainly not denying the utter relevance of hierarchical structures in language comprehension and production, they advocate that it emerges from the fact that language must be processed linearly and is subject to constrains posed by general-purpose memory and cognitive mechanisms (Christiansen and Chater, 2016). The existence and facilitatory role of higher-order structures in unquestioned and consistent with general observations about memory, such as the well known constraints on our ability to recall stimuli (Miller, 1956). The emergence of language-like structure from purely linear signal has for example been shown in recent experiments such as the one carried by Cornish et al. (2017), where the authors have demonstrated how important aspects of the sequential structure of language, as its char-

acteristic reusable parts, may derive from adaptations to the cognitive limitations of human learners and users.

One of the major issues that chunking models have to face is the existence of non-adjacent structures with very variable aspect on the surface. These kind of long-distance dependencies are common in language, involving verbal structures (e.g., the *progressive* and*perfective construction* in English, that involve a dependency between an auxiliar verb and the appropriate morphological mark), as well as higher order structures like the *correlative construction* (i.e., *the X-er, the Y-er*, as in *the more, the merrier*), or even more subtle things like agreement throughout the sentence or event-level dependencies: while it is intuitive that we, as speakers, are able to detect this kind of discontinuous patterns, evidence coming primarily from artificial grammar learning is not so strong about it (Gomez, 2002; Newport and Aslin, 2004; Gómez and Maye, 2005), being influenced by a great number of factors such as internal variability, the nature of the elements in the pattern.

The discovery and treatment of non-adjacent dependencies have therefore a central role in the theories that subserve language comprehension and production. Be they rules or actual chunks, and be they managed by a dedicated mechanism or a general statistical process (Peña et al., 2002), they embody the building blocks that bridge the traditional lexical level to the sentence and discourse level, being therefore central to the issues of linguistic creativity and compositionality.

Proposal

A number of questions emerge from the aforementioned literature: the emergence of non-adjacent dependencies still represents a puzzle both from a linguistic and computational point of view, and it appears to be strictly tied to two aspects that cannot be disentangled or detached from linguistic research, namely the **time-dependent** nature of the linguistic material, and the constrained posed on it by cognitive processing and human memory.

The question about *how do we attach meaning representations to linguistic symbols* has been central to usage-based models of language acquisition. In order to be better integrated with the statistical learning and cognitive-based community, we propose to pose the same question in a different formulation: *how do we identify the linguistic structures that are better suited, or more likely to cue the desired meaning?*

In other words, the problem of **segmentation**, which has been largely taken for granted by computational semanticists, should be more deeply investigated. At the same time, research on statistical learning and chunking has mainly focused on symbols, leaving aside issues concerning the function that the chunks have in the utterance.

Learning, also irrespective of the linguistic level, entails in fact two different aspects: **finding** (i.e., segmenting) the most relevant units to encode information and **representing** (i.e., compressing) information so as to make is efficient to store and to reproduce. The key issue is that these two processes should be mutually informative to one another and should be both considered when modeling or analyzing language.

The proposal is to provide a distributional model of non-adjacent dependencies (i.e., construction), as they emerge from the linear linguistic stream through general purpose statistical mechanisms.

Methods

Segmenting the signal: Spiking Neural Networks

Artificial Neural Networks (ANN), although having represented a sensible paradigm shift in many communities and having proven themselves as extremely powerful modelling tools, have also been accused of biological implausibility for a number of reasons, most commonly the fact that they involve non-local transfer of real-valued errors and

weights, while biological neuronal systems assume a kind of firing rate code for transmitting information throughout the brain.
Regularities are usually and most effectively extracted through overlapping representations, but as the Schapiro et al. (2017) model and complementary learning systems (McClelland et al., 1995; Schapiro et al., 2017) theory (CLS) have shown, non-overlapping representations are equally valuable tools for learning. In other words, while most neural network models seek generalization through the creation of prototypical items, but exemplars require modeling as well.
Spiking Neural Networks represent an emerging computational framework that could help overcome these drawbacks (Maass, 1997), moreover naturally incorporating the concept of time and therefore promising to be valuable candidates to model phenomena such as the linguistic ones, whose theorized hierarchical structures are highly constrained in a stream that develops over time.

Finding and representing the units: Distributional Construction Grammar

The idea of having different levels of abstraction with different levels of representation is directly reflected in linguistic items such as constructions, where fully instantiated elements coexist with partially filled structures. One of the areas where the co-existence of some sort of deterministic symbolic rules and subsymbolic mechanisms has emerged and has been widely explored is that of morphological structures (Bybee, 1995; Hay and Baayen, 2005), with frequency of exposure playing a key role in the organization and recognition of relevant morphological units and their combination (Bybee and McClelland, 2005). At higher levels than words various levels of idiomaticity and unpredictability have been recognized (e.g., multiword expressions and collocations), but they are still widely treated as special cases that depart from standard compositionality. From a computational perspective, even though the presence of subword and idiosyncratic units have proven to be effective in performance (Bojanowski et al., 2017; Ramisch and Villavicencio, 2018; Salle and Villavicencio, 2018), a more comprehensive and linguistically informed computational approach to the coexistence of different levels of segmentation is still missing.

The attempt to explain structural properties of language by means of distributional or linear patterns of co-occurrence has a long-standing history in linguistic research. Distributional semantics, that has now become one of the most influential frameworks for the representation and analysis of meaning in computational linguistics (Erk, 2012; Lenci, 2018), has one of its many roots in the structuralist distributional analysis such as the works of Harris (1954): a similar methodology is also at the core of the first attempts to identify the items and structures in children's language, such as *pivot grammar* (Braine, 1963). Linguistic distributional information, besides being a quantitative method for semantic analysis, could as well be regarded as a cognitive hypothesis about the form and origin of semantic representations (Miller and Charles, 1991; Lenci, 2008), an hypothesis that has been tested also in language acquisition studies (Twomey et al., 2014, 2016).

Conclusion

While distributional semantics could provide a solid framework for the representation of a wider spectrum of elements, such as the ones recognized as constructions, the process of identification of constructions itself could be informed by the acquired distributional knowledge, thus implementing the mutually informative cycle between newly processed and stored pieces of information. Moreover, the use of SNN in the processing phase allows for biological plausibility, while keeping crucial properties of the input, such as linearity, as central and motivating features of the model.

References

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

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

MDS Braine. 1963. The ontogeny of english phrase structure. *Language*, 39:1–13.

Joan Bybee. 1995. Regular morphology and the lexicon. *Language and cognitive processes*, 10(5):425–455.

Joan Bybee and James L McClelland. 2005. Alternatives to the combinatorial paradigm of linguistic theory based on domain general principles of human cognition. *The linguistic review*, 22(2-4):381–410.

Noam Chomksy. 1959. Review of skinner's verbal behaviour. *Language*, 35:26–58.

Noam Chomsky. 1968. *Language and Mind*. New York: Harcourt Brace Jovanovich.

Morten H Christiansen and Nick Chater. 2016. The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, 39.

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):e0168532.

Jeffrey L Elman. 1990. Finding structure in time. *Cognitive science*, 14(2):179–211.

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

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

Rebecca Gómez and Jessica Maye. 2005. The developmental trajectory of nonadjacent dependency learning. *Infancy*, 7(2):183–206.

Rebecca L Gomez. 2002. Variability and detection of invariant structure. *Psychological Science*, 13(5):431–436.

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

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Marc D Hauser, Noam Chomsky, and W Tecumseh Fitch. 2002. The faculty of language: what is it, who has it, and how did it evolve? *science*, 298(5598):1569–1579.

Jennifer B Hay and R Harald Baayen. 2005. Shifting paradigms: Gradient structure in morphology. *Trends in cognitive sciences*, 9(7):342–348.

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.

Wolfgang Maass. 1997. Networks of spiking neurons: the third generation of neural network models. *Neural networks*, 10(9):1659–1671.

James L McClelland, Bruce L McNaughton, and Randall C O'reilly. 1995. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist

models of learning and memory. *Psychological review*, 102(3):419.

George A Miller. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81.

George A Miller and Walter G Charles. 1991. Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1):1–28.

Elissa L Newport and Richard N Aslin. 2004. Learning at a distance i. statistical learning of non-adjacent dependencies. *Cognitive psychology*, 48(2):127–162.

Marcela Peña, Luca L Bonatti, Marina Nespor, and Jacques Mehler. 2002. Signal-driven computations in speech processing. *Science*, 298(5593):604–607.

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

Carlos Ramisch and Aline Villavicencio. 2018. Computational treatment of multiword expressions. In *The Oxford Handbook of Computational Linguistics 2nd edition*.

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

Jenny R Saffran, Janet F Werker, and Lynne A Werner. 2006. The infant's auditory world: Hearing, speech, and the beginnings of language. *Handbook of child psychology*.

Alexandre Salle and Aline Villavicencio. 2018. Incorporating subword information into matrix factorization word embeddings. In *Proceedings of the Second Workshop on Subword/Character LEvel Models*, pages 66–71.

Anna C Schapiro, Nicholas B Turk-Browne, Matthew M Botvinick, and Kenneth A Norman. 2017. Complementary learning systems within the hippocampus: a neural network modelling approach to reconciling episodic memory with statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1711):20160049.

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

Katherine E Twomey, Franklin Chang, and Ben Ambridge. 2014. Do as i say, not as i do: A lexical distributional account of english locative verb class acquisition. *Cognitive Psychology*, 73:41–71.

Katherine E Twomey, Franklin Chang, and Ben Ambridge. 2016. Lexical distributional cues, but not situational cues, are readily used to learn abstract locative verb-structure associations. *Cognition*, 153:124–139.