Computational models of language and processing

Final report for the Machine Learning for NLP course

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

July 13, 2019

People use language creatively. This ability in manipulating conceptual units, despite seeming a very superficial, maybe even naive and intuitive aspect of human linguistic ability, is actually at the core of many properties that natural language exhibits and should be taken as both the starting point and the guiding light of any theory aimed at explaining how natural language, broadly speaking, develops.

Creativity, which we simply define as the ability to reuse existing, small linguistic bits to build up new, unseen blocks, has been in fact mentioned as 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. The nature and origin of these tools will be discussed in the following part of this work. One crucial claim to which we stick throughout this analysis is that such tools need not to be tailored to linguistic competence, as the generative framework poses, on the contrary many of the studies that are being mentioned here regard it as a general-purpose learning mechanism.

1 Language brick by brick

The latter mentioned group of theories, that broadly fall under the category of **usage-based models**, 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). Not only has it been shown that language is a rich-enough signal for learners to pick up on, but also that children dispose of mechanisms of **attention** and **memory** that explain and constrain many phenomena in language learning.

Once the question of whether infants are able to track statistics in the input has been reasonably settled (Gómez and Gerken, 2000; Saffran et al., 2006), interest has shifted to a whole array of new issues concerning how children use the acquired **patterns** (Romberg and Saffran, 2010) and about the nature (Perruchet and Pacteau, 1990; Perruchet et al., 2002) and content (Estes et al., 2007; Yu and Ballard, 2007) of the representations, as well as the nature of the learning process itself.

1.1 Through the Processing Glass

From a more strict linguistic standpoint, this kind of approaches have contributed to blur the traditional, manichaeistic distinction between **lexicon and syntax** (Elman, 2009), the former being the repository of meaning and the latter being the grammatical device subserving the composition processes. A number of new architectures have been introduced in order to fill the gap left by the traditional dualistic model (MacDonald et al., 1994; Goldberg, 2003; Jackendoff, 2007; Christiansen and Chater, 2016).

This argumentation has been strongly supported also with neural evidence (Kuperberg, 2007), mainly coming from ERPs studies: as stated in Kuperberg (2007), N400 and P600, that had traditional been associated to semantic congruity (Kutas and Hillyard, 1980) and syntactic anomalies (Osterhout and Holcomb, 1992; Hagoort et al., 1993) respectively, have been observed also in response to a number of different phenomena, including and not limited to real world expectations (Hagoort et al., 2004), discourse level information (Berkum et al., 1999; Nieuwland and Van Berkum, 2006), organization of lexical items in semantic memory (Petten, 1993), and, strikingly, P600 was found also in pure semantic violations, with no sign of the N400 effect, all suggesting that, even in a model with two separate streams devoted to semantic memory and combinatorial properties respectively, different types of relationships among linguistic stored items influence each other, in a continuum that points to a much more cognitively plausible organizing schema.

In a nutshell, **processing**, rather than *abstract linguistic* competence, with its physical and cognitive underpinnings, has gained centrality in linguistic research.

1.1.1 ...and what linguists found there

This comes with a number of consequences, such as the fact that traditional categories like *grammaticality* has to be approached through the lens of cognitive salience. As it has been shown, it is the notion of **typicality** the one that turns out to be cognitively more salient, which is much affected by properties like frequency of co-occurrence or transitional probabilities, also giving relevance to the role of prediction in language comprehension and production (McRae and Matsuki, 2009; Misyak et al., 2010; Lupyan and Clark, 2015). Although questioned (Huettig, 2015; Huettig and Mani, 2016), **prediction** has undoubtedly played a leading role in linguistic research, both at a theoretical (Altmann and Mirković, 2009) and computational (Elman, 1990; Mikolov et al., 2013) level. Altmann and Mirković (2009), as Elman (1990) does, distinguish between prediction as a learning task and prediction, or rather *anticipation*, as an ability acquired as a by-product of the task: despite the fact that the task in Elman's model was modeled on the contingency between the input at time t and the input at time $t+1$, higher-order contingencies (e.g., between words) have emerged as well, resulting in a more varied and multi-level (i.e., hierarchical) representation.

This leads to mention the other aspect that distinguishes the usage-based approached and the generative ones, namely the emphasis that usage-based models pose on the linear and **time-dependent** nature of the linguistic signal. 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. 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).

On the other hand, the emergence of language-like structure from purely linear signal has 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 characteristic reusable parts, may derive from adaptations to the cognitive limitations of human learners and users. In a letter-string recall task, participants were asked to reproduce a series of 15 string that they had been previously been trained on. The recalled strings were used as inputs for the next participants, in a series of 10 subjects for chain. The authors report that, across generations, not only does learnability increase (i.e., the overall accuracy of the recalled items in terms of normalized edit distance increases, and not at the cost of a collapse of the string sets into very short sequences), but the amount of reuse of chunks also significantly differs from what one would expect from random strings, and structure similar to natural language generally emerge. In other to determine the increase of distributional structure, Cornish et al. adopt a metric which is frequently used in artificial grammar learning studies: *Associative Chunk Strength* (ACS) (Knowlton and Squire, 1994): for a given test sequence consisting of x bigrams, and $x-1$ trigrams, ACS is calculated as the relative frequency with which those chunks occur in the training items. For example, ACS for the recalled item ZVX in generation t is calculated as the sum of the frequencies of the fragments ZV, VX and ZVX in generation $t - 1$ divided by 3. By means of averaging, the authors find that the next generation tends to reuse these chunks successfully, and more so as generations proceed, thus incrementally developing re-usable units.

1.1.2 Meanwhile, in distribution-land

The attempt to explain structural properties of language by means of **distributional patterns of co-occurrence** has indeed 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).

What Miller and Charles claims, and Lenci underlines is, in fact, that:

> Knowing how to use words is a basic component of knowing a language, and how that component is acquired is a central question for linguists and cognitive psychologists alike. The search for an answer can begin with the cogent assumption that people learn how to use words by observing how words are used. And because words are used together in phrases and sentences, this starting assumption directs attention immediately to the importance of context.

This idea that knowledge is primarily *knowledge of use* is widespread in many other theories, and in line with the claim by Tomasello, who frames linguistic competence in terms of

> the mastery of a structured inventory of meaningful linguistic constructions

Although neither *pivot grammar* nor the subsequent attempts have proven to be able to explain children's language acquisition, some of the fundamental principles permeated into subsequent literature, in particular the idea that the child is able to make his or her own abstractions from exposure to adult behaviors. What's interesting is that these patterns of learning have survived especially in the non language-related research about children's cognitive development.

In particular, **statistical learning** (SL), which had initially focused on word learning (Reber, 1967; Saffran et al., 1996), has extended to treating the processing of regularities in sensory input in general, therefore offering a more comprehensive theory of information processing (Armstrong et al., 2017). As

Armstrong et al. point out, the outcome of statistical learning is not constrained to a representation of the statistics in the input, but rather it claims that experiencers possess the cognitive abilities to take track of distributional patterns, and that this contributed to shaping expectations and behavioral responses, much alike to what distributional semantic theories claim.

2 Process global, utter local

As **use** is the main aspect that has to be factored in, from our standpoint this cannot be accomplished without an overall model of *information processing*. A series of consequences depend on bringing processing back at the core of distributional linguistic theory: it undoubtedly affects decisions concerning the nature of linguistic conceptualization and abstractions and, most importantly, the processing mechanisms cannot be possibly faced from a pure linguistic standpoint, as it has been shown how general and unsupervised learning mechanisms are also active in language.

Some important remarks need to be mentioned about this: the claim for a domain-general ability such as statistical learning (or *implicit statistical learning*, as Christiansen (2018) frames it) does not rule out the possibility of modality- and stimulus- specific constraints Frost et al. (2015). As a matter of fact, the evicence for cross-modality transfer learning is scarce (Redington and Chater, 1996; Gomez et al., 2000), while the evidence persistently shows patterns of modality specificity or even stimulus specificity (Johansson, 2009) learning. Although this might seem to undermine the claim of SL as a unitary learning system, as Frost et al. note providing an operational definition of domain generality, the two aspect are better explained when seen together: while varying with respect to different sets of modality-specific constraints, the computational principles by which the learning happens are shared, and, consistently with neurobiological findings, some portion of the neuronal system are tied to a given modality while others form a multi-domain cognitive system that modulates or operates on inputs provided in form of modality-specific representations.

With respect to the general **underlying computational mechanisms**, Thiessen (2017) distinguishes between two distinct groups, aimed at detecting *conditional* and *distributional* regularities respectively. Conditional regularities are used to inform a chunk-based memory processes that ultimately stores **exemplars**, while distributional regularities, that refer to frequency and variability of the found exemplars in the input, are employed to capture **central tendencies** and group elements into categories. Thiessen (2017) argues how taking memory-based perspective on the topic is helpful in connecting both sides of statistical learning in the same framework, as it allows to focus on processes that are endemic to memory, such as activation, decay, interference and prototype formation. This kind of approach is also able to account for developmental changes in learning outcomes: as memory decays when getting older, the way the input is represented and the nature of this representation is likely to change with age. Earlier stages of learning could be slower but more flexible to adapt to new environments, while later stages gain stability and efficiency but are also more constrained by regularities of the environment (Thiessen and Saffran, 2003).

Thiessen et al. (2013), much like Christiansen and Chater (2016), highlight the necessity for a mechanism that is able to take into account for two distinct tendencies, namely an **extraction** process that supports chunking and an **integration** process that supports the aggregation of exemplars in coherent clusters, mediated by an attention mechanism.

One of the major consequences of this approach is dealing with the relation between item-based and categorical knowledge: one of the basic tenets of statistical learning and usage-based models more broadly is the fact that individual episodes can be integrated into more abstract representations that echo the statistical distributional properties of the input. This process of incremental abstraction implies a certain amount of loss of detail about the individual episodes. Altmann (2017) offers an operationalized definition of the **abstraction** process as

> accumulation of experience across trials which leads to generalizable knowledge (the 'correct' object-label mappings) not available to the organism at the start of this accumulation, and in which episode-specific details (including non-systematic, accidental, co-occurrences, including the 'incorrect' object-label pairings as well as the object-object and label-label pairings) become less salient than more systematic details, reflecting structure (or regularities) across episodes.

2.1 The word, violà l'ennemi

Altmann also offers a review of computational models of memory that embody principles relevant to the exemplar/schema distinction (Lund and Burgess, 1996; Elman, 1990; Jones and Mewhort, 2007; Landauer and Dumais, 1997; Altmann and Mirković, 2009), and they further argue that such models, including Elman's SRN, lack of a specific tool for modeling the relationship between episodic and semantic memory, which is instead a feature of neurobiologically-inspired models that rely on **complementary learning systems** (McClelland et al., 1995; Schapiro et al., 2017) (CLS): different structures, namely **hippocampal structures** and **neocortex**, support rapid encoding of different instances and slow recognition of regularities respectively. The difference between the two different kinds of semantic memory is however not sharp, and neuronal structures themselves have shown support for a *gradient of abstraction.*

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. Similarly, as more modern models, especially in the artificial neural network trend, tend to be less and less supervised and agnostic about the linguistic items they are modeling (Kalchbrenner et al., 2014), some insights have been provided about the amount of structural knowledge (i.e., knowledge about the underlying grammatical structure of the text used to train the model, rather than its content in terms of words) injected into the produced semantic representations (Shi et al., 2016; Belinkov et al., 2017a,b; Blevins et al., 2018). In this landscape, the success of models like BERT (Devlin et al., 2018), for example, paves the way for a new generation of distributional semantic models and sheds light on a different formulation of the problem of **learning form-meaning mappings**.

Learning, also irrespective of the linguistic level, entails in fact two different aspects:

- **finding** (i.e., segmenting) the most relevant units to encode information
- **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.

3 Computational models for Statistical Learning

3.1 PARSER

PARSER (Perruchet and Vinter, 1998) is among the first attempts to give a computational account of word segmentation. The aim of the authors is to show how parsing emerges as a natural consequence of the on-line attentional processing of the input, thanks to basic laws of memory and associative learning. They argue that chunking is an ubiquitous phenomenon, which people use to decompose a complex sequence of elements into simpler processing primitives: their aim is therefore to explain how chunks turn out to match linguistic items such as words. Perruchet et al. hypothesis is that the parsing emerges as a consequence of the organization of the cognitive system, which is characterized by the interplay of two principles: the fact that *perception shapes internal representations*, meaning that primitives that are perceived within one attentional focus as a consequence of their proximity become the constituents of one new representational unit, and the fact that *internal representations guide perception*, meaning that perception involves an active coding of the incoming information and this coding schema is constrained by the perceiver's already acquired knowledge. The frequency of repetition of certain chunks is then what reinforces the representation, and leads to lasting chunks that match words or subwords rather than between-words segments.

PARSER is initialized with an alphabet (i.e., a mental lexicon) of primitives (i.e., syllables), each assigned with a weight that gets updated in order to reflect the person's familiarity with the item. At each timestep, also processes of forgetting and retroactive interference take place.

The results obtained match well human performances on artificial grammars, and the authors claim that this is due to principles general enough to scale up to natural language experiments. However, the application of such a mechanisms to naturalistic data has proven to be non-trivial.

The relevance of the PARSER model for the subsequent works in the field of statistical learning is out of question. The assumptions made (i.e., taking syllables as primitives and the idea that material is perceived as a succession of small and disjunctive chunks composed of a few primitives) is debatable: it seems reasonable to suppose that such primitives, provided they exist and people rely on them for language processing, lie at a much higher and more abstract level than linguistic, symbolic material. Moreover, although forgetting is implemented in the model, the representation of the chunks remains discrete throughout the whole process, without any form of abstraction or compression of the encountered information.

3.2 CAPPUCCINO

The aim of the CAPPUCCINO (i.e., Comprehension and Production Performed Using Chunks Computed Incrementally, Non-categorically and On-line) model of language acquisition, introduced in McCauley and Christiansen (2011), is to provide a test of the assumption that children's language use involves explicitly stored chunks. It has a number of features that mirror key psychological properties, such as *incremental learning* from *naturalistic input*, *on-line processing*, simple *statistical measures* that we know children are able to track, and *comprehension* and *production* are taken into account at once.

In practice, the models builds an inventory of chunks (i.e., a *chunkatory*) used to segment phrases, and that are afterwards used to reproduce children's utterances. The model relies on backward transitional probabilities (BTPs) to discover chunks: high BTPs are taken as a cue that words belong to the same phrase, while low BTPs mark phrase boundaries. Utterances are processed on a word-by-word basis and, when BTP is higher than the running average, the current word-pair is grouped together as a part of a chunk; whereas when the BTP falls below the running average, a chunk is created and added to the chunkatory. The key aspect is that items of the chunkatory are in turn used to assist processing on the same word-to-word basis.

The major standpoint of McCauley and Christiansen is however that, given better performances of the model to fit artificial grammar learning data when exposed to words as opposed to lexical categories, knowledge of concrete words and chunks may be more important to early language acquisition than abstract rules operating over word classes.

While this methodology has proven itself to be viable, it heavily relies on previous segmentation of the text in words, thus making it impossible to account for a greater number of structures such as subword chunks.

3.3 iMinerva

iMinerva, short for *Integrative Minerva*, was introduced in Thiessen and Pavlik Jr (2013) with the aim of understanding whether domain-general principles (i.e., *activation* of similar memories, *decay*, *integration*, and *abstraction*) can account for a variety of linguistically relevant learning tasks: the authors test their model against three fairly different tasks, namely category learning reproducing the patterns found in Maye et al. (2002), simulating the effects of distribution of exemplars in children's use of categorial distrinctions (Thiessen, 2007) in word learning, and the effect of variability in non-adjacent dependency learning (Gomez, 2002).

The dissimilarity of the three tasks, along with the fact that iMinerva employs general-purpose memory mechanisms, suggest the possibility that the underlying human processes could be explainable through the basic operations that iMinerva performs, in particular comparing between current and prior exemplars, and integrating them into a representation that is sensitive to the central tendencies of prior experience.

One limitation, highlighted by the authors themselves, concerns the fact that iMinerva does not attempt to model production, nor infants' behavioral responses. Instead, iMinerva is focused on identifying the representations underlying performance in learning tasks and the processes that lead to the formation of those representations. Related to this issue is the fact that iMinerva does not chunk the input either, but relies on the provided segmentation, tailored to the task. In other words, it is just a model of distributional statistical learning, while it sets aside the issue of conditional statistical learning. The two processes, as Thiessen et al. (2013) argues, cannot be truly disentangled and are mutually informative.

3.4 R-Grams

A quite different approach to the issue of chunking the input in a set of relevant linguistic units is introduced in Ekgren et al. (2018). Although their aim is probably very far from that of the previous studies, we find it nonetheless interesting to mention here, as it shares some of the main features of the previous studies (i.e., being completely usage-based, input-driven and unsupervised), while showing some of the traits of the computational linguistic tradition, as for example a greater scalability to larger amounts of data.

Another interesting feature of their model is that it is based on the *Re-Pair* algorithm (Moffat and Larsson, 2000), a compression algorithm in the family of dictionary-based compression. The idea that the processes of extraction of more and more abstract chunks or schemas from the input must involve a notion of compression is quite widespread in the statistical learning literature and highlighted also in Christiansen and Chater (2016).

The chunk identification procedure is pretty straightforward, and involves just a few steps, namely, given an initial alphabet of symbols, i. find the pair *ab* that occurs most frequently in text, ii.) replace all occurrences of *ab* with a new symbol A, iii.) add the rule $A \to ab$ in the grammar, iv.) repeat until no pair occurs more than a defined threshold or the vocabulary size exceeds memory limits.

The implementation, as is, has a number of drawbacks, such as the fact that throughout the whole process the entire text must be maintained available, and the inability to account for non-adjacent chunks. Furthermore, the model presents a mixture of grammar rules induction and fragments storing: it remains therefore unclear how a subsequent parsing phase would be performed.

3.5 The issue of non-adjacent chunks

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. It is the case of, for example, the *progressive* (e.g., *is running*) and *perfective construction* (e.g., *has stopped*) in English, that is shaped as set of discontinuous chunks consisting of a form of the verb *to be* or *to have* respectively, an empty slot with some high level constraint at the level of linguistic category (i.e., it must be filled by a *verb*) and the appropriate morphological mark (e.g., the *-ing* ending for the progressive construction). The same holds for 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.

A seminal work on this issue (Gomez, 2002) underlined the striking role played by **internal variability**: while discrimination was poor in low-variability conditions, it significantly increased in high-variability conditions, going towards the principle of *reduction of uncertainty* (Gibson, 1991). When transitional probabilities are high, adjacent elements are perceived as invariant, whereas when high variability disrupts adjacent dependencies, learners tend to seek alternative sources of predictability.

Newport and Aslin (2004) showed that adult learners are highly selective in the types of non-adjacent regularities they are readily able to compute: in particular, they found that non-adjacent syllable regularities are extremely hard to acquire, while segment regularities (i.e., co-occurrences of sets of consonants like roots of semitic languages, but note that their results also applied so segments made up of vowels) are much easier, and provide various possible explanation to this, namely **element similarity** (i.e., homogeneity of the elements forming the pattern), or the interaction between distance and elements' representations.

Such results are further confirmed by Gómez and Maye (2005), that also investigate differences in the detection of nonadjacent dependencies in respect of the different development stages of their subjects and in different conditions of variability (i.e., *set-size*, in stings like aXb, the pool from which the element X was drawn varied in size). In spite of the nonadjacent elements being always perfectly predictable, learners showed to be able to track the less reliable adjacent probabilities in all but the highest set-size conditions ($|\{X \text{ in } aXb\}| = 18, 24$), showing to be attracted by adjacent probabilities also if it would have been more useful to track the nonadjacent ones.

An important and related point raised in Mintz (2002, 2003) and Onnis et al. (2004) about the results on variability provided by Gomez (2002), is how high variability of context may be instrumental in the identification of frequent frames, that lead in turn to inferences on syntactic categories.

Gómez and Maye (2005) also mention that results such as the ones reported in Peña et al. (2002), which are apparently at odds with the aforementioned studies, can be explained in light of some cues in the stimuli that Peña et al. did not control for, although they recognize that the existence of two different learning mechanisms, one pertaining to component discovery and one pertaining to structural regularities discovery, is still not ruled out by the current literature. Peña et al. (2002) argue in fact, in a more generative fashion, that, while the process of chunking the linguistic stream relies on statistical computations, the process of projecting generalizations on grammatical regularities that go beyond stored items in memory could be non statistical in nature.

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, they embody the building blocks that bridge the traditional lexical level to the sentence level, being therefore central to the issues of linguistic creativity and compositionality.

4 Connectionist approaches in statistical learning and usage-based language models

Computational connectionism has provided a solid framework to implement many of the theories of statistical learning and grammar induction. The approaches are so many and so varied that it would be out of the scope of this work to have a proper review of them. We've already mentioned the path-breaking study of Elman (1990), which has also been widely discussed in the subsequent literature.

In the following paragraphs we will therefore summarize a few models that have been proposed concerning statistical learning and chunking, and introduce the spiking paradigm, that has been successfully applied to auditory and visual tasks and shows interesting properties with respect to its cognitive plausibility.

4.1 TRACX2

The TRACX2 model (Mareschal and French, 2017) tackles the issue of statistical learning by arguing that both transitional probabilities learning and chunking can coexist in one system, as it is one single mechanism that underlies sequantial learning, Hebbian-style learning.

In a previous version of the model (French et al., 2011), the authors had shown that a connectionist autoencoder, augmented with conditional recurrence, could extract chunks from a stream, successfully capturing data from the adult and infant auditory statistical learning literature.

Both TRACX and TRACX2 consist of an autoencoder with two identical banks of inputs units, two identical banks of output units (each of which is the same size as each of the banks of input units), and a bank of hidden units with the same dimensions as one of the input/output unit banks: however, while in TRACX a threshold value indicating successful recognition of the current pair of input elements was used to decide whether the input on the right bank was being transferred to the left bank or not, TRACX2 removes the use of an all-or-nothing threshold, making the contribution of the hidden-unit activation vector to the left bank of input units graded and depending on the level of learning already achieved.

The key aspect of TRACX2 is that it is encoding and recognizing previously seen chunks of information, rather than just internalizing the overall statistical structure of the sequence. This fits with the idea that infant statistical learning can be explained through a memory-based chunking model, and also endorses the fact that sequence processing can emerge from the application of fairly general associative mechanisms.

4.2 Reconciling episodic memory with statistical learning

Moving to a fairly different scenario, the work of Schapiro et al. (2017) aims at modeling the function of the hippocampus, reconciling its rapid learning function with the idea that it specializes in memorizing individual episodes.

Therefore, the authors exposed a neural network model that instantiates known properties of hippocampal projections and subfields to sequences of items with temporal regularities and asked whether it is possible for the hippocampus to handle both statistical learning and memorization of individual episodes.

The model they are providing is particularly relevant in the Complementary Learning Systems (CLS) theory (McClelland et al., 1995), which provides itself a computational framework for understanding the distinct roles that the hippocampus and neocortex play in representing memories. The theory posits the existence of different systems: the *hippocampus*, with high learning rate and sparse, non-overlapping representations to quickly store memory traces for recent experience, and *cortical areas*, informed by the hippocampus during offline periods, with slow learning rate and overlapping representations that allow to represent regularities across experiences. It has been claimed, however, that the hippocampus is also involved in rapid statistical learning (Schapiro and Turk-Browne, 2015), that requires learning regularities over a short period of time (e.g., minutes), posing a challenge for the theories that the theories that view the hippocampus as solely supporting memory for distinct episodes.

Schapiro et al. therefore investigate the role of hippocampal structures across different subfields, and do this through three learning paradigms that require the extraction of regularities on different timescales: in particular, they test *classic statistical learning*, in which a continuous sequence of items is presented to participants during passive viewing or a cover task, *community structure*, which still involves a continuous sequence of items but with uninformative transitional probabilities among adjacent items and therefore requiring sensitivity to higher-level associations, and a third class of tasks including *transitive inference*, *acquired equivalence*, and *associative inference* that also require indirect associations, this time not based on segmentation but rather on rapid integration of experience over time in order to uncover regularities.

Their results suggest a modification of the CLS framework, allowing for both novel episodic and novel statistical information to be quickly learned in the hippocampus, but different substructures subserve different functions, thus reconciling the trade-off between episodic memory and statistical learning by suggesting that the hippocampus itself contains complementary learning systems.

4.3 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.

According to the previous section, also another source of implausibility has to be dealt with: in neural network models, regularities are usually and most effectively extracted through overlapping representations, but as the Schapiro et al. (2017) model and CLS theory 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.

Just like traditional ANNs, Spiking Neural Networks are directed graphs made of nodes (*neurons*) and edges (*synapses*). What differentiates the two frameworks is that SNNs operate using *spikes*, discrete events that take place at points in time, rather than continuous values, that are produced in ANNs by the activation function. In the biological metaphor, each neuron of a SNN has a time-dependent variable that serves as the biological membrane, which integrates the received inputs over a portion of the stream and determines the response produced by the neuron (i.e., it regulates the production of a spike). Various models have been proposed for the neuron, the simplest being the *Leaky integrate-and-fire* (LIF) model.

Also supervised learning algorithms such as the well-known backpropagation are to be revised for the SNN framework: differently from ANNs, Spiking Neural Networks encode input sequences of spikes in output sequences of spikes (Jeong, 2018), thus opening the path for different solutions to learning, some of them implementing bio-inspired local training rules.

4.3.1 Neuronal structure

As an example[1], we will consider the case of a neuron involving just one state variable, which represents the membrane potential, similarly to the LIF model (Delorme et al., 1999), which is among the most popular neuron models.

The idea is that, over time, the value of this activation potential increases when a spike is received and decays during the inter-spike intervals. The state variable is therefore expressed as follows:

$$u(t) = u_0 + a \int_0^t D(s) \cdot w \cdot \sigma(t-s) \mathrm{d}s \tag{1}$$

where u_0 and a are the initial membrane potential and a positive constant respectively, $D(s)$ is a linear filter, w the synaptic weight and σ a series of N input spikes, namely $\sigma(t) = \sum_{i=1}^{N} \delta(t - t_i)$. In other words, the state of the membrane at time t is given by its initial state u_0 plus some additional potential due to the received spike stream. A spike is then elicited at time t if the value of $u(t)$ reaches a threshold u_{th}, and the potential is consequently reset to u_0.

The function of the linear filter D allows for modulations in how the integration over spikes works, therefore mimicking memory loss.

The weight w characterizes the synapse as excitatory or inibitory

More complex models are of course possible, such as the ones that take into consideration excitatory post-synaptic current or account for other specific biological aspects (see for example Izhikevich (2003)).

4.3.2 Encoding the input

The brain encodes external, analog information into electrical pulses (i.e., spikes): the schemes used to encode real data into spike streams are therefore a relevant aspect of the SNN architecture (Kasabov, 2018).

[1] We refer to Jeong (2018) notation for this section.

Rate coding implies encoding a sequence of spikes based on the average number of spikes (or spikes count) over time i.e. how many spikes are emitted within a time window. In literature, the expression *rate coding* is used to refer to three different notions, that reflect three different views of the code: an average over time, or an average over a population of neurons, or an average over several experimental repetitions.

Temporal coding is mostly used to encode real value data (e.g., sound, pixel images, temperature...) into spike sequences. A spike is generated only if a change in the input data occurs beyond a threshold. An example is a code where for each neuron the timing of the first spike after the reference signal contains all information about the new stimulus: firing shortly after the reference signal means a strong stimulation, while delaying the firing would signal a weaker stimulation (Thorpe et al., 1996). Similarly to *rate coding*, this strategy can be viewed at the population level, for example distributing the input into multiple neurons with overlapping receptive fields, represented as a continuous function (e.g., Gaussian).

Many variants are discussed in Kasabov (2018), and it is hard to draw a clear boundary between rate based codes and pulse-based temporal codes. As pointed out in Gerstner and Kistler (2002), the minimum requirement for any code is that it is able to offer the possibility to react quickly to changes in the input, in order to march behavioral reaction times.

4.3.3 Learning

As in traditional ANNs, the learning is performed by adjusting the weights of the synapses. Differently from standard ANNs, in SNNs learning is local both with respect to the neighborhood of the synapse and in time, and in particular the weight of the synapse connecting two neurons is adjusted following various variants of the spike-timing-dependent plasticity (STDP) algorithm, which are largely inspired by the basic Hebb rule of learning, well summarized in its most popular formulation *"cells that fire together wire together"*.

Another difference to consider is that the spike trains are expressed as sums of Dirac delta functions (Tavanaei et al., 2018), that make derivative optimization such as backpropagation difficult to apply. Therefore, different solutions have to be implemented for multi-layered networks.

Unsupervised learning via STDP

The idea of STDP is that the temporal relation between the pre- and post-synaptic spike influences the strength of the connection. In particular, if the post-synaptic neuron fires shortly after the pre-synaptic one, the synapse is strenghtened, while if the opposite happens the relation is considered spurious and the connection is depressed. An example is reported in the following equation[2]:

$$\Delta w = \begin{cases} Ae^{\frac{-(\|t_{pre}-t_{post}\|)}{\tau}} & t_{pre} - t_{post} \leq 0, A > 0 \\ Be^{\frac{-(\|t_{pre}-t_{post}\|)}{\tau}} & t_{pre} - t_{post} > 0, B < 0 \end{cases} \tag{2}$$

where A, B are constant parameters indicating learning rates and τ is the temporal window.

Supervised learning

In SNNs, the objective function that is minimized during the learning is the so-called *readout* error, a cost function that capures the difference between the desired and output spike streams.

Here we cite an example of updating rules which is among the most intuitive ones, namely the one of the ReSuMe (remote supervised learning) algorithm (Ponulak and Kasiński, 2010), that is an adaptation of the standard Delta rule:

$$\Delta w = (y^d - y^0)x \tag{3}$$

where x is the pre-synaptic input and y^d and y^0 are the desired and observed outputs respectively. When reformulated for SNNs, the rule for training excitatory synapses takes the form:

$$\Delta w = \Delta w^{STDP}(S^{in}, S^d) + \Delta w^{aSTDP}(S^{in}, S^0) \tag{4}$$

where Δw^{STDP} is a function of the correlation of the presynaptic and desired spike trains, while Δw^{aSTDP} depends on presynaptic and observed spike trains.

In another model with a similar framework, the Chronotron (Florian, 2012), the Victor-Purpura (VP) distance metric (similar to Levenshtein distance) is introduced in order to compute the difference between two spike trains, in terms of *the minimum cost of transforming one spike train into the other by creating, removing or moving spikes.* This was made fully differentiable and therefore used as a cost function.

4.3.4 Deep Learning and applications

An accurate and extensive review of the most recent deep learning models involving SNNs is reported in Tavanaei et al. (2018).

The general picture has seen big strides in the SNN framework in the most recent years, however, in spite of overall good and competitve results, the tasks that have been explored are still very few, and researchers have mostly focused on the Modified National Institute of Stan- dards and Technology (MNIST) dataset (LeCun et al., 2010).

Moreover, the approaches have mainly focused on adapting backpropagation to the SNN architechture, while upscaling biologically inspired algorithms such as STPD to more complex architectures still represents a challenge.

A special mention is needed for Liquid State Machines (LSM) architechtures (Maass et al., 2002), born natively with spiking neurons not to the purpose

[2] We refer to Tavanaei et al. (2018) notation for this section.

of modeling a specific task but rather of reproducing the dynamics of cortical circuits.

One of the few applications of SNN to language modeling can be found in Costa et al. (2017), who attempted to adapt the architecture of a conventional LSTM to be a plausible model of a cortical microcircuit, involving LIF neurons. Moreover, the gating operations, which are multiplicative in a traditional LSTM, were replaced by substractions (hence, *subtractive LSTM* or *subLSTM*), which are closer to biological functions. Although not outperforming LSTMs, subLSTMs achieved a comparable level of perplexity in a simple word-prediction task, therefore opening up promising paths for future research.

5 Bridging the gap

The picture we tried to draw throughout this brief summary leaves many questions unanswered. First and foremost, the issue of the emergence of non-adjacent dependencies still represents a puzzle both from a linguistic and computational point of view. Moreover, 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. We however think that the same question could be posed in a different fashion, in order to be better integrated with the statistical learning and more cognitive-based community: *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, could 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.

Building meaning representations and segmenting the input are in fact deeply related and mutually informative tasks, and tackling the two issues together would could be beneficial as well as lead to a model that has a more cognitively informed approach to the process of acquisition of linguistic structures.

References

Gerry TM Altmann. 2017. Abstraction and generalization in statistical learning: implications for the relationship between semantic types and episodic tokens. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1711):20160060.

Gerry TM Altmann and Jelena Mirković. 2009. Incrementality and prediction in human sentence processing. *Cognitive science*, 33(4):583–609.

Blair C Armstrong, Ram Frost, and Morten H Christiansen. 2017. The long road of statistical learning research: past, present and future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1711).

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.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017a. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872.

Yonatan Belinkov, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2017b. Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10.

Jos JA van Berkum, Peter Hagoort, and Colin M Brown. 1999. Semantic integration in sentences and discourse: Evidence from the n400. *Journal of cognitive neuroscience*, 11(6):657–671.

Terra Blevins, Omer Levy, and Luke Zettlemoyer. 2018. Deep rnns encode soft hierarchical syntax. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 14–19.

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.

M. H. Christiansen. 2018. Implicit statistical learning: A tale of two literatures. *Topics in cognitive science.*

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.

Rui Costa, Ioannis Alexandros Assael, Brendan Shillingford, Nando de Freitas, and TIm Vogels. 2017. Cortical microcircuits as gated-recurrent neural networks. In *Advances in neural information processing systems*, pages 272–283.

Arnaud Delorme, Jacques Gautrais, Rufin Van Rullen, and Simon Thorpe. 1999. Spikenet: A simulator for modeling large networks of integrate and fire neurons. *Neurocomputing*, 26:989–996.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ariel Ekgren, Amaru Cuba Gyllensten, and Magnus Sahlgren. 2018. R-grams: Unsupervised learning of semantic units in natural language. *arXiv preprint arXiv:1808.04670*.

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

Jeffrey L Elman. 2009. On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive science*, 33(4):547–582.

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

Katharine Graf Estes, Julia L Evans, Martha W Alibali, and Jenny R Saffran. 2007. Can infants map meaning to newly segmented words? statistical segmentation and word learning. *Psychological science*, 18(3):254–260.

Răzvan V Florian. 2012. The chronotron: a neuron that learns to fire temporally precise spike patterns. *PloS one*, 7(8):e40233.

Robert M French, Caspar Addyman, and Denis Mareschal. 2011. Tracx: A recognition-based connectionist framework for sequence segmentation and chunk extraction. *Psychological Review*, 118(4):614.

Ram Frost, Blair C Armstrong, Noam Siegelman, and Morten H Christiansen. 2015. Domain generality versus modality specificity: the paradox of statistical learning. *Trends in cognitive sciences*, 19(3):117–125.

Wulfram Gerstner and Werner M Kistler. 2002. *Spiking neuron models: Single neurons, populations, plasticity.* Cambridge university press.

Eleanor Jack Gibson. 1991. *An odyssey in learning and perception.* The MIT Press.

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 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.

Rebecca L Gomez, Louann Gerken, and Roger W Schvaneveldt. 2000. The basis of transfer in artificial grammar learning. *Memory & Cognition*, 28(2):253–263.

Peter Hagoort, Colin Brown, and Jolanda Groothusen. 1993. The syntactic positive shift (sps) as an erp measure of syntactic processing. *Language and cognitive processes*, 8(4):439–483.

Peter Hagoort, Lea Hald, Marcel Bastiaansen, and Karl Magnus Petersson. 2004. Integration of word meaning and world knowledge in language comprehension. *science*, 304(5669):438–441.

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.

Falk Huettig. 2015. Four central questions about prediction in language processing. *Brain research*, 1626:118–135.

Falk Huettig and Nivedita Mani. 2016. Is prediction necessary to understand language? probably not. *Language, Cognition and Neuroscience*, 31(1):19–31.

Eugene M Izhikevich. 2003. Simple model of spiking neurons. *IEEE Transactions on neural networks*, 14(6):1569–1572.

Ray Jackendoff. 2007. A parallel architecture perspective on language processing. *Brain research*, 1146:2–22.

Doo Seok Jeong. 2018. Tutorial: Neuromorphic spiking neural networks for temporal learning. *Journal of Applied Physics*, 124(15):152002.

Tobias Johansson. 2009. Strengthening the case for stimulus-specificity in artificial grammar learning: no evidence for abstract representations with extended exposure. *Experimental Psychology*, 56(3):188–197.

Michael N Jones and Douglas JK Mewhort. 2007. Representing word meaning and order information in a composite holographic lexicon. *Psychological review*, 114(1):1.

N Kalchbrenner, E Grefenstette, and Philip Blunsom. 2014. A convolutional neural network for modelling sentences. In *52nd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Nikola K Kasabov. 2018. *Time-space, spiking neural networks and brain-inspired artificial intelligence*, volume 7. Springer.

Barbara J Knowlton and Larry R Squire. 1994. The information acquired during artificial grammar learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20(1):79.

Gina R Kuperberg. 2007. Neural mechanisms of language comprehension: Challenges to syntax. *Brain research*, 1146:23–49.

Marta Kutas and Steven A Hillyard. 1980. Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207(4427):203–205.

Thomas K Landauer and Susan T Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.

Yann LeCun, Corinna Cortes, and CJ Burges. 2010. Mnist handwritten digit database. *AT&T Labs [Online]. Available: http://yann. lecun. com/exdb/mnist*, 2:18.

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.

Kevin Lund and Curt Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior research methods, instruments, & computers*, 28(2):203–208.

Gary Lupyan and Andy Clark. 2015. Words and the world: Predictive coding and the language-perception-cognition interface. *Current Directions in Psychological Science*, 24(4):279–284.

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

Wolfgang Maass, Thomas Natschläger, and Henry Markram. 2002. Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural computation*, 14(11):2531–2560.

Maryellen C MacDonald, Neal J Pearlmutter, and Mark S Seidenberg. 1994. The lexical nature of syntactic ambiguity resolution. *Psychological review*, 101(4):676.

Denis Mareschal and Robert M French. 2017. Tracx2: a connectionist autoencoder using graded chunks to model infant visual statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1711):20160057.

Jessica Maye, Janet F Werker, and LouAnn Gerken. 2002. Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, 82(3):B101–B111.

Stewart M McCauley and Morten H Christiansen. 2011. Learning simple statistics for language comprehension and production: The cappuccino model. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33.

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.

Ken McRae and Kazunaga Matsuki. 2009. People use their knowledge of common events to understand language, and do so as quickly as possible. *Language and linguistics compass*, 3(6):1417–1429.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

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.

Toben H Mintz. 2002. Category induction from distributional cues in an artificial language. *Memory & Cognition*, 30(5):678–686.

Toben H Mintz. 2003. Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, 90(1):91–117.

Jennifer B Misyak, Morten H Christiansen, and J Bruce Tomblin. 2010. Sequential expectations: The role of prediction-based learning in language. *Topics in Cognitive Science*, 2(1):138–153.

NJ Larsson'_and A Moffat and J Larsson. 2000. Offline dictionary-based compression. In *Data Compression Conference*, pages 296–305.

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.

Mante S Nieuwland and Jos JA Van Berkum. 2006. When peanuts fall in love: N400 evidence for the power of discourse. *Journal of cognitive neuroscience*, 18(7):1098–1111.

Luca Onnis, Padraic Monaghan, Morten H Christiansen, and Nick Chater. 2004. Variability is the spice of learning, and a crucial ingredient for detecting and generalizing in nonadjacent dependencies. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 26.

Lee Osterhout and Phillip J Holcomb. 1992. Event-related brain potentials elicited by syntactic anomaly. *Journal of memory and language*, 31(6):785–806.

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

Pierre Perruchet and Chantal Pacteau. 1990. Synthetic grammar learning: Implicit rule abstraction or explicit fragmentary knowledge? *Journal of experimental psychology: General*, 119(3):264.

Pierre Perruchet and Annie Vinter. 1998. Parser: A model for word segmentation. *Journal of memory and language*, 39(2):246–263.

Pierre Perruchet, Annie Vinter, Chantal Pacteau, and Jorge Gallego. 2002. The formation of structurally relevant units in artificial grammar learning. *The Quarterly Journal of Experimental Psychology: Section A*, 55(2):485–503.

Cyma Van Petten. 1993. A comparison of lexical and sentence-level context effects in event-related potentials. *Language and Cognitive Processes*, 8(4):485–531.

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

Filip Ponulak and Andrzej Kasiński. 2010. Supervised learning in spiking neural networks with resume: sequence learning, classification, and spike shifting. *Neural computation*, 22(2):467–510.

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

Arthur S Reber. 1967. Implicit learning of artificial grammars. *Journal of verbal learning and verbal behavior*, 6(6):855–863.

Martin Redington and Nick Chater. 1996. Transfer in artificial grammar learning: A reevaluation. *Journal of experimental psychology: general*, 125(2):123.

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, Richard N Aslin, and Elissa L Newport. 1996. Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928.

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.

A Schapiro and Nicholas Turk-Browne. 2015. Statistical learning. *Brain mapping*, 3:501–506.

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.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural mt learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1526–1534.

Amirhossein Tavanaei, Masoud Ghodrati, Saeed Reza Kheradpisheh, Timothée Masquelier, and Anthony Maida. 2018. Deep learning in spiking neural networks. *Neural Networks*.

Erik D Thiessen. 2007. The effect of distributional information on children's use of phonemic contrasts. *Journal of Memory and Language*, 56(1):16–34.

Erik D Thiessen. 2017. What's statistical about learning? insights from modelling statistical learning as a set of memory processes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1711):20160056.

Erik D Thiessen, Alexandra T Kronstein, and Daniel G Hufnagle. 2013. The extraction and integration framework: A two-process account of statistical learning. *Psychological bulletin*, 139(4):792.

Erik D Thiessen and Philip I Pavlik Jr. 2013. iminerva: A mathematical model of distributional statistical learning. *Cognitive Science*, 37(2):310–343.

Erik D Thiessen and Jenny R Saffran. 2003. When cues collide: use of stress and statistical cues to word boundaries by 7-to 9-month-old infants. *Developmental psychology*, 39(4):706.

Simon Thorpe, Denis Fize, and Catherine Marlot. 1996. Speed of processing in the human visual system. *nature*, 381(6582):520.

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.

Chen Yu and Dana H Ballard. 2007. A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, 70(13-15):2149–2165.