Citation: Endress, A. (2020). A simple, biologically plausible feature detector for language acquisition. Journal of Cognitive Neuroscience, 32(3), pp. 435-445. doi: 10.1162/jocn_a_01494

Link to published version: https://doi.org/10.1162/jocn_a_01494

A simple, biologically plausible feature detector for language acquisition

Ansgar D. Endress
City, University of London

Running head: A feature detector for language acquisition

Abstract

Language has a complex grammatical system we still have to understand computationally and biologically (Hauser et al., 2002; Yang, 2013). However, some evolutionarily ancient mechanisms have been repurposed for grammar (Dehaene & Cohen, 2007; Endress, Cahill, et al., 2009; Endress, Nespor, et al., 2009; Fitch, 2017) so that we can use insight from other taxa into possible circuit-level mechanisms of grammar. Drawing upon recent evidence for the importance of disinhibitory circuits across taxa and brain regions (Chevalier & Deniau, 1990; Letzkus et al., 2015; Hangya et al., 2014; Xu et al., 2013; Goddard et al., 2014; Mysore & Knudsen, 2012; Koyama et al., 2016; Koyama & Pujala, 2018), I suggest a simple circuit that explains the acquisition of core grammatical rules used in 85% of the world's languages (Rubino, 2013): grammatical rules based on sameness/difference relations. This circuit acts as a sameness-detector. *Different* items are suppressed through inhibition, but presenting two *identical* items leads to inhibition of inhibition. The items are thus propagated for further processing. This sameness-detector thus acts as a feature detector for a grammatical rule. I suggest that having a set of feature detectors for elementary grammatical rules might make language acquisition feasible based on relatively simple computational mechanisms.

Keywords: Language Acquisition; Rule Learning; Perceptual or Memory Primitives; Disinhibition; Circuit Motifs; Reduplication

A simple, biologically plausible feature detector for language acquisition

Language acquisition is fast, largely based on positive evidence (or sometimes no evidence at all; Goldin-Meadow & Mylander, 1998; Senghas et al., 2004), goes far beyond what learners hear or see in their environment (Chomsky, 1959; Pinker, 1984) and results in a uniquely complex grammatical system that stands out in the animal kingdom (Hauser et al., 2002; Yang, 2013). Even seemingly straightforward "memory" problems such as learning the meanings of words hide complexities that call for human-specific grammatical adaptations (Medina, Snedeker, Trueswell, & Gleitman, 2011; Pinker & Jackendoff, 2005). Unsurprisingly, we know very little about the underlying computational mechanisms at the circuit level.

However, some linguistic mechanisms are evolutionarily ancient and have been repurposed for linguistic use (Dehaene & Cohen, 2007; Endress, Cahill, et al., 2009; Endress, Nespor, et al., 2009; Fitch, 2017). In such cases, it might be possible to identify core linguistic mechanism whose systems-level implementation might be tractable due to its evolutionary history.

Here, I use sameness/difference relations as a case in point. I will first show that many grammatical rules are based on such relations, especially in morphology and phonology, but that similar relations are critical in many other domains and animals, suggesting that they reflect a linguistic core mechanism with evolutionarily ancient roots. I will then suggest that such relations can be computed using an ubiquitous processing motif: disinhibition among neurons or neural populations.

Sameness/difference relations in language and other domains and animals

Sameness/difference relations are critical for many aspects of linguistic structure, especially in phonology and morphology. For example, some 85% of the world's languages use some form of reduplication (Rubino, 2013). Among many other uses, reduplications can signal changes in word class (e.g., from noun to verb, as in the Marshallese contrast between "ta*kin* – sock" and "ta*kinkin* – to wear socks"; Moravcsik, 1978), attenuation (as in the Alabama contrast between "ka*s*atka – cold" and "ká*ss*atka – cool"; Hardy & Montler, 1988) or intensification; they can mark differences in number (e.g., singular vs. plural), tense (e.g., past vs. present), aspect (e.g., continued vs. repeated occurrence or temporary vs. permanent), size or case (see Rubino, 2013, and references therein).

Phonological processes also often appeal to sameness/difference relations, with some processes requiring some features to be identical within a relevant constituent, and others requiring them to be different. Processes that require identical features include vowel harmony and assimilation. Specifically, in languages with vowel harmony, vowels within words (or smaller domains) need to have one or more features in common (Rose & Walker, 2011). For example, Hungarian words generally have either only back vowels or only front vowels; grammatical suffixes thus come in two varieties, one with back vowels and one with front vowels. Accordingly, the dative suffix is –*nak* for words like "ablak – window" (resulting in forms like "ablaknak") and –*nek* for words like "bíró – judge" (resulting in forms like "bírónek"; Hayes & Londe, 2006). Likewise, in languages with consonant assimilation, consonants must share a feature with other surrounding consonants. For example, in English, "football" might be pronounced as “foo*p*ball” because the place of articulation of the [t] at the end of [foot] gets assimilated to the place of articulation of the [b] at the start of “ball”; in contrast, in French, “football” might be pronounced as “foo*d*ball” because the voicing feature of the [t] (but not the place feature) gets assimilated to the following [b] (Darcy, Ramus, Christophe, Kinzler, & Dupoux, 2009). Both vowel harmony and assimilation thus introduce sameness relations among phonemes. Listeners use these sameness relations not only in word recognition (Darcy et al., 2009; Mitterer & Blomert, 2003; Suomi, McQueen, & Cutler, 1997), but also as cues to learn new words (Vroomen, Tuomainen, & de Gelder, 1998). Further, sameness relations in the form of vowel harmony often interact with other area of grammar, such as stress assignment or morphology (Rose & Walker, 2011).

While vowel harmony and assimilation require sameness relations among phonemic features, other phonological processes impose difference relations. Such processes include the Obligatory Contour Principle (Frisch, Pierrehumbert, & Broe, 2004; McCarthy, 1986). Initially, the Obligatory Contour Principle was proposed to account for the observation that, in certain tone languages, tones cannot be repeated within words, but it also applies to other phonological phenomena. For example, in Semitic languages like Arabic and Hebrew, the basic meaning of verbs is given by their consonantal root; roots like /k t b/ are then transformed into surface forms such as “*k*a*t*a*b*a – he wrote” and “*k*u*t*i*b*a – it was written” (Frisch et al., 2004). The OCP prevents consonantal roots from having repeated consonants, while other morphological processes can *create* (rather than prevent) sameness relations among consonants (Frisch et al., 2004; McCarthy, 1986). Such rules might also interact with other areas of grammar (Yip, 1988) and speakers apply them even when presented with novel non-sense words (e.g., Berent & Shimron, 1997; Frisch & Zawaydeh, 2001).

Sameness relations are also important during language acquisition. Reduplications are prominent in child-directed speech across languages (Ferguson, 1964) and children themselves “invent” forms with reduplicated syllables; these reduplicated forms might be important for acquiring multisyllabic words (Schwartz, Leonard, Wilcox, & Folger, 1980) and syllable-final consonants that would otherwise be lost (Fee & Ingram, 1982).

More generally, sameness relations have been critical for defining the computational complexity of phonological rules (Culy, 1985; Manaster-Ramer, 1986), and, in developmental psychology, rules based on sameness relations have been the most prominent assay for studying rule-learning in human infants (Marcus et al., 1999), to the extent that in a recent meta-analysis of “rule-learning” in infancy, rule-learning was treated as synonymous with the learning of sameness relations (Rabagliati, Ferguson, & Lew-Williams, 2019).

Sameness relations are also important for other forms of language use. Not only are rhymes and alliterations important in poetry (Fabb, 2015), but many language games that spontaneously arise in children also make extensive use of sameness relations in the form of reduplications (Bagemihl, 1995). For example, in the Chinese May-ka language game, syllables are duplicated and then the vowel of the first duplicate is replaced by “ay” and the consonant of the second duplicate by “k”; *ma* (mother) thus becomes *may-ka* (Bao, 1990; Yip, 1982).

Despite their simplicity, sameness relations thus appear to be a core part of the language faculty.

However, sameness/difference rules are clearly not specific to language. They are crucial for many other aspects of cognition, including motor learning (Brooks, 1986), any comparison of sensory input to predictions or internal state (e.g., novelty detection in the hippocampus; Kumaran & Maguire, 2007) and short-term memory tasks such as delayed-match to sample tasks (Cope et al., 2018; Engel & Wang, 2011). Accordingly, grammar-like rules based on sameness/difference relations can be learned in many non-linguistic domains in humans (Dawson & Gerken, 2009; Endress, Dehaene-Lambertz, & Mehler, 2007; Marcus, Fernandes, & Johnson, 2007; Saffran, Pollak, Seibel, & Shkolnik, 2007) and by many non-human animals (de la Mora & Toro, 2013; Hauser & Glynn, 2009; Martinho & Kacelnik, 2016; Murphy, Mondragon, & Murphy, 2008; Neiworth, 2013; Pepperberg, 1987; Smirnova, Zorina, Obozova, & Wasserman, 2015; Versace, Spierings, Caffini, Ten Cate, & Vallortigara, 2017; but see Heijningen, Visser, Zuidema, & Cate, 2009; Hupé, 2017; Langbein & Puppe, 2017), possibly through a specialized sameness-detector (Endress, 2013; Endress et al., 2007) that might exist from birth (Antell, Caron, & Myers, 1985; Gervain, Berent, & Werker, 2012; Gervain, Macagno, Cogoi, Peña, & Mehler, 2008). The computations underlying sameness/difference relations thus reflect a core linguistic mechanism whose systems-level implementation might be tractable due to its evolutionary history.

Disinhibition-based computations

Here, drawing upon recent evidence stressing the importance of disinhibitory circuits (neurons that inhibit other inhibitory neurons) across a variety of taxa and brain regions (Chevalier & Deniau, 1990; Goddard et al., 2014; Hangya et al., 2014; Koyama et al., 2016; Mysore & Knudsen, 2012; Xu et al., 2013), I suggest a simple circuit that acts as a sameness-detector. Disinhibition has been observed in a variety of brain areas (Chevalier & Deniau, 1990; Letzkus et al., 2015), and some interneuron populations specifically inhibit other inhibitory interneurons (Hangya et al., 2014; Xu et al., 2013). Critically, some interneuron types receive both local and long-range input; such interneurons have been found to inhibit other inhibitory interneurons in auditory (Pi et al., 2013), visual (Pfeffer, Xue, He, Huang, & Scanziani, 2013), somatosensory (Lee, Kruglikov, Huang, Fishell, & Rudy, 2013) and prefrontal cortex (Pi et al., 2013), from where they can exert spatially remarkably specific disinhibition on other populations (Zhang et al., 2014). Accordingly, Hangya et al. (2014) argued that this disinhibitory circuit might be a cortical circuit motif. Other authors suggested a more local disinhibitory circuit motif with mutual inhibition among inhibitory neurons (Goddard et al., 2014; Koyama et al., 2016; Koyama & Pujala, 2018; Mysore & Knudsen, 2012).

Disinhibitory circuits have been proposed to account for a variety of cognitive phenomena, including attentional selection (van Der Velde & de Kamps, 2001; Zhang et al., 2014), gain control (Fu et al., 2014), sequential discriminations of stimulus strength of stimuli (Machens, Romo, & Brody, 2005; Miller & Wang, 2006; but see Barak, Sussillo, Romo, Tsodyks, & Abbott, 2013) categorization of stimuli (Goddard et al., 2014; Kusunoki, Sigala, Nili, Gaffan, & Duncan, 2010; Mysore & Knudsen, 2012), behavioral response selection (Jovanic et al., 2016; Zhao et al., 2019), associative learning (Letzkus et al., 2011), plasticity (Fu, Kaneko, Tang, Alvarez-Buylla, & Stryker, 2015) and social behavior (Marlin, Mitre, D'amour, Chao, & Froemke, 2015; Owen et al., 2013). Here, I suggest that the same biological mechanisms might provide a circuit-level mechanism for a core grammatical computation based on sameness vs. difference computations.

Models of sameness/difference relations

A number of models of how sameness-relations might be computed have been proposed in the literature (Arena et al., 2013; Carpenter & Grossberg, 1987; Cope et al., 2018; Engel & Wang, 2011; Hasselmo & Wyble, 1997; J. S. Johnson, Spencer, Luck, & Schöner, 2009; Ludueña & Gros, 2013; Wen, Ulloa, Husain, Horwitz, & Contreras-Vidal, 2008). The underlying principles and assumptions vary substantially across models. Some rely on the fact that repeatedly activated representations suffer some form of neural "fatigue" (Grill-Spector, Henson, & Martin, 2006; Kumaran & Maguire, 2007), others on circuitry where the *combined* input from some form of memory and from sensory representations matching (or mismatching) the memory representations must be sufficiently strong (Carpenter & Grossberg, 1987; Hasselmo & Wyble, 1997; Wen et al., 2008) or where the *difference* between input from memory and from sensory representations is the critical variable (Engel & Wang, 2011). Still other models detect reduced levels inhibition for novel compared to previously encountered items (Cope et al., 2018; J. S. Johnson et al., 2009). I discuss these models in more detail in Supplementary Material 1, where I show that they fall short on at least one of two criteria of grammar learning: they either do not generalize to unseen exemplars or they require labeled counter-examples.

To better illustrate the computational principles underlying the current dishibition-based circuit, I will first present a version of the model that can detect sameness relations in sequentially presented stimuli. Following this, I will sketch a version of the model that can detect sameness relations in spatially distributed, simultaneously presented stimuli, and finally a model that can detect sameness relations in both simultaneously presented stimuli and sequentially presented stimuli.

Results

Sameness detection for sequential stimuli

Figure 1a shows a possible disinhibition-based architecture of how sameness might be detected for sequentially presented items. (Model equations are given in Appendix A; an R implementation is available online). The model comprises two populations of neurons (hereafter "layers") that encode features of items (e.g., frequency, color and so on; in Figure 1, the features are represented as geometric shapes).

The *source layer* receives input; input can be sensory or non-sensory, depending on where this circuit is located in the brain. Units in the *copy layer* receive excitatory one-to-one input from units in the source layer that code for the same feature. However, they also receive feature-specific tonic inhibition from an *inhibition layer* (which might consist of interneurons); tonic inhibition has been observed in a variety of brain regions, and might subserve functions such as maintaining an appropriate level of excitability or the suppression of undesirable motor programs (Benjamin, Staras, & Kemenes, 2010; Farrant & Nusser, 2005; Semyanov, Walker, Kullmann, & Silver, 2004).

Due to the inhibition from the inhibition layer to the copy layer, input from the source layer is not propagated to the copy layer with a single stimulation. The critical aspect of this circuit is that each feature in the source layer also *inhibits* the corresponding feature in the inhibition layer, which, in turn, reduces inhibitory input to the copy layer for that feature. A similar phenomenon has been observed in auditory fear conditioning, where inhibition of (inhibitory) parvalbumin-positive interneurons allowed for associations between sounds and aversive stimuli to be formed (Letzkus et al., 2011).

Accordingly, once the inhibitory input to the copy layer ceases, there will be a time window during which the excitatory input from the source layer can drive the corresponding units in the copy layer. As a result, only repeated items will be propagated to the copy layer. Any readout mechanism for the copy layer (e.g., a population of thresholded neurons) could thus act as a sameness-detector.[1]

---

[1] While I model disinhibition across different neural populations, the same computational principles could be implemented using reciprocal inhibition among inhibitory neurons as in earlier models of stimulus selection and categorization (Goddard, Mysore, Bryant, Huguenard, & Knudsen, 2014; Koyama et al., 2016; Koyama & Pujala, 2018; Mysore & Knudsen, 2012). To do so, one would simply replace the inhibitory connections from the source layer to the inhibition layer with inhibition in the source layer that is itself subject to lateral inhibition.

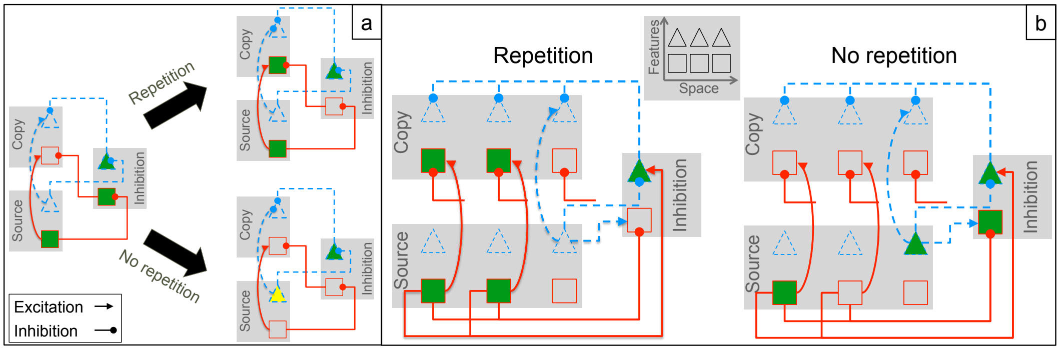

Figure 1: A disinhibition-based sameness-detector for (a) sequentially (b) simultaneously presented identical items. The geometric shapes (squares and triangles) stand for populations of neurons that encode features of the items (e.g., frequency, shape etc.); filled shapes are currently active while empty shapes are currently inactive. (a) Units in the *source layer* (bottom gray box) receive (sensory or other) input. Units in the *copy layer* (top gray box) receive one-to-one excitatory input from the source layer. Critically, units from the *inhibition layer* (right gray box) exert tonic inhibition on the copy layer. (a, left) Upon initial presentation of a feature (represented here as a square), all units in the inhibition layer are active. As a result, excitatory input from the source layer is not propagated to the copy layer. (a, right, top) Feature-specific inhibition from the source layer to the corresponding units in the inhibition layer shuts down the inhibitory input to the copy layer. If the same item is presented again during the time window of reduced inhibition, input from the source layer is propagated to the copy layer. (a, right, bottom) If a new, non-identical item is presented, the source layer cannot drive the copy layer because the corresponding units in the inhibition layer have not been inhibited. Sameness-detection thus proceeds by reading out the copy layer, as only repeated items are propagated to the copy layer. (b) Sameness-detection in simultaneously presented, spatially arranged items. The source layer consists of populations of neurons coding for features (arranged in the y-direction), but these units encode space as

well (arranged in the x-direction). Tonically active inhibitory (inter-)neurons (small gray box on the right) prevent activation in the copy layer (top gray box). Critically, they receive inhibitory input from those units in the source layer that code for the same feature, and excitatory input from units coding for other features. For example, units representing squares in the input layer inhibit all units representing squares in the inhibition layer, and excite all other units. (b, left) If the stimuli consist of two identical items (squares), the combined inhibitory input from the identical items in the source layer shuts down the corresponding units in the inhibition layer, which lets identical items "pass through" to the copy layer (b, right) In contrast, when the stimuli consist of two different items, these singleton features are insufficient to drive the copy population due to inhibition from the inhibition layer.

I simulated this model at various levels of noise; at each noise level, I ran 50 simulations, representing 50 virtual participants. Figure 2 (left) shows that, in the copy layer, activation for repeated features is high, while activation for non-repeated features is low. Repeated items are thus highly discriminable from non-repeated items. This result is robust to the simulated noise level. A simple disinhibition-based circuit can thus act as a sameness-detector that discriminates repeated features from not repeated features.

While the primary goal of this model is to detect when two temporarily adjacent items are identical, whether or not it can detect the sameness of two objects with intervening material depends on the time constants of the disinhibitory effects. If disinhibition is sufficiently long-lasting, the model will also detect the sameness of two non-adjacent items (e.g., of the two *A*'s in the sequence ***ABA***). If so, it would predict that, the further two items are separated (in terms of the amount of intervening time and/or the number of intervening items, which might or might not have separable effects), the harder it should become to detect the sameness of the two items. At least in infants, it might be harder to detect non-adjacent repetitions compared to adjacent repetitions (S. P. Johnson et al., 2009; Kovács & Mehler, 2008, 2009).

That being said, the separation of two items is unlikely to be the only determinant of how it easy it is to detect whether they are the same. For example, in a longer sequence like ***ABCDEDFGA***, the two *A*'s are further apart than the two *D*'s. Still, it might easier to detect the sameness of the two *A*'s than of the two *D*'s despite their greater distance because initial and final items are more salient than medial items (Benavides-Varela & Mehler, 2015; Endress, Scholl, & Mehler, 2005). As a result, the representations of initial items are likely stronger than those of medial items and thus create stronger and longer-lasting disinhibition. However, the goal of the current model is just to show that a simple and ubiquitous mechanism such as disinhibition can serve as the basis of a sameness detector, while more detailed predictions require a biophysically more realistic model.

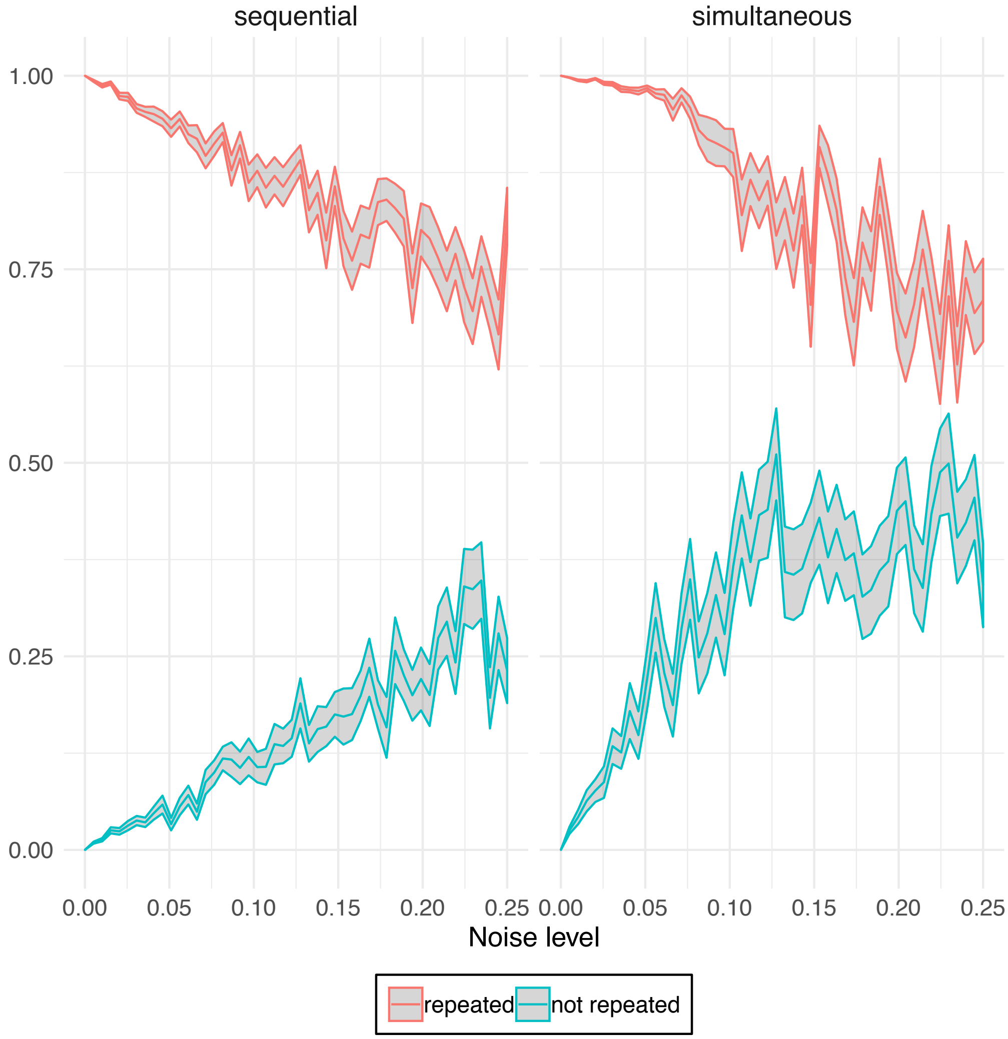

Figure 2: Activation of repeated or non-repeated items in the copy layer. The noise level is the standard deviation of normally distributed noise centered at zero. In each curve, the middle line shows the average activation across 50 simulations, representing 50 participants. The shaded areas represent standard errors from the mean. (Top) Activation in the models shown in Figure 1 that detect either sequentially (Figure 1a) or simultaneously presented (Figure 1b) identical items. (Left). In the *sequential* sameness-detector (Figure 1a), the

activity of repeated items is highly discriminable from that from non-identical items even for high noise levels. (Right). In the *simultaneous* sameness-detector (Figure 1b), the activity of repeated items is highly discriminable from that of non-repeated items even for high noise levels.

Sameness detection for simultaneous stimuli

In its current stage, the model can detect the sameness of sequentially presented stimuli, but not of spatially distributed, simultaneously presented stimuli, simply because space is not represented. Figure 1b shows a version of the model where items are presented simultaneously rather than sequentially. Again, there is a source layer, a copy layer, and an inhibition layer. The model differs from the sequential model in three critical aspects. First, all layers now represent space. In Figure 1b, the vertical axis represents the features as before, while the horizontal axis represents the spatial locations of the items (though space is presumably represented in some topological order in real neuronal populations). This change is necessary so that two simultaneously presented identical objects can be represented.

Second, the connectivity between the source layer and the inhibition layer has been changed. Units in the source layer send (i) inhibitory input to all units in the inhibition layer that code for the same feature across all locations and (ii) excitatory input to all units in the inhibition layer that code for different features; in other words, there is center-surround disinhibition among features. This ensures that, in the copy layer, different-feature input from the source layer stays inhibited, while same-feature input is disinhibited.

Third, the sequential model needs to update the activation of the copy layer before that of the inhibition layer; if the inhibition layer were updated first, a single presentation of a feature would be sufficient to produce disinhibition. In contrast, the simultaneous model needs to update the inhibition layer before the copy layer; if the copy layer were updated first, there would be no disinhibition for identical features.

I simulated this architecture using 50 virtual participants. As shown in Figure 2, identical items are highly discriminable from non-identical items even at high levels of noise. A simple, disinhibition-based circuit can thus detect sameness relations among simultaneously presented identical objects.

A combined model of sameness detection for simultaneous and sequential stimuli

While the main differences between the sequential and the simultaneous circuit are simply due to how stimuli are presented (i.e., spatial representations and lateral inhibition among features could be added to the sequential model, but are not necessary), the different update orders raise the question of whether a combined model can be developed that detects both sequential and simultaneous sameness relations. Practically speaking, sequential and simultaneous presentation might not be as different as they seem. For example, if observers attend simultaneously presented items one after the other (Liu & Becker, 2013; Vogel, Woodman, & Luck, 2006; but see Mance, Becker, & Liu, 2012), we need a sequential model to account for *simultaneous* sameness-detection; conversely, if sequential items are placed in some kind of (short-term) memory before being compared, we need a simultaneous model for sameness-detection in sequentially presented items. As such, a combined sequential/simultaneous model might be neither necessary nor desirable.

Be that as it might, such a combined model is shown in Figure 3.

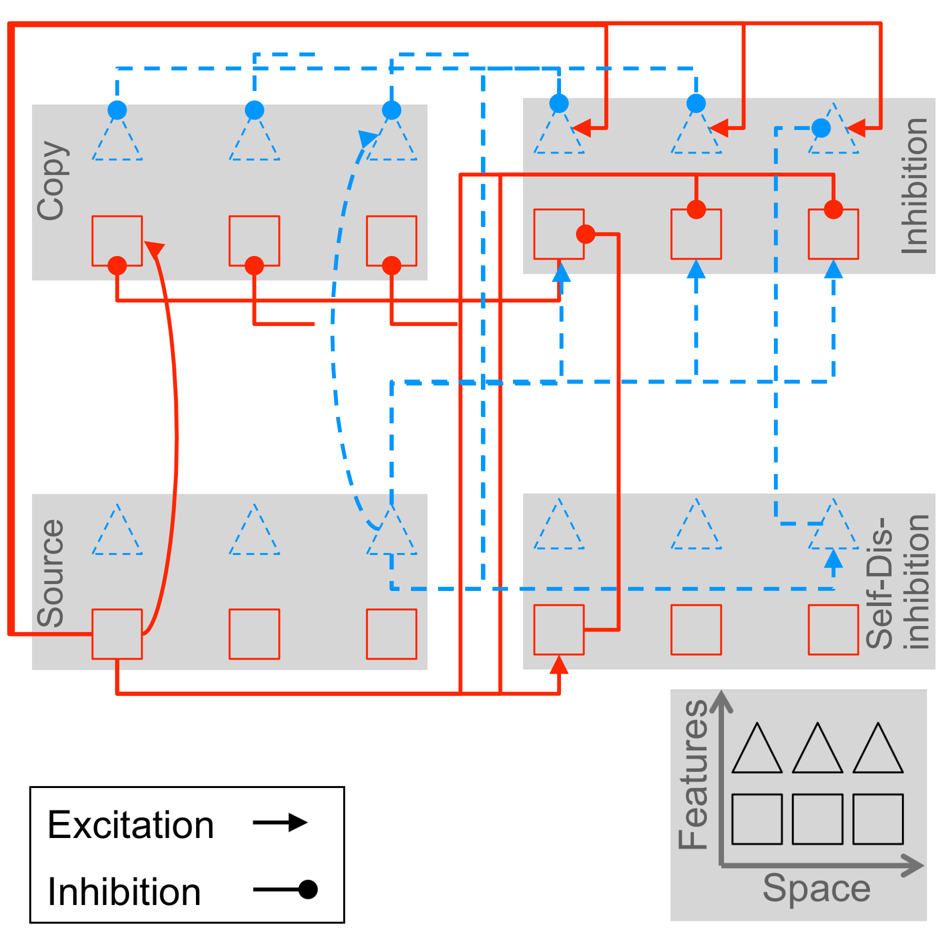

Figure 3: Combined disinhibition-based sameness-detector for both sequential and simultaneous sameness relations. As in the simultaneous circuit from Figure 1b, the source layer (bottom left gray box) consists of populations of neurons coding for features (arranged in the y-direction) and spatial locations (arranged in the x-direction). Tonically active units in the inhibition layer (top right gray box) prevent activation in the copy layer (top left gray box). Units in the inhibition layer receive (i) inhibitory input from the source layer for units coding for the same feature and (ii) excitatory input for units coding for other features, leading to center-surround disinhibition among features and, in the copy layer, to inhibition

for different-feature input and disinhibition for same-feature input. Critically, and in contrast to the simultaneous model from Figure 1b, units in the source layer do not inhibit units in the inhibition layer that code for features at their own spatial location; they disinhibit features only at other locations. To obtain disinhibition at the spatial location of a given unit, a self-inhibition layer (bottom right gray box) was added that receives one-to-one input from the source layer, and that specifically inhibits units in the inhibition layer that code for the same feature *at the same spatial location*. This delays same-feature/same-location disinhibition to prevent a single sequential presentation of a feature from disinhibiting that feature.

This “combined” sameness-detector is similar to the simultaneous sameness-detector in that it comprises a source layer, a copy layer and an inhibition layer, and that the copy layer receives excitatory input from the source layer. However, (dis-)inhibition is organized differently. The copy layer still receives tonic inhibition from those units in the inhibition layer that code for the same feature and spatial position. Further, each feature of the input layer inhibits the corresponding feature in the inhibition layer across spatial positions (i.e., it disinhibits this feature in the copy layer), and excites all other features.

The critical difference is that disinhibition of features at the same location is delayed. To do so, I removed direct connections between the source layer and the inhibition layer that coded for the same feature at the same location (while keeping the center-surround disinhibition at other locations). Instead, I added a *self-disinhibition layer* where each unit (i) receives excitatory input from the corresponding feature and location in the source layer and (ii) sends inhibitory input to all units coding for the same feature (across locations) in the inhibition layer. (While these modifications might seem to some extent *ad-hoc*, as mentioned above, it is not clear if a combined sequential/simultaneous model is necessary or desirable in the first place.)

As shown in Figure 4, identical items were highly discriminable from non-identical items in the simultaneous situation across noise levels; in contrast, in the sequential situation, discriminability suffered as noise increased.

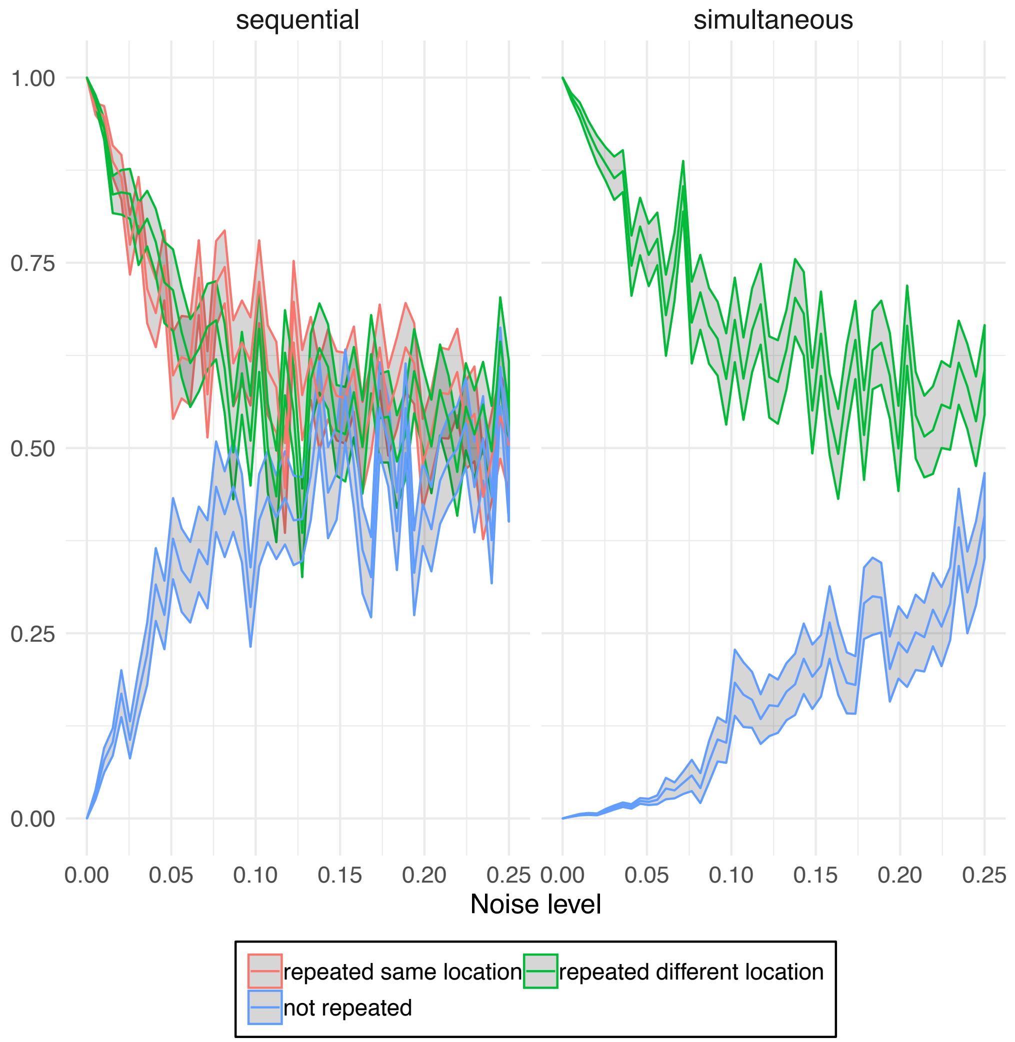

Figure 4: Activation in the copy layer of the combined sequential/simultaneous sameness-detector (Figure 3). (Left) In the combined sequential/simultaneous sameness-detector, repeated features can be repeated either at the same location or at a different location. While activation of (same or different location) repeated items is highly discriminable from activation for non-repeated items for moderate noise levels, discriminability becomes much poorer at high noise levels, when the standard deviation of the noise reaches about 15% of the activation level of active neurons. (Right) The combined sequential/simultaneous

sameness-detector (Figure 3) shows that the activation in the copy layer is highly discriminable between simultaneously repeated items and non-repeated items, even for high noise levels.

Discussion

The current results thus show that a simple and biologically realistic circuit can support a core grammatical computation that is used in more than 80% of the world's languages: grammatical rules based on sameness/difference relationships. In this circuit, non-identical items are filtered out through tonic inhibition as well as center-surround inhibition. In contrast, when identical items are presented sequentially or simultaneously, inhibition is inhibited; this disinhibition of identical items then allows them to be propagated for further processing.

Unlike previous models of sameness-detection (Arena et al., 2013; Carpenter & Grossberg, 1987; Cope et al., 2018; Engel & Wang, 2011; Hasselmo & Wyble, 1997; Johnson, Spencer, Luck, & Schöner, 2009; Ludueña & Gros, 2013; Wen, Ulloa, Husain, Horwitz, & Contreras-Vidal, 2008; see Supplementary Material 1), the model satisfies critical criteria of grammar acquisition: (1) It generalizes to unseen stimuli and (2) does not require any labeled counterexamples for learning, simply because this circuit architecture does not require any learning at all.

Once such a sameness-detector is available, it can be used for building more complex grammatical rules. For example, after exposure to syllable sequences such as *dubaba*, seven-month-olds notice that the last two syllables are identical, and generalize this sameness-relation to new items (Marcus et al., 1999). Critically, they do not only have to detect the sameness relation between the last two syllables, but also have to associate it with the *correct serial* position (Endress et al., 2007; Gervain et al., 2012). Once a sameness-detector is available, it can form associations with representations of sequential positions or other stimuli (Kabdebon & Dehaene-Lambertz, 2019), allowing learners to acquire more complex, composite rules, which is one of the hallmarks of complex cognition (Corballis, 2014; Dehaene, Meyniel, Wacongne, Wang, & Pallier, 2015; Fitch & Martins, 2014; Hauser & Watumull, 2017).

This, in turn, suggests a fundamentally new view on language acquisition. Learners might be equipped with a potentially large number of potentially complex detectors for a variety of rules that act as feature detectors for grammatical rules (Endress, Nespor, et al., 2009). Learning then involves combining these features, potentially through the use of associative mechanisms. This would be consistent with results from formal language theory, where suitable pre-processing (e.g., through feature detectors) can reduce the complexity of the required computational mechanism. For example, a finite state automaton operating on trees can recognize context-free languages (Morgan, 1986) and even humble rules based on sameness relations can be shown to be beyond the reach of even context-free grammars (Culy, 1985; Manaster-Ramer, 1986).

Feature detectors for elementary grammatical rules might thus expand the range of grammars that even simple learning mechanisms (such as associative mechanisms) can learn, which, in turn might make language acquisition feasible using relatively simple computational machinery.

References

Antell, S. E., Caron, A. J., & Myers, R. S. (1985). Perception of relational invariants by newborns. *Developmental Psychology*, *21*(6), 942–948.

Arena, P., Patané, L., Stornanti, V., Termini, P. S., Zäpf, B., & Strauss, R. (2013). Modeling the insect mushroom bodies: Application to a delayed match-to-sample task. *Neural Networks*, *41*, 202–211. https://doi.org/10.1016/j.neunet.2012.11.013

Bagemihl, B. (1995). Language games and related areas. In J. A. Goldsmith (Ed.), *Handbook of Phonological Theory* (1st ed., pp. 697–712). Cambridge, MA: Blackwell.

Bao, Z. (1990). Fanqie Languages and Reduplication. *Linguistic Inquiry*, *21*(3), 317–350.

Barak, O., Sussillo, D., Romo, R., Tsodyks, M., & Abbott, L. F. (2013). From fixed points to chaos: Three models of delayed discrimination. *Progress in Neurobiology*, *103*, 214–222. https://doi.org/10.1016/j.pneurobio.2013.02.002

Benavides-Varela, S., & Mehler, J. (2015). Verbal Positional Memory in 7-Month-Olds. *Child Development*, *86*(1), 209–223. https://doi.org/10.1111/cdev.12291

Benjamin, P. R., Staras, K., & Kemenes, G. (2010). What roles do tonic inhibition and disinhibition play in the control of motor programs? *Frontiers in Behavioral Neuroscience*, *4*, 30. https://doi.org/10.3389/fnbeh.2010.00030

Berent, I., & Shimron, J. (1997). The representation of Hebrew words: Evidence from the obligatory contour principle. *Cognition*, *64*(1), 39–72.

Brooks, V. B. (1986). How does the limbic system assist motor learning? A limbic comparator hypothesis. *Brain, Behavior and Evolution*, *29*(1–2), 29–53.

Carpenter, G. A., & Grossberg, S. (1987). A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics, and Image Processing*, *37*(1), 54–115. http://dx.doi.org/10.1016/S0734-189X(87)80014-2

Chevalier, G., & Deniau, J. M. (1990). Disinhibition as a basic process in the expression of striatal functions. *Trends in Neurosciences*, *13*(7), 277–280.

Chomsky, N. (1959). A Review of B. F. Skinner's Verbal Behavior. *Language*, *35*(1), 26–58.

Cope, A. J., Vasilaki, E., Minors, D., Sabo, C., Marshall, J. A. R., & Barron, A. B. (2018). Abstract concept learning in a simple neural network inspired by the insect brain. *PLoS Computational Biology*, *14*(9), e1006435. https://doi.org/10.1371/journal.pcbi.1006435

Corballis, M. C. (2014). *The Recursive Mind: The Origins of Human Language, Thought, and Civilization*. Princeton, NJ: Princeton University Press.

Culy, C. (1985). The complexity of the vocabulary of Bambara. *Linguistics and Philosophy*, *8*(3), 345–351. https://doi.org/10.1007/BF00630918

Darcy, I., Ramus, F., Christophe, A., Kinzler, K. D., & Dupoux, E. (2009). Phonological knowledge in compensation for native and non-native assimilation. In F. Kügler, C. Féry, & R. van de Vijver (Eds.), *Variation and Gradience in Phonetics and Phonology* (pp. 265–309). Berlin: Mouton De Gruyter.

Dawson, C., & Gerken, L. (2009). From domain-generality to domain-sensitivity: 4-month-olds learn an abstract repetition rule in music that 7-month-olds do not. *Cognition*, *111*(3), 378–382. https://doi.org/10.1016/j.cognition.2009.02.010

de la Mora, D. M., & Toro, J. M. (2013). Rule learning over consonants and vowels in a non-human animal. *Cognition*, *126*(2), 307–312. https://doi.org/10.1016/j.cognition.2012.09.015

Dehaene, S., & Cohen, L. (2007). Cultural recycling of cortical maps. *Neuron*, *56*(2), 384–398. https://doi.org/10.1016/j.neuron.2007.10.004

Dehaene, S., Meyniel, F., Wacongne, C., Wang, L., & Pallier, C. (2015). The Neural Representation of Sequences: From Transition Probabilities to Algebraic Patterns and Linguistic Trees. *Neuron*, *88*(1), 2–19. https://doi.org/10.1016/j.neuron.2015.09.019

Endress, A. D. (2013). Bayesian learning and the psychology of rule induction. *Cognition*, *127*(2), 159–176. https://doi.org/10.1016/j.cognition.2012.11.014

Endress, A. D., Cahill, D., Block, S., Watumull, J., & Hauser, M. D. (2009). Evidence of an evolutionary precursor to human language affixation in a nonhuman primate. *Biol Lett*, *5*(6), 749–751.

Endress, A. D., Dehaene-Lambertz, G., & Mehler, J. (2007). Perceptual constraints and the learnability of simple grammars. *Cognition*, *105*(3), 577–614.

Endress, A. D., Nespor, M., & Mehler, J. (2009). Perceptual and memory constraints on language acquisition. *Trends Cogn Sci*, *13*(8), 348–353.

Endress, A. D., Scholl, B. J., & Mehler, J. (2005). The role of salience in the extraction of algebraic rules. *J Exp Psychol Gen*, *134*(3), 406–419.

Engel, T. A., & Wang, X.-J. (2011). Same or different? A neural circuit mechanism of similarity-based pattern match decision making. *Journal of Neuroscience*, *31*(19), 6982–6996. https://doi.org/10.1523/JNEUROSCI.6150-10.2011

Fabb, N. (2015). *What is Poetry? Language and Memory in the Poems of the World*. https://doi.org/10.1017/CBO9780511736575

Farrant, M., & Nusser, Z. (2005). Variations on an inhibitory theme: Phasic and tonic activation of GABA(A) receptors. *Nature Reviews. Neuroscience*, *6*(3), 215–229. https://doi.org/10.1038/nrn1625

Fee, J., & Ingram, D. (1982). Reduplication as a strategy of phonological development. *J Child Lang*, *9*(1), 41–54.

Ferguson, C. (1964). Baby talk in six languages. *American Anthropologist*, *66*, 103–14.

Fitch, W. T. (2017). Empirical approaches to the study of language evolution. *Psychonomic Bulletin & Review*, *24*(1), 3–33. https://doi.org/10.3758/s13423-017-1236-5

Fitch, W. T., & Martins, M. D. (2014). Hierarchical processing in music, language, and action: Lashley revisited. *Annals of the New York Academy of Sciences*, *1316*, 87–104. https://doi.org/10.1111/nyas.12406

Frisch, S. A., Pierrehumbert, J. B., & Broe, M. B. (2004). Similarity Avoidance and the OCP. *Nat Lang Ling Theory*, *22*(1), 179–228.

Frisch, S. A., & Zawaydeh, B. A. (2001). The psychological reality of OCP-Place in Arabic. *Language*, *77*, 91–106.

Fu, Y., Kaneko, M., Tang, Y., Alvarez-Buylla, A., & Stryker, M. P. (2015). A cortical disinhibitory circuit for enhancing adult plasticity. *ELife*, *4*, e05558. https://doi.org/10.7554/eLife.05558

Fu, Y., Tucciarone, J. M., Espinosa, J. S., Sheng, N., Darcy, D. P., Nicoll, R. A., … Stryker, M. P. (2014). A cortical circuit for gain control by behavioral state. *Cell*, *156*(6), 1139–1152. https://doi.org/10.1016/j.cell.2014.01.050

Gervain, J., Berent, I., & Werker, J. F. (2012). Binding at birth: The newborn brain detects identity relations and sequential position in speech. *Journal of Cognitive Neuroscience*, (Early Access), 1–11.

Gervain, J., Macagno, F., Cogoi, S., Peña, M., & Mehler, J. (2008). The neonate brain detects speech structure. *Proc Natl Acad Sci U S A*, *105*(37), 14222–14227. https://doi.org/10.1073/pnas.0806530105

Goddard, C. A., Mysore, S. P., Bryant, A. S., Huguenard, J. R., & Knudsen, E. I. (2014). Spatially reciprocal inhibition of inhibition within a stimulus selection network in the avian midbrain. *PloS One*, *9*(1), e85865. https://doi.org/10.1371/journal.pone.0085865

Goldin-Meadow, S., & Mylander, C. (1998). Spontaneous sign systems created by deaf children in two cultures. *Nature*, *391*(6664), 279–281. https://doi.org/10.1038/34646

Grill-Spector, K., Henson, R. N., & Martin, A. (2006). Repetition and the brain: Neural models of stimulus-specific effects. *Trends Cogn Sci*, *10*(1), 14–23. https://doi.org/10.1016/j.tics.2005.11.006

Hangya, B., Pi, H.-J., Kvitsiani, D., Ranade, S. P., & Kepecs, A. (2014). From circuit motifs to computations: Mapping the behavioral repertoire of cortical interneurons. *Current Opinion in Neurobiology*, *26*, 117–124. https://doi.org/10.1016/j.conb.2014.01.007

Hardy, H. K., & Montler, T. (1988). Imperfective Gemination in Alabama. *International Journal of American Linguistics*, *54*, 399–475. https://doi.org/10.1086/466094

Hasselmo, M. E., & Wyble, B. P. (1997). Free recall and recognition in a network model of the hippocampus: Simulating effects of scopolamine on human memory function. *Behavioural Brain Research*, *89*(1–2), 1–34.

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, *298*(5598), 1569–1579. https://doi.org/10.1126/science.298.5598.1569

Hauser, M. D., & Glynn, D. (2009). Can free-ranging rhesus monkeys (Macaca mulatta) extract artificially created rules comprised of natural vocalizations? *J Comp Psychol*, *123*(2), 161–167. https://doi.org/10.1037/a0015584

Hauser, M. D., & Watumull, J. (2017). The Universal Generative Faculty: The source of our expressive power in language, mathematics, morality, and music. *Journal of Neurolinguistics*, *43*, 78–94. https://doi.org/10.1016/j.jneuroling.2016.10.005

Hayes, B., & Londe, Z. C. (2006). Stochastic phonological knowledge: The case of Hungarian vowel harmony. *Phonology*, *23*, 59–104.

Heijningen, C. A. A. van, Visser, J. de, Zuidema, W., & Cate, C. ten. (2009). Simple rules can explain discrimination of putative recursive syntactic structures by a songbird species. *Proc Natl Acad Sci U S A*, *106*(48), 20538–20543. https://doi.org/10.1073/pnas.0908113106

Hupé, J.-M. (2017). Comment on “Ducklings imprint on the relational concept of ‘same or different.’” *Science*, *355*(6327), 806. https://doi.org/10.1126/science.aah6047

Johnson, J. S., Spencer, J. P., Luck, S. J., & Schöner, G. (2009). A dynamic neural field model of visual working memory and change detection. *Psychological Science*, *20*(5), 568–577. https://doi.org/10.1111/j.1467-9280.2009.02329.x

Johnson, S. P., Fernandas, K. J., Frank, M. C., Kirkham, N., Marcus, G., Rabagliati, H., & Slemmer, J. A. (2009). Abstract Rule Learning for Visual Sequences in 8- and 11-Month-Olds. *Infancy*, *14*(1), 2–18. https://doi.org/10.1080/15250000802569611

Jovanic, T., Schneider-Mizell, C. M., Shao, M., Masson, J.-B., Denisov, G., Fetter, R. D., … Zlatic, M. (2016). Competitive Disinhibition Mediates Behavioral Choice and Sequences in Drosophila. *Cell*, *167*(3), 858-870.e19. https://doi.org/10.1016/j.cell.2016.09.009

Kabdebon, C., & Dehaene-Lambertz, G. (2019). Symbolic labeling in 5-month-old human infants. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(12), 5805–5810. https://doi.org/10.1073/pnas.1809144116

Kovács, Á. M., & Mehler, J. (2008). *Regularity learning in 7-month-old infants under `noisy’ conditions: Adjacent repetitions vs. Non-adjacent repetitions*.

Kovács, Á. M., & Mehler, J. (2009). Flexible learning of multiple speech structures in bilingual infants. *Science*, *325*(5940), 611–612. https://doi.org/10.1126/science.1173947

Koyama, M., Minale, F., Shum, J., Nishimura, N., Schaffer, C. B., & Fetcho, J. R. (2016). A circuit motif in the zebrafish hindbrain for a two alternative behavioral choice to turn left or right. *ELife*, *5*. https://doi.org/10.7554/eLife.16808

Koyama, M., & Pujala, A. (2018). Mutual inhibition of lateral inhibition: A network motif for an elementary computation in the brain. *Current Opinion in Neurobiology*, *49*, 69–74. https://doi.org/10.1016/j.conb.2017.12.019

Kumaran, D., & Maguire, E. A. (2007). Which computational mechanisms operate in the hippocampus during novelty detection? *Hippocampus*, *17*(9), 735–748. https://doi.org/10.1002/hipo.20326

Kusunoki, M., Sigala, N., Nili, H., Gaffan, D., & Duncan, J. (2010). Target detection by opponent coding in monkey prefrontal cortex. *Journal of Cognitive Neuroscience*, *22*(4), 751–760. https://doi.org/10.1162/jocn.2009.21216

Langbein, J., & Puppe, B. (2017). Comment on “Ducklings imprint on the relational concept of ‘same or different.’” *Science*, *355*(6327), 806. https://doi.org/10.1126/science.aai7431

Lee, S., Kruglikov, I., Huang, Z. J., Fishell, G., & Rudy, B. (2013). A disinhibitory circuit mediates motor integration in the somatosensory cortex. *Nature Neuroscience*, *16*(11), 1662–1670.

https://doi.org/10.1038/nn.3544

Letzkus, J. J., Wolff, S. B. E., & Lüthi, A. (2015). Disinhibition, a Circuit Mechanism for Associative Learning and Memory. *Neuron*, *88*(2), 264–276. https://doi.org/10.1016/j.neuron.2015.09.024

Letzkus, J. J., Wolff, S. B. E., Meyer, E. M. M., Tovote, P., Courtin, J., Herry, C., & Lüthi, A. (2011). A disinhibitory microcircuit for associative fear learning in the auditory cortex. *Nature*, *480*(7377), 331–335. https://doi.org/10.1038/nature10674

Liu, T., & Becker, M. W. (2013). Serial Consolidation of Orientation Information Into Visual Short-Term Memory. *Psychol Sci*. https://doi.org/10.1177/0956797612464381

Ludueña, G. A., & Gros, C. (2013). A self-organized neural comparator. *Neural Computation*, *25*(4), 1006–1028. https://doi.org/10.1162/NECO_a_00424

Machens, C. K., Romo, R., & Brody, C. D. (2005). Flexible control of mutual inhibition: A neural model of two-interval discrimination. *Science*, *307*(5712), 1121–1124. https://doi.org/10.1126/science.1104171

Manaster-Ramer, A. (1986). Copying in Natural Languages, Context-freeness, and Queue Grammars. *Proceedings of the 24th Annual Meeting on Association for Computational Linguistics*, 85–89. https://doi.org/10.3115/981131.981145

Mance, I., Becker, M. W., & Liu, T. (2012). Parallel consolidation of simple features into visual short-term memory. *J Exp Psychol Hum Percept Perform*, *38*(2), 429–438. https://doi.org/10.1037/a0023925

Marcus, G. F., Fernandes, K. J., & Johnson, S. P. (2007). Infant rule learning facilitated by speech. *Psychol Sci*, *18*(5), 387–391. https://doi.org/10.1111/j.1467-9280.2007.01910.x

Marcus, G. F., Vijayan, S., Rao, S. B., & Vishton, P. (1999). Rule learning by seven-month-old infants. *Science*, *283*(5398), 77–80.

Marlin, B. J., Mitre, M., D'amour, J. A., Chao, M. V., & Froemke, R. C. (2015). Oxytocin enables maternal behaviour by balancing cortical inhibition. *Nature*, *520*(7548), 499–504. https://doi.org/10.1038/nature14402

Martinho, A., & Kacelnik, A. (2016). Ducklings imprint on the relational concept of "same or different." *Science*, *353*(6296), 286–288. https://doi.org/10.1126/science.aaf4247

McCarthy, J. J. (1986). OCP Effects: Gemination and Antigemination. *Linguistic Inquiry*, *17*(2), 207–263.

Medina, T. N., Snedeker, J., Trueswell, J. C., & Gleitman, L. R. (2011). How words can and cannot be learned by observation. *Proc Natl Acad Sci U S A*, *108*(22), 9014–9019. https://doi.org/10.1073/pnas.1105040108

Miller, P., & Wang, X.-J. (2006). Inhibitory control by an integral feedback signal in prefrontal cortex: A model of discrimination between sequential stimuli. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(1), 201–206. https://doi.org/10.1073/pnas.0508072103

Mitterer, H., & Blomert, L. (2003). Coping with phonological assimilation in speech perception: Evidence for early compensation. *Perception & Psychophysics*, *65*(6), 956–969.

Moravcsik, E. (1978). Reduplicative constructions. In J. H. Greenberg (Ed.), *Universals of human language: Word structure* (Vol. 3, pp. 297–334).

Stanford, CA: Stanford University Press.

Morgan, J. L. (1986). *From Simple Input to Complex Grammar*. Cambridge, MA: MIT Press.

Murphy, R. A., Mondragon, E., & Murphy, V. A. (2008). Rule Learning by Rats. *Science*, *319*(5871), 1849–1851. https://doi.org/10.1126/science.1151564

Mysore, S. P., & Knudsen, E. I. (2012). Reciprocal inhibition of inhibition: A circuit motif for flexible categorization in stimulus selection. *Neuron*, *73*(1), 193–205. https://doi.org/10.1016/j.neuron.2011.10.037

Neiworth, J. J. (2013). Chasing sounds. *Behav Processes*, *93*, 111–115. https://doi.org/10.1016/j.beproc.2012.11.009

Owen, S. F., Tuncdemir, S. N., Bader, P. L., Tirko, N. N., Fishell, G., & Tsien, R. W. (2013). Oxytocin enhances hippocampal spike transmission by modulating fast-spiking interneurons. *Nature*, *500*(7463), 458–462. https://doi.org/10.1038/nature12330

Pepperberg, I. M. (1987). Acquisition of the same/different concept by an African Grey parrot (Psittacus erithacus): Learning with respect to categories of color, shape, and material. *Anim Learn Behav*, *15*, 421–32.

Pfeffer, C. K., Xue, M., He, M., Huang, Z. J., & Scanziani, M. (2013). Inhibition of inhibition in visual cortex: The logic of connections between molecularly distinct interneurons. *Nature Neuroscience*, *16*(8), 1068–1076. https://doi.org/10.1038/nn.3446

Pi, H.-J., Hangya, B., Kvitsiani, D., Sanders, J. I., Huang, Z. J., & Kepecs, A. (2013). Cortical interneurons that specialize in disinhibitory control. *Nature*, *503*(7477), 521–524. https://doi.org/10.1038/nature12676

Pinker, S. (1984). *Language Learnability and Language Development*. Cambridge, MA: MIT Press.

Pinker, S., & Jackendoff, R. (2005). The faculty of language: What's special about it? *Cognition*, *95*(2), 201–236. https://doi.org/10.1016/j.cognition.2004.08.004

Rabagliati, H., Ferguson, B., & Lew-Williams, C. (2019). The profile of abstract rule learning in infancy: Meta-analytic and experimental evidence. *Developmental Science*, *22*(1), e12704. https://doi.org/10.1111/desc.12704

Rose, S., & Walker, R. (2011). Harmony Systems. In J. Goldsmith, J. Riggle, & A. C. L. Yu (Eds.), *The Handbook of Phonological Theory* (2nd ed., pp. 240–290). https://doi.org/10.1002/9781444343069.ch8

Rubino, C. (2013). Reduplication. In M. S. Dryer & M. Haspelmath (Eds.), *The World Atlas of Language Structures Online*. Retrieved from http://wals.info/chapter/27

Saffran, J. R., Pollak, S. D., Seibel, R. L., & Shkolnik, A. (2007). Dog is a dog is a dog: Infant rule learning is not specific to language. *Cognition*, *105*(3), 669–680. https://doi.org/10.1016/j.cognition.2006.11.004

Schwartz, R. G., Leonard, L. B., Wilcox, M. J., & Folger, M. K. (1980). Again and again: Reduplication in child phonology. *Journal of Child Language*, *7*(1), 75–87. https://doi.org/10.1017/S0305000900007030

Semyanov, A., Walker, M. C., Kullmann, D. M., & Silver, R. A. (2004). Tonically active GABA A receptors: Modulating gain and maintaining the tone. *Trends in Neurosciences*, *27*(5), 262–269. https://doi.org/10.1016/j.tins.2004.03.005

Senghas, A., Kita, S., & Ozyürek, A. (2004). Children creating core properties of language: Evidence from an emerging sign language in Nicaragua. *Science*, *305*(5691), 1779–1782. https://doi.org/10.1126/science.1100199

Smirnova, A., Zorina, Z., Obozova, T., & Wasserman, E. (2015). Crows spontaneously exhibit analogical reasoning. *Current Biology*, *25*(2), 256–260. https://doi.org/10.1016/j.cub.2014.11.063

Suomi, K., McQueen, J. M., & Cutler, A. (1997). Vowel Harmony and Speech Segmentation in Finnish. *J Mem Lang*, *36*(3), 422–444.

van Der Velde, F., & de Kamps, M. (2001). From knowing what to knowing where: Modeling object-based attention with feedback disinhibition of activation. *Journal of Cognitive Neuroscience*, *13*(4), 479–491.

Versace, E., Spierings, M. J., Caffini, M., Ten Cate, C., & Vallortigara, G. (2017). Spontaneous generalization of abstract multimodal patterns in young domestic chicks. *Animal Cognition*, *20*(3), 521–529. https://doi.org/10.1007/s10071-017-1079-5

Vogel, E. K., Woodman, G. F., & Luck, S. J. (2006). The time course of consolidation in visual working memory. *J Exp Psychol Hum Percept Perform*, *32*(6), 1436–1451. https://doi.org/10.1037/0096-1523.32.6.1436

Vroomen, J., Tuomainen, J., & de Gelder, B. (1998). The Roles of Word Stress and Vowel Harmony in Speech Segmentation. *J Mem Lang*, *38*(2), 133–149.

Wen, S., Ulloa, A., Husain, F., Horwitz, B., & Contreras-Vidal, J. L. (2008). Simulated neural dynamics of decision-making in an auditory delayed match-to-sample task. *Biological Cybernetics*, *99*(1), 15–27.

https://doi.org/10.1007/s00422-008-0234-0

Xu, H., Jeong, H.-Y., Tremblay, R., & Rudy, B. (2013). Neocortical somatostatin-expressing GABAergic interneurons disinhibit the thalamorecipient layer 4. *Neuron*, *77*(1), 155–167. https://doi.org/10.1016/j.neuron.2012.11.004

Yang, C. (2013). Ontogeny and phylogeny of language. *Proc Natl Acad Sci U S A*, *110*(16), 6324–6327. https://doi.org/10.1073/pnas.1216803110

Yip, M. (1982). Reduplication and C-V Skeleta in Chinese Secret Languages. *Linguistic Inquiry*, *13*(4), 637–661.

Yip, M. (1988). The Obligatory Contour Principle and Phonological Rules: A Loss of Identity. *Linguistic Inquiry*, *19*(1), 65–100.

Zhang, S., Xu, M., Kamigaki, T., Hoang Do, J. P., Chang, W.-C., Jenvay, S., … Dan, Y. (2014). Selective attention. Long-range and local circuits for top-down modulation of visual cortex processing. *Science*, *345*(6197), 660–665. https://doi.org/10.1126/science.1254126

Zhao, W., Zhou, P., Gong, C., Ouyang, Z., Wang, J., Zheng, N., & Gong, Z. (2019). A disinhibitory mechanism biases Drosophila innate light preference. *Nature Communications*, *10*(1), 124. https://doi.org/10.1038/s41467-018-07929-w

Appendix A: Model equations

A.1 Sequential model

The feature *f* is encoded in the source layer, the inhibition layer and the copy layer; the corresponding activations, are $S_f\,(t)$ for a unit encoding feature *f* in the source layer, $I_f\,(t)$ for such a unit in the inhibition layer and (3) $C_f\,(t)$ such for a unit in the copy layer. $E_f\,(t)$ is the external input, $\mathcal{N}(\mu,\sigma)$ is a random value drawn from a normal distribution with mean μ and standard deviation σ.

Before stimulation, the activation in the source layer and in the copy layer are initialized to zero (plus noise), while the activation in the inhibition layer is initialized to some value a_I (here arbitrarily set to 1):

$$\begin{array}{l} S_f(t=0) \sim \mathcal{N}(0,\sigma_{activation}) \\ C_f(t=0) \sim \mathcal{N}(0,\sigma_{activation}) \\ I_f(t=0) \sim \mathcal{N}(a_I,\sigma_{activation}) \end{array} \qquad (1)$$

The connection weights between units in the different layers are indicated by w: $w^{I,S}$ from the source layer to the inhibition layer, $w^{C,S}$ from the source layer to the copy layer and $w^{C,I}$ from the inhibition layer to the copy layer. A connection between a source layer unit coding for feature *f* and a copy layer unit coding for feature *f'* is indicated by $w^{C,S}_{f\,f}$. The weights are given as follows:

$$
\begin{aligned}
w_{f',f}^{C,S} &\sim \begin{cases} \mathcal{N}(1,\sigma_{weight}) & f = f' \\ 0 & f \neq f' \end{cases} \\
w_{f',f}^{C,I} &\sim \begin{cases} \mathcal{N}(-1,\sigma_{weight}) & f = f' \\ 0 & f \neq f' \end{cases} \\
w_{f',f}^{I,S} &\sim \begin{cases} \mathcal{N}(-1,\sigma_{weight}) & f = f' \\ 0 & f \neq f' \end{cases}
\end{aligned}
\tag{2}
$$

At each time step, the activations in the different layers are then updated as follows; as mentioned in the main text, the update order is critical.

$$
\begin{aligned}
&S_f(t) = E_f(t) + \mathcal{N}(0,\sigma_{activation}) \\
&C_f(t) = w_f^{C,S} S_f(t) + w_f^{C,I} I_f(t) + \mathcal{N}(0,\sigma_{activation}) \\
&I_f(t) = \mathcal{N}(a_I,\sigma_{activation}) + w_f^{I,S} S_f(t)
\end{aligned}
\tag{3}
$$

At the end of each update cycle, the activations are curtailed to be between zero and one.

A.2. Simultaneous model

In the simultaneous model, units represent both features and spatial locations. $S_{f,l}$ (t) is thus the activation of a unit in the source layer that encodes feature f at location l, $I_{f,l}$ (t) is the corresponding activation in the inhibition layer and (3) $C_{f,l}$ (t) is the corresponding activation in the copy layer. $E_{f,l}$ (t) is the external input.

Before stimulation, the activation in the source layer and in the copy layer are initialized to zero (plus noise), while the activation in the inhibition layer is initialized to some value a_I (here arbitrarily set to 1):

$$\begin{aligned} S_{f,l}(t=0) &\sim \mathcal{N}(0,\sigma_{activation}) \\ C_{f,l}(t=0) &\sim \mathcal{N}(0,\sigma_{activation}) \\ I_{f,l}(t=0) &\sim \mathcal{N}(a_I,\sigma_{activation}) \end{aligned} \tag{4}$$

Connection weights now carry indices for both features and spatial locations. For example, a connection between a source layer unit coding for feature *f* at location *l* and a copy layer unit coding for feature *f*' at location *l*' is indicated by $w^{\mathrm{C,S}}_{f\,f,l',l}$. The weights are given as follows:

$$\begin{aligned} w^{C,S}_{f',f,l',l} &\sim \begin{cases} \mathcal{N}(1,\sigma_{weight}) & f=f', l=l' \\ 0 & \text{otherwise} \end{cases} \\ w^{C,I}_{f',f,l',l} &\sim \begin{cases} \mathcal{N}(-1,\sigma_{weight}) & f=f' \\ 0 & f\neq f' \end{cases} \\ w^{I,S}_{f',f,l',l} &\sim \begin{cases} \mathcal{N}(-1,\sigma_{weight}) & f=f' \\ \mathcal{N}(1,\sigma_{weight}) & f\neq f' \end{cases} \end{aligned} \tag{5}$$

At each time step, the activations in the different layers are then updated as follows; as mentioned in the main text, the update order is critical.

$$\begin{aligned} S_{f,l}(t) &= E_{f,l}(t) + N(0,\sigma_{activation}) \\ I_{f,l}(t) &= N(a_I,\sigma_{activation}) + \sum_{f_S,l_S} w^{I,S}_{f,l,f_S,l_s} S_{f_S,l_S}(t) \\ C_{f,l}(t) &= \sum_{f_S,l_S} w^{C,S}_{f,l,f_S,l_s} S_{f_S,l_S}(t) + \sum_{f_I,l_I} w^{C,I}_{f,l,f_I,l_I} I_{f_I,l_I}(t) + N(0,\sigma_{activation}) \end{aligned} \tag{6}$$

At the end of each update cycle, the activations are curtailed to be between zero and one.

A.3. Combined model

The combined sequential/simultaneous model is similar to the simultaneous model in that it comprises a source layer, a copy layer and an inhibition layer and that the copy layer receives excitatory input from the source layer as well as tonic inhibition from those units in the inhibition layer that code for the same feature and spatial position. Further, each feature of the input layer inhibits the corresponding feature in the inhibition layer across spatial positions and excites all other features. The critical difference between the simultaneous and the combined model is that there are no connections between the source layer and the inhibition layer that code for the same feature *at the same location* (while disinhibition occurs for other locations), and that same-location disinhibition of features proceeds through a *self-disinhibition layer* where each unit (1) receives excitatory input from the corresponding feature and location in the source layer (2) sends inhibitory input to all units coding for the same feature (across locations) in the inhibition layer.

The symbols for the activation in the source, inhibition and copy layers are the same as in the simultaneous model; activation in the self-disinhibition layer for a unit coding for feature f at location l is designated as $D_{f,l}\,(t)$ and is initialized using random values around zero.

The symbols for the connection weights are similar to those in the simultaneous model, but the weights reflect the changes above:

$$
\begin{gathered}
w^{C,S}_{f',f,l',l} \sim \begin{cases} \mathcal{N}(1,\sigma_{weight}) & f=f', l=l' \\ 0 & \text{otherwise} \end{cases} \\
w^{D,S}_{f',f,l',l} \sim \begin{cases} \mathcal{N}(1,\sigma_{weight}) & f=f', l=l' \\ 0 & \text{otherwise} \end{cases} \\
w^{C,I}_{f',f,l',l} \sim \begin{cases} \mathcal{N}(-1,\sigma_{weight}) & f=f' \\ 0 & f\neq f' \end{cases} \\
w^{I,S}_{f',f,l',l} \sim \begin{cases} \mathcal{N}(-1,\sigma_{weight}) & f=f', l\neq l' \\ 0 & f=f', l=l' \\ \mathcal{N}(1,\sigma_{weight}) & f\neq f' \end{cases} \\
w^{I,D}_{f',f,l',l} \sim \begin{cases} \mathcal{N}(-1,\sigma_{weight}) & f=f' \\ 0 & f\neq f' \end{cases}
\end{gathered} \tag{7}
$$

At each time step, the activations in the different layers are then updated as follows; again, the update order is critical.

$$
\begin{aligned}
S_{f,l}(t) &= E_{f,l}(t) + N(0,\sigma_{activation}) \\
I_{f,l}(t) &= N(a_I,\sigma_{activation}) + \sum_{f_S,l_S} w^{I,S}_{f,l,f_S,l_s} S_{f_S,l_S}(t) + \sum_{f_D,l_D} w^{I,D}_{f,l,f_D,l_D} D_{f_D,l_D}(t) \\
C_{f,l}(t) &= \sum_{f_S,l_S} w^{C,S}_{f,l,f_S,l_s} S_{f_S,l_S}(t) + \sum_{f_I,l_I} w^{C,I}_{f,l,f_I,l_I} I_{f_I,l_I}(t) + N(0,\sigma_{activation}) \\
D_{f,l}(t) &= \sum_{f_S,l_S} w^{D,S}_{f,l,f_S,l_s} S_{f_S,l_S}(t) + N(0,\sigma_{activation})
\end{aligned} \tag{8}
$$

At the end of each update cycle, the activations are curtailed to be between zero and one.