Language Learning and Development

ISSN: 1547-5441 (Print) 1547-3341 (Online) Journal homepage: www.tandfonline.com/journals/hlld20

Patterns Bit by Bit. An Entropy Model for Rule Induction

Silvia Radulescu, Frank Wijnen & Sergey Avrutin

To cite this article: Silvia Radulescu, Frank Wijnen & Sergey Avrutin (2020) Patterns Bit by Bit. An Entropy Model for Rule Induction, Language Learning and Development, 16:2, 109-140, DOI: 10.1080/15475441.2019.1695620

To link to this article: https://doi.org/10.1080/15475441.2019.1695620

LANGUAGE LEARNING AND DEVELOPMENT
2020, VOL. 16, NO. 2, 109–140
https://doi.org/10.1080/15475441.2019.1695620

Routledge
Taylor & Francis Group

Patterns Bit by Bit. An Entropy Model for Rule Induction

Silvia Radulescu, Frank Wijnen, and Sergey Avrutin

UiL-OTS, Utrecht University, Utrecht, The Netherlands

ABSTRACT

From limited evidence, children track the regularities of their language impressively fast and they infer generalized rules that apply to novel instances. This study investigated what drives the inductive leap from memorizing specific items and statistical regularities to extracting abstract rules. We propose an innovative entropy model that offers one consistent information-theoretic account for both learning the regularities in the input and generalizing to new input. The model predicts that rule induction is an encoding mechanism gradually driven as a natural automatic reaction by the brain's sensitivity to the input complexity (entropy) interacting with the finite encoding power of the human brain (channel capacity). In two artificial grammar experiments with adults we probed the effect of input complexity on rule induction. Results showed that as the input becomes more complex, the tendency to infer abstract rules increases gradually.

Introduction

The induction problem for language acquisition

When acquiring the rules of their language from a limited number of examples, children not only learn how particular linguistic items (sounds, words, etc.) are associated, but they also infer generalized rules that apply productively to novel instances. This inductive leap is a powerful phenomenon because it enables learners to create and understand an infinite number of sentences. From memorizing sequences like *Dad walked slowly* and *Mom talked nicely*, to learning generalizations of the type "add – *ed*" to express a past action, and to generalizing to abstract categories (Noun, Verb, Adverb), and inducing a general rule that the sequence Noun-Verb-Adverb is well-formed, learners take a qualitative step from encoding exemplars to forming abstract categories and acquiring relations between them. This paper addresses this qualitative step from items to categories.

Following previous proposals in the literature (Gómez & Gerken, 2000), we will distinguish between two types of rule induction: *item-bound generalizations* and *category-based generalizations*. An *item-bound generalization* is a relation between perceptual features[1] of items, e.g. a relation based on physical identity, like *ba-ba* (*ba* follows *ba*), or "add – *ed*". *Category-based generalization* operates beyond the physical items; it abstracts over categories (variables), e.g. *Y* follows *X*, where *Y* and *X* are variables taking different values. In

CONTACT Silvia Radulescu S.Radulescu@uu.nl UiL-OTS, Utrecht University, Utrecht, The Netherlands

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/hlld.

The entropy model and parts of the findings were presented at the following conferences:
Architectures and Mechanisms for Language Processing: Edinburgh, UK, 2014
Architectures and Mechanisms for Language Processing: Valletta, Malta, 2015
CUNY Conference on Human Sentence Processing - University of Florida, Gainesville, USA, 2016
Statistical Learning Conference - Bilbao, Spain, 2017
Architectures and Mechanisms for Language Processing - Lancaster, UK, 2017
Architectures and Mechanisms for Language Processing - Berlin, Germany, 2018
The first version of the entropy model and the results of the first experiment were described in Silvia Radulescu's 2014 Master's Thesis, which is available from the Utrecht University Online Repository: https://dspace.library.uu.nl/handle/1874/294595

[1]Perceptual features are any physical characteristics specific to the respective perception modality (auditory, visual etc.).

natural language, the grammatical generalization that a sentence consists of a Noun-Verb-Noun sequence is based on recognizing an identity relation over the abstract linguistic category of noun (which can be construed as a variable that takes specific nouns as values). *Category-based generalization* is a very powerful phenomenon, because it enables processing a potentially infinite number of sentences, making it crucial to linguistic productivity. Thus, a fundamental mechanism that needs to be investigated to thoroughly understand language acquisition is how learners converge on these higher-order *category-based generalizations*.

Statistical learning vs. algebraic rules

An ongoing debate in psycholinguistics revolves around the learning mechanisms underlying *item-bound* and *category-based generalizations*. Studies focusing on *item-bound generalization* argue that the learning mechanism at stake is a lower-level item-bound mechanism that relies on memorization of the specific items (i.e. their physical features), and on the *statistical relations* between them. For example, it was shown that children detect patterns of specific auditory/visual items, e.g. phonotactic information (Chambers, Onishi, & Fisher, 2003), and word boundaries (Aslin, Saffran, & Newport, 1998; Saffran, Aslin, & Newport, 1996), by *statistical learning*. As defined in Saffran et al. (1996), *statistical learning* denotes statistical computation about probabilistic distributions of items, such as transitional probabilities (e.g. the probability that a certain item occurs after another). While such basic statistical computations were shown to suffice for *item-bound generalizations*, some researchers argued (Endress & Bonatti, 2007; Marcus, Vijayan, Rao, & Vishton, 1999) that this mechanism alone cannot account for generalizing beyond specific items. Marcus et al. (1999) showed that 7-month olds recognize the AAB structure underlying strings such as "*leledi*", "*kokoba*", as they were able to discriminate new strings, consisting of novel syllables, with the same AAB structure, from novel strings with a different structure (e.g. ABA). Marcus et al. argue that infants are equipped with an abstract symbolic ("algebraic") system that comprises variables and relations between these variables. Thus, they proposed that children possess two *separate* learning mechanisms, which are different in nature: *statistical learning* for tracking co-occurrence probabilities of specific items, and an *abstract rule learning mechanism* that creates and operates on variables. Although an algebraic system might enable generalizing to novel input, the authors do not explain how learners tune into such algebraic rules, and what factors facilitate or impede this process.

In contrast to the proposition put forth by Marcus et al. and Endress and Bonatti, that statistical learning and abstract rule learning are separate and distinct mechanisms, Aslin & Newport (2012) argued that statistical learning accounts for learning both statistical regularities of specific items and abstract rules that apply to novel instances. Recent computational models suggest that learners might be "adding generalization to statistical learning" when inducing phonotactic knowledge (Adriaans & Kager, 2010), and that neither a "pure statistics" position, nor a "rule-only position" would suffice for explaining the phenomenon of generalization, but rather an interaction between the two mechanisms in which "statistical inference is performed over rule-based representations" (Frank & Tenenbaum, 2011).

In the studies summarized above, the terminology was used to refer to both the two types of encoding (statistical regularities vs. abstract rules), and to the underlying learning mechanisms, i.e. *statistical learning* vs. *abstract rule learning*. But we posit that the processes (i.e. learning mechanisms) should be disentangled from their results (i.e. forms of encoding). Drawing this distinction allows for more specific questions to be formulated:

(1) Are these forms of encoding outcomes of two separate mechanisms, with *statistical learning* underlying *item-bound generalizations*, and *abstract rule learning* accounting for the higher-order *category-based generalizations*?
(2) Or, are these forms of encoding two different outcomes of the same mechanism?

(a) If they are outcomes of the same mechanism, are the two types of generalizations stages of a phased mechanism that *gradually* transitions from a lower-level item-bound generalization to a higher-order abstract one?
(b) Or is it a mechanism that switches *abruptly* from one form of encoding to the other?

(3) What triggers the change in form of encoding, be it a gradual transition from *item-bound* into *category-based generalization*, or a sudden leap from one form of encoding to the other one?

Rule induction in infants

Gerken (2006) took a step toward understanding the relation between the two forms of encoding and the triggering factors, by showing that the nature of generalization that learners form depends crucially on the statistical properties of the input. Gerken (2006) modified the design used by Marcus et al. (1999) and reconsidered their argument. She asked whether 9-month-olds presented with two different subsets of the strings used by Marcus et al. (1999) would make the same generalization. To answer this question, she presented one group of infants with four AAB strings ending in different syllables (*je/li/di/we*) and another group with four AAB strings ending only in *di*. Gerken argues that infants in the second group had two equally plausible generalizations at hand: the broader AAB rule (a *category-based generalization*, according to our terminology), and the narrower "ends in *di*" generalization (an *item-bound generalization*). The results showed that the second group only generalized to novel AAB strings that ended in *di* (so, not *ko_ko_ba, etc.*), while the first group made the broader generalization to all AAB strings. Gerken surmises that (1) the learners in the AAdi condition did not see evidence that strings could end in any other syllable, and, therefore, (2) they posited the only (minimal) rule that reliably generated the set of AAB strings ending in the same syllable *di*, namely, the "ends in *di*" rule. The implication of this study is that generalization is apparently graded, and that the degree to which learners generalize depends on the variability of the input.

However, this account is incomplete. Gerken argues that only the second group had two equally plausible generalizations at hand, but we think that, formally, both groups were presented with input that evidenced both a narrower generalization ("ends in *je/li/di/we*" in the first group; and "ends in *di*" in the second one), and a broader *AAB* generalization, but in one case the narrower *item-bound generalization* was made, and in the other case the broader *category-based generalization*. In fact, both groups were presented with input that provided no direct evidence that strings could also end in a new syllable (i.e. none of the strings in the input ended in *ba*). However, learners in the first group accepted a new *AAB* string ending in *ba* (instead of sticking to the narrower "ends in *je/li/di/we*" generalization), while the second group stuck to "ends in *di*". As the authors argue that the second group made the narrower generalization "ends in *di*" because there is no direct evidence from the input that a string could end in a new syllable (e.g. *ba)*, then the other group should be expected to do the same, i.e. stick to the narrower generalization "ends in *je/li/di/we*", because their input also showed no direct evidence that a string could end in a new syllable (e.g. *ba*). Hence it is still not clear from these results what exactly triggered a broader *category-based generalization* and what kind of evidence is needed to support it. Also, if input variability is a factor, as argued by Gerken, how much variability is needed to trigger a *category-based generalization*?

A subsequent study by Gerken (2010) may help finding answers. In this study, she exposed 9-month-olds to the same "ends in *di*" condition as in Gerken (2006), but – crucially – added three strings ending in "*je/we/li*" at the end of the familiarization. The participants subsequently made the broader AAB generalization. The author hypothesizes that the factor driving generalization is not the mere number of examples, but the logical structure of the input. She proposes that infants entertain incremental learning models (by updating their hypothesis in real time), and that they use rational decision criteria, in a process that resembles Bayesian learning. But we ask: would they make a broader generalization also if these 3 "divergent" strings were presented at the beginning of the 2-minute familiarization? Would infants not "forget" those 3 strings, and rather update their model based on the more strongly evidenced and recent "end in *di*" input? As Gerken (2010) did not include this control condition, the study cannot decisively show that infants are

incremental and "rational" learners, as there is no online measure or intermediate checkpoint into their models before and after each batch of stimuli. Nonetheless, it clearly shows that little evidence and variability is needed for them to move to a broader generalization. However, surprisingly, the results of Gerken, Dawson, Chatila, and Tenenbaum (2015) suggest that variability is not needed. An input consisting of a single item ("*ledi*") is enough for 9-month-olds to make a broader generalization (AAB), if there is a surprising repetition pattern ("*lele*") which is very rare in their prior language model. However, when the single item was ("*lelezhi*") – "*zhi*" is considered another surprising feature (due to its very low frequency in end position in English) – the infants did not make the broader generalization, but kept with the narrower *AAzhi* pattern. Gerken et al. argue that infants only generalized if both surprising features were present. However, the authors make no comments on what would be the psychological reason or "rational" criterion that accounts for this behavior. They also do not take into consideration as a possible factor for their results the extremely short exposure time (21 seconds vs 2 minutes in their previous studies), and learning from a much longer test phase with a lot of added variability (4 different test strings were added in the test phase). We will come back to this apparently surprising finding in the General Discussion section.

These studies and others (Gerken & Bollt, 2008; Gómez, 2002) show that input variability is a strong factor driving generalization. However, it seems that it is not mere variability that is critical, but a specific pattern of variable input. How can this specific pattern be captured and defined by incorporating all variables?

Rule induction in adults

In research with adults, a study that aimed to elucidate the relation between the two forms of encoding (*item-bound* and category-based), and to further show that the type of encoding learners make depends on input properties is Reeder, Newport, and Aslin (2009, 2013). In a series of eight artificial language experiments (Exp. 1–4, 5A-5D), adults were familiarized with nonsense strings having the underlying structure: *(Q)AXB (R)*[2], in order to probe whether they can generalize *X* as a category, rather than just memorize the exact strings. Participants heard different subsets of strings from this grammar, which displayed different combinations of items. In the test phase, participants were tested on the withheld (novel) grammatical strings, as well as on ungrammatical strings (*AXA* or *BXB* strings). In our terminology, participants' ability to recognize the novel strings as grammatical implies that they made the correct *category-based generalization* (i.e. *AXB*). Reeder, Newport, and Aslin (2013) found four factors with different effects on generalization: *richness of contexts* (all *As* and *Bs* concatenated with all *Xs*) drives generalization (Exp. 1), *reduced number of exemplars* does not impede generalization (Exp.2), but *incomplete overlap of contexts* (*Xs* concatenated only with 2/3 *As* and 2/3 *Bs* – in Exp.3) and *longer exposure time* (increased frequency of items – in Exp.4) reduce the likelihood of generalization. In Experiments 5A – 5D, the input mirrored that of Experiments 1–4, respectively, but they added a minimally overlapping *X*-word that occurred in only a single *A1_B1* context. They found a similar pattern of results as in Experiments 1– 4, i.e. subjects generalized the novel minimally overlapping *X* to the full range of the X category. However, when exposure increased in Experiment 5D, learners were less likely to generalize, mirroring the results found in Experiment 4. However, the authors gave no consistent explanation for the different effects of these factors on generalization. Are they independent factors? Why did participants still make *category-based generalizations* when exposed to the input in Experiment 3, but were significantly less inclined to do so when they had increased exposure to the same input (with the same statistical properties; Experiment 4 and 5D)? These results suggest that statistical properties of the input interact with degree of exposure. The authors also suggest that at some degree of sparseness and overlap of contexts, there must be a threshold for shifting from word-by-word learning to category generalization. We propose that finding an approach to calculate this threshold would explain how the *item-bound generalization* and the *category-based generalization* are related, and help answer the question

[2]Each letter stands for a category of words and those in brackets mark optional categories. Each category had three words.

whether the learning mechanisms underlying these two types of generalizations are the same, or different. While this study found some factors that trigger or impede generalization, the authors did not capture the specific pattern of variability and exposure that drives generalization.

Aslin and Newport (2012) argue that for both Reeder et al. (2009) and Gerken (2006) the key point is the reliability of the distributional cues: the consistency/inconsistency of the distribution of context cues determines whether a generalization is formed, or specific instances are learned. In other words, they hypothesize that statistical learning is the mechanism that underlies both *item-bound generalizations* and *category-based generalizations*. Their view is very much in line with the model we propose in the next section. However, they do not give an account as to how the same mechanism outputs two qualitatively different forms of generalization, what kind of context cue distribution leads to one or the other generalization, and why it is the case that the same mechanism can have two different outcomes. Also, if the distribution of the context cues is the factor driving generalization, why does increased exposure to the same statistical distribution negatively impact generalization (Experiments 4 and 5D in Reeder et al., 2013)?

Summarizing, while these studies provided important insights into generalization, showing that infants and adults can tune into both forms of encoding, *item-bound generalizations* and *category-based generalizations*, they do not explain how learners converge on higher-order *category-based generalizations*. Are the two forms of encoding outcomes of two separate mechanisms? Or are they two outcomes of the same mechanism, with either a gradual transition or an abrupt switch from a lower-level item-bound to a higher-order abstract one? What are the independent factors that trigger the transition from *item-bound* to *category-based generalizations*? Below we sketch a new model that captures the specific pattern of variable input interacting with cognitive constraints, to give a clear and complete picture of the mechanism underlying rule induction and to unify previous findings in one consistent account.

An entropy model for linguistic generalization

Introduction to the model

We present a new approach to generalization from an information-theoretic perspective, and we propose a new entropy model for rule induction. Our entropy model is designed to unify the findings of the artificial grammar studies discussed so far under one consistent account. The basic intuition of our model is that the factor triggering the transition from *item-bound* to *category-based generalizations* is *input complexity*, as measured by the information-theoretic concept of entropy. Intuitively, entropy quantifies the complexity of a set of items, and it varies depending both on the number of items and their frequency distribution. Entropy increases if the number of items increases, and it also increases if items have a homogeneous frequency distribution. Entropy can also be defined as uncertainty, in this context uncertainty (or surprise) about the occurrence of specific items or configurations of items. Both factors (number and frequency distribution of items) contribute to the uncertainty of the occurrence of specific items or configurations.

The concept of entropy is not new to this domain. Pothos (2010) proposed an information-theoretic model to describe performance in acquiring knowledge about a finite-state grammar. He employed Shannon's entropy (Shannon, 1948) as a measure of quantifying the ease of predicting if a string of items is consistent with a trained language, i.e. if a string would possibly be part of the trained language. However, this model tackles *item-bound generalizations* only, as finite-state grammars contain a finite number of items, and they define regularities in terms of specific items (rather than categories).

Unlike Pothos's model, the entropy model we propose gives a conceptual analysis that encompasses both *item-bound generalizations* and *category-based generalizations*. In addition to *entropy, channel capacity* (Shannon, 1948) is another critical factor, as our model hypothesizes that *rule induction is an encoding mechanism gradually driven as a natural automatic reaction by the brain's*

sensitivity to the input complexity (entropy) interacting with the finite encoding power of the human brain (*channel capacity*). Thus, our model is based on the following tenets:

(1) *Item-bound generalization* and *category-based generalization* are not independent; they are outcomes of the same encoding mechanism that *gradually* goes from lower-level item-bound to higher-order abstract generalizations.
(2) The independent factors that drive the gradual transition from *item-bound* to *category-based generalization* are *input complexity (entropy)* and *the finite encoding power of the human brain* (*channel capacity*).

This model thus specifies a quantitative measure for the gradual transition from *item-bound* to *category-based generalization* by capturing the specific pattern of variable input interacting with cognitive mechanisms.

Entropy, as an information-theoretic concept, varies as a function of the number of items in the input and their probability of occurrence (which is a function of their relative frequency).For a random variable *X*, with *n* values $\{x_1, x_2 \ldots x_n\}$, Shannon's entropy (Shannon, 1948), denoted by *H(X)*, is defined as:

$$H(X) = \; -\sum_{i=1}^{n} p(x_i) log p(x_i)$$

where Σ denotes the sum, and $p(x_i)$ is the probability that x_i occurs. Probability shows how likely it is that a value x_i occurs. *Log* should be read as *log* to the base 2 here and throughout the paper. Entropy is used in our model to capture and describe a property of the input – *a specific pattern of complexity (or variability)*, and as a measure of this property, i.e. a measure of *input complexity*. Entropy has the following properties:

(1) For a given set of *n* items from the input, entropy (H) is zero, if the probability of one item is 1 and the probabilities of all the other items are zero. Intuitively, this is a set with the lowest complexity, and *uncertainty*. In psychological terms, an event with only one outcome with a maximum probability of occurrence is totally predictable, i.e. the amount of surprise when that outcome occurs is zero.
(2) For a given set of *n* items, the entropy is maximal if the distribution of the items' probabilities is uniform, i.e. when all the probabilities are equal (for example, for $n = 4$ and each item has $p = .25$, H = 2). Due to the equal probabilities, intuitively this set has the highest uncertainty about specific items' occurrence. In psychological terms, an event that has many outcomes which are equally probable to happen creates the highest amount of surprise.
(3) If all the probabilities are equal, the entropy of a set of items increases as a function of the number of discrete items.
(4) Any change to render the probabilities of the items unequal (i.e. some items are more probable than others) causes a decrease in entropy.

Taken together, these properties capture the unique dynamics between both factors (number and probability distribution of items) that defines a specific pattern of variability that our model proposes to be relevant for the process of rule induction.

C*hannel capacity* (C) describes the amount of entropy that can be sent through the channel per unit of time (Shannon, 1948). If H < C, information can be sent through the channel at the channel rate (*C)* with an arbitrarily small frequency of errors (equivocations) by using a proper encoding method. If H > C, it is possible to find an encoding method to transmit the signal over the channel, but the rate of transmission can never be higher than C. *Channel capacity* is employed here to model the finite encoding power of the information encoding system. Intuitively, the capacity to encode specific items and relations between them is finite. Thus, depending on the degree of input complexity and the finite encoding power (i.e. channel capacity), different forms of information encoding are necessary to encode the complexity of a given input.

Predictions of the model

(1) *Item-bound generalization* and *category-based generalization* are not independent mechanisms. Rather, they are outcomes of the same information encoding mechanism that *gradually* goes from a lower-level form of encoding (*item-bound generalization)* to a higher-order abstract encoding (*category-based generalization)*, as triggered by the interaction between *input complexity* and the finite encoding power of the brain. The encoding mechanism moves gradually from an *item-bound* to a *category-based generalization* as a function of increasing *input complexity* (entropy), as follows:

(a) If the input entropy is low – that is below or matches the *channel capacity*, then the input can be encoded using an encoding method that matches the input statistical structure, i.e. the probability distribution of the specific items in the input. Thus, the items with their specificity defined by their uniquely identifying features (acoustic, phonological, phonotactic, prosodic, distributional, etc.) and their specific probability distribution can be transmitted through the channel (i.e. encoded) at the default channel rate (i.e. amount of entropy per unit of time) and stored by *item-bound encoding* (i.e. probability matching to the input).

Examples of *item-bound encoding* would include rules like "*ends in di*", or rules specifying what specific items would follow each other (e.g. *ba* or *ge* follows *wo*).

(b) If the finite *channel capacity* of the encoding system is exceeded by the *input entropy*, it is possible to find a proper method that encodes more information (entropy), but the rate of encoding cannot be higher than the default channel capacity (Shannon, 1948). It is precisely this essential design feature of the *channel capacity* which "forces" the information processing system to re-structure the information to *gradually* – bit by bit – shape the *item-bound encoding* into another form of encoding. Remember the *channel capacity* theorem (Shannon, 1948): if H >C, another encoding method can be found to transmit the signal, but the rate of transmission stays constant. Re-structuring the information entails re-observing the item-specific features and structural properties of the input and identifying similarities and differences in order to compress the message by *gradually* reducing the number of specific features that individual items are coded for (i.e. to erase or "forget" statistically insignificant differences, that is low probability features). As a result of reducing ("forgetting") the specific features, i.e. differences, items are grouped in "buckets" (i.e. categories) based on nonspecific shared features, thus, a new form of encoding is created, which allows for higher *input entropy* to be encoded using the same given *channel capacity*, thus yielding higher-level *category-based encodings*. This would be the case for generalizations made over abstract categories: such as *AAB* or *AXB* patterns, which allow for more novel items to be included in these categories. Thus, the *channel capacity* promotes re-structuring (in accord with Dynamic Systems Theory invoked also in studies of other cognitive mechanisms – e.g. Stephen, Dixon, & Isenhower, 2009) for the purpose of adapting to noisier environments (i.e. in our terminology, increasingly entropic environments).

(2) An increase of *channel capacity*, (e.g. resulting from growth/development), reduces the need, and thus the tendency to move to a higher-order *category-based* form of encoding. Therefore, if infants and adults are exposed to the same input entropy, adults will have a lower tendency to make a *category-based generalization* than infants, because adults' *channel capacity* is higher.

(3) *Channel capacity* is used to model the finite encoding power of the human mind. We hypothesize that it is modulated by (unintentional) incidental memory capacity, attention and a general pattern-recognition capacity.

Therefore, the model hypothesizes that there is a *gradient of generalization*, in line with previous suggestions (Aslin & Newport, 2014), but it refines and extends this proposal, by further explaining how and why this gradual process happens. Sensitivity to entropy means a sensitivity to a specific pattern of variability in the input given by the degree of similarity/dissimilarity between items and their features and also their probability distribution, which assigns significance to specific items and their features. The more differences are encoded between specific items (i.e. many different specific features encoded for each item – measured in bits of information), the higher the degree of specificity of the encoding (i.e. *item-bound* specificity). Conversely, since the *channel capacity* places an upper bound on the number of bits encoded per unit of time, a reduction – *"gradual forgetting"* – of the encoded differences highlights more similarities, hence the lower the degree of specificity and the higher the degree of generality. Entropy captures this dynamics of specificity vs generality, and quantifies it in bits of information. Thus, a gradient of specificity/generality on a continuum from *item-bound* to *category-based encodings* can be envisaged in terms of less or more bits of information encoded in the representation.[3]

Application of the model to AGL

Given that entropy is defined as a property of a variable[4], the input must be organizable in variables that can take certain values. In artificial grammar studies using patterns like AAB, AXB, each position of the patterns creates a variable (a category of items), whose possible values are the specific items: for example, variable A in a study on learning an AAB pattern (*le_le_di*) is filled by *le, wi, ji, de*, etc. Each category of bigrams and trigrams creates a variable, whose possible values are the specific bigrams and trigrams: for example, *lele* is a value of the AA category of bigrams, *ledi* is a possible value of the AB category of bigrams, while *wiwije* is one of the values taken by the AAB category of trigrams. Similarly, in finite-state grammars, the strings generated by the grammar can be segmented in groups of bigrams and trigrams, which can be construed as variables in a similar way. Given this set of variables, we can calculate the entropy of the familiarization input.

For an entropy model to be relevant for the encoding mechanism under scrutiny here, evidence is needed that learners acquire knowledge about categories of items that can be construed as variables: there is extensive evidence that grammaticality judgments in artificial grammar learning are shaped by knowledge acquired about bigrams and trigrams (Knowlton & Squire, 1996; Perruchet & Pacteau, 1990). Studies also showed that performance is predicted by the frequency of these chunks (Knowlton & Squire, 1994). There is also evidence for transfer of the knowledge to novel chunks, based on abstract analogy to the specific familiarization items (Brooks & Vokey, 1991; Vokey & Higham, 2005).

Pothos (2010) proposed an implementation method for his entropy model by suggesting that the entropy level (complexity) of each string can be calculated based on the probability that specific items will follow each other to form grammatical strings[5]. A lower entropy of a sequence of items (given by high probability bi-/trigrams and a low number of items) triggers a higher tendency to endorse it as possible in the familiarization language. Pothos's conclusions are in line with one of the predictions of our entropy model: a low entropy of the set of items enables *item-bound generalizations* (rules about which items follow each other).

[3]In terms of strength of neural networks, this degree of specificity vs. generality can be thought as the degree of strength of the memory pathways underlying the representations, i.e. in terms of stability vs. plasticity of memory networks (Kumaran et al., 2016).

[4]A variable X is a set of x values, where x ranges from $\{0, x_1, x_2 \ldots x_n\}$.

[5]The author provides a method for calculating entropy of every test string based on the familiarization items. We had some difficulty implementing his model, given that his method of calculating entropy of each test string based on the familiarization stimuli differs conceptually from our vision on how the entropy of the familiarization set has an effect on the mechanism of generalization. These conceptual differences might be due to the fact that his model addresses only item-bound generalizations, while our model encompasses both item-bound and category-based encoding. However, we will not discuss these differences here, as we think that they do not fall under the scope of this paper.

A unified account for previous studies. A brief proof of concept

A reinterpretation according to our entropy model can be given to Gerken's findings, to help answer the unanswered questions mentioned in the first section of this paper. Tables 1 and 2 display the familiarization stimulus sets for the two conditions tested by Gerken (2006), plus additional entropy calculations as per the entropy model presented in this paper. In our entropy calculations, each string contains four bigrams ([**begin**-A], [AA], [AB], [B-**end**]), to include the crucial information carried by the beginning and ending of a string by modeling an empty slot in the first and last bigram of the string. Likewise each string contains three trigrams ([**begin**-AA], [AAB], [AB-**end**]). The entropy values of the stimulus set include the bigram entropy for all bigram sets (H[**begin**-A], H[AA], H[AB], H[B-**end**]) and the trigram entropy for all sets of trigrams (H[**begin**-AA], H[AAB], H[AB-**end**]), as well as the average bigram entropy ($H[\text{bigram}] = \frac{H[\textbf{begin}-A]+H[AA]+H[AB]+H[B-\textbf{end}]}{4}$), the average trigram entropy ($H[\text{trigram}] = \frac{H[\textbf{begin}-AA]+H[AAB]+H[AB-\textbf{end}]}{3}$). Since there is evidence that learning of grammars is shaped by knowledge acquired about bigrams and trigrams, as discussed in the previous section, and also because some learners might be parsing only some parts of the set of all bigrams/trigrams, while others might be parsing other sets of bigrams/trigrams, we deem an average of bigram entropies and an average of trigram entropies to be the relevant measure. Also, based on the results reported by Pothos (2010) an average bigram/trigram entropy seems to be a better predictor for performance than the sum of all bigram/trigram entropies.

In Gerken (2006), the experiment condition that had an input characterized by a higher entropy (Table 1) yielded generalization to the broader category-based *AAB* generalization, while the one with lower entropy (Table 2) resulted in a narrower item-bound generalization "ends in *di*".

Table 1. Entropy values of the input in the diagonal condition in Gerken (2006).

Diagonal condition
[A A B]
le le di
wi wi je
ji ji li
de de we
Entropy values
H[begin|A] = – [(*p(le)**$\log_2$*p(le)*) + (*p(wi)**$\log_2$*p(wi)*) + (*p(ji)**$\log_2$*p(ji)*) + (*p(de)**$\log_2$*p(de)*)] = – [*.25* * $\log_2$*(.25)* + *.25* * $\log_2$*(.25)*+ *.25* * $\log_2$*(.25)* + *.25* * $\log_2$*(.25)*] = 2
H[B|end] = – [(*p(di)$\log_2$*p(di)*) + (*p(je)**$\log_2$*p(je)*) + (*p(li)**$\log_2$*p(li)*) + (*p(we)**$\log_2$*p(we)*)] = 2**
H[AA] = – [(*p(lele)**$\log_2$*p(lele)*) + (*p(wiwi)**$\log_2$*p(wiwi)*) + (*p(jiji)**$\log_2$*p(jiji)*) + (*p(dede)**$\log_2$*p(dede)*)] = 2
H[AB] = – [(*p(ledi)**$\log_2$*p(ledi)*) + (*p(wije)**$\log_2$*p(wije)*) + (*p(jili)**$\log_2$*p(jili)*) + (*p(dewe)**$\log_2$*p(dewe)*)] = 2
H[AAB] = – [(*p(leledi)**$\log_2$*p(leledi)*) + (*p(wiwije)**$\log_2$*p(wiwije)*) + (*p(jijili)**$\log_2$*p(jijili)*) + (*p(dedewe)**$\log_2$*p(dedewe)*)] = 2
H[bigram] = 2 H[trigram] = 2

Table 2. Entropy values of the input in the column condition in Gerken (2006).

Column condition
[A A B]
le le di
wi wi di
ji ji di
de de di
Entropy values
H[bA] = – [(*p(le)**$\log_2$*p(le)*) + (*p(wi)**$\log_2$*p(wi)*) + (*p(ji)**$\log_2$*p(ji)*) + (*p(de)**$\log_2$*p(de)*) = 2
H[Be] = – [*p(di)$\log_2$*p(di)*] = 0**
H[AA] = – [(*p(lele)**$\log_2$*p(lele)*) + (*p(wiwi)**$\log_2$*p(wiwi)*) + (*p(jiji)**$\log_2$*p(jiji)*) + (*p(dede)**$\log_2$*p(dede)*) = 2
H[AB] = – [(*p(ledi)**$\log_2$*p(ledi)*) + (*p(widi)**$\log_2$*p(widi)*) + (*p(jidi)**$\log_2$*p(jidi)*) + (*p(dedi)**$\log_2$*p(dedi)*) = 2
H[AAB] = – [(*p(leledi)**$\log_2$*p(leledi)*) + (*p(wiwidi)**$\log_2$*p(wiwidi)*) + (*p(jijidi)**$\log_2$*p(jijidi)*) + (*p(dededi)**$\log_2$*p(dededi)*)] = 2
H[bigram] = 1.5 H[trigram] = 2

An entropy-based reinterpretation of the results by Reeder et al. (2009, 2013) eliminates the need for the four factors proposed by the authors, which are not independent, and they modulate generalization inconsistently (as we argued in the first section of this paper). We suggest that it is one factor (i.e. the amount of entropy contained by each set of stimuli) that consistently accounts for the results of all these experiments. Table 3 shows that the two data sets used in the first two experiments are similar in terms of entropy values, which explains the absence of a significant difference in learners' tendency to generalize, even though in Experiment 2 exposure is half as long and only half the number of exemplars were presented. The factor proposed by the authors (i.e. reduced number of exemplars) is insufficiently constrained and cannot account for this unchanged tendency in generalization. Consequently, their results are unexplained under their hypothesis. Just as Gerken (2010) suggested, it is not the mere number of exemplars that has an effect on generalization, but a specific pattern of variability. As we show in Table 3, this pattern of variability can be captured by input entropy. Even though the input was reduced to half the number of exemplars, the total entropy was only slightly reduced, which explains why learners' tendency to generalize remained almost the same. The entropy values of the set of stimuli used in Experiment 3 were significantly reduced as compared to the first two experiments, which can explain learners' lower likelihood to generalize the categories. The effect of increased exposure to the same stimulus set in the fourth experiment cannot be explained by the authors' hypothesis, as the input displayed the same statistical properties as in Experiment 3, but the tendency to generalize was significantly reduced. We would argue that increased exposure leads to stronger memory traces of the items, which allows for *item-bound generalization*, hence to a suppression of category-based generalization, which is in line with the predictions of our entropy model. The entropy values for Experiment 5 series (from A to D) are slightly higher than those for Experiment 1– 4, respectively, which explains the slightly higher tendencies to generalize.

In conclusion, our entropy model accounts for all the findings of these experiments and gives a complete and unifying picture of rule induction by capturing the specific pattern of input variability (*entropy*) interacting with exposure time (which affects working memory and therefore modulates *channel capacity*[6]). The predictions made by our entropy model are borne out: a low *input complexity* enables *item-bound generalizations*, while a high *input complexity* exceeding *channel capacity* increases the tendency toward *category-based generalizations*.

Testing the predictions of the entropy model

In the remainder of this paper we present two AGL experiments that test specific predictions made by our entropy model. To the best of our knowledge, these are the first AGL experiments that investigate the role of *input complexity* in linguistic generalization by specifically testing entropy-based predictions. The experiments presented here focus on the effect of *input complexity*, without

Table 3. Entropy values for all conditions in Reeder, Newport and Aslin (2013).

	Experiment_1	Experiment_2	Experiment_3	Experiment_4
Entropy values	H[AX] = 3.169 H[bA]/[Be] = 1.584 H[XB] = 3.169 H[AXB] = 4.169 H[bigram] = 2.376 H[trigram] = 3.502	H[AX] = 3.169 H[bA]/[Be] = 1.584 H[XB] = 3.169 H[AXB] = 3.169 H[bigram] = 2.376 H[trigram] = 3.169	H[AX] = 2.503 H[bA]/[Be] = 1.584 H[XB] = 2.503 H[AXB] = 2.584 H[bigram] = 2.043 H[trigram] = 2.530	H[AX] = 2.503 H[bA]/[Be] = 1.584 H[XB] = 2.503 H[AXB] = 2.584 H[bigram] = 2.043 H[trigram] = 2.530
	Experiment_5A	**Experiment_5B**	**Experiment_5C**	**Experiment_5D**
Entropy values	H[AX] = 3.32 H[bA]/[Be] = 1.584 H[XB] = 3.32 H[AXB] = 4.24 H[bigram] = 2.452 H[trigram] = 3.626	H[AX] = 3.32 H[bA]/[Be] = 1.584 H[XB] = 3.32 H[AXB] = 3.32 H[bigram] = 2.452 H[trigram] = 3.32	H[AX] = 2.807 H[bA]/[Be] = 1.584 H[XB] = 2.807 H[AXB] = 2.807 H[bigram] = 2.193 H[trigram] = 2.807	H[AX] = 2.807 H[bA]/[Be] = 1.584 H[XB] = 2.807 H[AXB] = 2.807 H[bigram] = 2.193 H[trigram] = 2.807

[6]Recall *channel capacity* quantifies the amount of entropy that can be processed per unit of time.

specifically measuring variations in channel capacity (i.e. individual biological/psychological capacities), which were assumed to be roughly insignificant since we tested participants of similar age and backgrounds. The following hypothesis was tested:

> *Item-bound generalization* and *category-based generalization* are not independent mechanisms. Rather, they are outcomes of the same information encoding mechanism that *gradually* goes from a lower-level *item-bound* encoding to a higher-order abstract encoding (*category-based generalization)*, as triggered by the *input complexity.*

This hypothesis allows for the following two specific predictions to be tested:

(i) the lower the *input complexity* (entropy), the higher the tendency toward *item-bound generalizations*, and, consequently, the lower the tendency to make a *category-based generalization*;
(ii) the higher the *input complexity* (entropy), the higher the tendency to make a *category-based generalization*.

To test these predictions, we designed several versions of the same artificial grammar (3-syllable XXY structure[7]) in order to expose participants to different input entropies in three groups: high, medium and low entropy. An ensuing test phase presented participants with a grammaticality judgment task, where they were asked a yes/no question to indicate if they accepted the test strings as possible in the familiarization language. The test included four types of test strings that were designed to test each type of rule induction, as presented below.

Familiar-syllable XXY (XXY structure with familiar X-syllables and familiar Y-syllables) – **correct answer: yes – accept** – this is a test case that is intended to check learning of the familiar strings. All groups are expected to accept these strings as grammatical, either due to having encoded a category-based generalization in the high and medium entropy conditions, or due to an item-bound generalization in the low entropy condition.

New-syllable XYZ (XYZ structure with new syllables) – **correct answer: no – reject** – this is the complementary test case, which is intended to check learning of the familiar strings and string pattern. It is designed to back up and complement results for the familiar-syllable XXY strings as follows: if the forms of encoding – either ITEM[8] or CATEG – trigger acceptance of familiar XXY strings, then they should trigger rejection of the structurally and item non-compliant test cases (new XYZ). Thus, all groups are expected to reject this test type, with no between-group difference. If these strings are not consistently rejected, the interpretation of the results for familiar XXY cannot be valid.

New-syllable XXY (XXY structure with new syllables) – **correct answer: yes – accept** – this is a test case that is intended to be the TARGET test string type to check generalization of rule to novel strings (*CATEG*). The number of correct answers is expected to be a function of entropy condition: the highest number of acceptances is expected in the high entropy group, followed by the medium, and the low entropy.

However, absolute mean rates (percentages) of acceptance of these strings do not constitute direct evidence for *category-based* vs *item-bound generalization*, unless they are compared against the mean rates of acceptance for the familiar XXY strings. Thus, if learners have an *item-bound* encoding of the set of specific syllables and/or their combinations in strings, they will be able to discriminate between **Familiar-syllable XXY** and **New-syllable XXY**, i.e. the rates of acceptance of these test types will be significantly different. A strong discrimination between these test types (**Familiar-syllable XXY** significantly more accepted than **New-syllable XXY**) would show that the encoding is highly *item-bound*. Conversely, similar rates of acceptance would show that the participants treat these test items as equally acceptable in the grammar, which means they encoded the items/strings as category-based generalizations. Given the first hypothesis of our model – that the encoding mechanism moves gradually from an *item-bound* to a *category-based generalization* as a function of increasing input entropy – a cross-condition comparison is predicted to

[7]An XXY pattern describes strings consisting of two identical syllables (XX) followed by another different syllable (Y): e.g. xoxoʃi; pypydy.
[8]For ease of presentation, *item-bound generalization* is denoted ITEM, and *category-based generalization* – CATEG).

show a gradually decreasing discrimination between these two test items: the low entropy group is expected to show the highest discrimination, followed by medium entropy, while the high entropy group is predicted to show the slightest discrimination.

Familiar-syllable XYZ (XYZ structure with familiar syllables[9]) – **correct answer: no – reject** – this is the complementary test case to the New-syllable XXY strings: if New-syllable XXY strings are accepted in a different proportion by the three groups due to hypothesized differences in types of encoding developed, then Familiar-syllable XYZ strings should also be treated differently across groups. We expect results for this test type to capture the two types of encoding competing against each other, because it is likely that the memory trace of familiar syllables drives acceptance of these ungrammatical strings with familiar syllables. Hence differences in performance are expected across groups, depending on the extent to which ITEM and CATEG are developed, i.e. to the *gradient of generalization*: the low entropy group is expected to yield the highest proportion of correct rejections, as (per hypothesis) they encoded the strings as frozen item-bound generalizations, which highlight clear mismatches between familiar and non-compliant combinations of specific items. In the high entropy group, *category-based generalization* will be predominant, and thus XYZ strings will be rejected for being inconsistent with the XXY pattern. The medium entropy group is expected to yield the lowest percentage of correct rejections, because it is likely that the memory traces of the individual familiar syllables work against a rejection, and because ITEM is too weak to have created a strong memory trace of the entire strings, while CATEG is not strongly developed to consistently reject the incorrect XYZ pattern: in this case, the two forms of encoding compete against each other with almost similar strength. Therefore, we expect a U-shape pattern of correct rejections as a function of increasing input entropy.

Experiment 1

Method

Participants

Thirty-five Dutch speaking adults (26 females and 9 males, age range 19–26, mean 22) participated in Experiment 1. One additional participant was tested, but excluded for being familiar with AGL setups. Only healthy participants that had no known language, reading or hearing impairment or attention deficit were included. They were paid 5 EUR for participation.

Familiarization stimuli

Participants were exposed (aurally) to 3-syllable strings that implemented a miniature artificial grammar, which closely resembled the structural pattern used by Gerken (2006), i.e. the strings had an underlying XXY structure, where each letter represents a set of syllables. All syllables consisted of a consonant followed by a long vowel, to resemble common Dutch syllable structure (e.g./xo/,/ʃi:/). The subset of syllables used in the two *X* slots of the pattern – to be called X-syllables – did not overlap with the subset of syllables used for the *Y* slot of the pattern – to be called Y-syllables. The subset of consonants used for the X-syllables did not overlap with the subset of consonants used for the Y-syllables.

A Perl script generated the syllables and strings, and checked the CELEX database (Baayen, Piepenbrock, & Gulikers, 1995), to filter out existing Dutch words. All the syllables were recorded in isolation by a female Dutch native speaker in a sound-proof booth, using a TASCAM DA-40 DAT-recorder. Syllables were recorded one by one, as they were presented to her on a screen, and she was instructed to use the same intonation for each syllable. The recorded syllables were spliced together to form the strings of the language using Praat (Boersma, 2001; Boersma & Weenink, 2014).

[9]A subset of the syllables used in familiarization were concatenated to create XYZ test strings with familiar syllables. Any of the X-syllables and Y-syllables were randomly assigned to the X, Y or Z slot of the XYZ pattern.

The experiment consisted of three exposure phases with intermediate test phases, followed by a final test phase. In the exposure phases, a total of 72 XXY strings were presented, 24 per each phase. The order of presentation was randomized for each participant separately (complete stimulus set in Appendix). Intermediate tests were included to gauge the learning process as a function of exposure. The experiment had a between-subjects design, and participants were assigned randomly to one of the three conditions: High Entropy, Medium Entropy and Low Entropy.

Entropy values of familiarization conditions

To obtain the desired variation in *input complexity* (entropy) across conditions, two factors were manipulated: (1) the number of X-syllables and Y-syllables; and (2) the number of repetitions of each syllable (i.e. syllable frequency). By applying Shannon's entropy formula as described in the previous sections, three different values for *input complexity* were obtained, as follows:

(1) **Low Entropy**: 6 X-syllables and 6 Y-syllables with each syllable used 4 times in each familiarization phase. To generate the XXY strings, all 6 XX pairs were concatenated with all 6 Y-syllables, but different subsets (consisting of 24 XX_Y combinations) were used for each familiarization phase. The same procedure was applied to the other conditions. All three familiarization phases had the same entropy values: the average bigram entropy (H[bigram]) was 3.08, the average trigram entropy (H[trigram]) was 3.91, and the total average entropy (H[total]) was 3.5 (the average bigram/trigram entropies were calculated here in the same way as presented in section 3. above for previous studies – see Table 4 for complete entropy calculations). Since there is evidence that learning of grammars is shaped by knowledge acquired about bigrams and trigrams, as discussed in section 2.3., and also because some learners might be parsing the familiarization set mostly at the level of bigrams, while others might parse it mostly at the level of trigrams, we deem an average between bigram and trigram entropy to be the relevant measure (based on Pothos (2010), as mentioned in section 3 above).

(2) **Medium Entropy**: 12 X-syllables and 12 Y-syllables (6 different X-syllables and 6 different Y-syllables were added to those in Low Entropy (Experiment 1) with each syllable used 2 times in each familiarization phase. All three familiarization phases had the same entropy values: the average bigram entropy (H[bigram]) was 3.83, the average trigram entropy (H[trigram]) was 4.25, and the total average entropy (H[total]) was 4.

(3) **High Entropy**: 24 X-syllables and 24 Y-syllables (12 X-syllables and 12 Y-syllables were added to those used for Medium Entropy (Experiment 1) with each syllable used one time. All three familiarization phases had the same entropy values: the average bigram entropy (H[bigram]) was 4.58, the average trigram entropy (H[trigram]) was 4.58, and the total average entropy (H[total]) was 4.58.

Table 4. Entropy values for experiment 1.

Low Entropy	Medium Entropy	High Entropy
H[bX] = H[6] = -Σ[0.167*log0.167] = 2.58	H[bX] =H[12] = -Σ[0.083*log0.083] = 3.58	H[bX] =H[24] = -Σ[0.042*log0.042] = 4.58
H[XX] = H[6] = 2.58	H[XX] = H[12] = 3.58	H[XX] = H[24] = 4.58
H[XY] = H[24] = 4.58	H[XY] = H[24] = 4.58	H[XY] = H[24] = 4.58
H[Ye] = H[6] = 2.58	H[Ye] = H[12] = 3.58	H[Ye] = H[24] = 4.58
H[bXX] = H[6] = 2.58	H[bXX] = H[12] = 3.58	H[bXX] = H[XXY] =H[XYe] =
H[XXY] = H[XYe] = H[24] = 4.58	H[XXY] = H[XYe] = H[24] = 4.58	H[24] = 4.58
H[bigram] = 3.08	**H[bigram] = 3.83**	**H[bigram] = 4.58**
H[trigram] = 3.91	**H[trigram] = 4.25**	**H[trigram] = 4.58**
H[total] = $\frac{H[bigram]+H[trigram]}{2}$ = 3.5	**H[total] = 4**	**H[total] = 4.58**

Procedure

Participants were tested in a sound-proof booth and were told that they would listen to a "forgotten language" that would not resemble any language that they might be familiar with, but which had its own rules and grammar. They were told that the language had its own rules for the forms of words, and that those words were not known to them from any other language they might be familiar with. The instructions were provided entirely in the beginning of the experiment. The instructions explained that the experiment had three phases, and during each phase several words from the language would be played. The participants were informed that the language had more words and syllables than what they heard in the familiarization phases. After each familiarization phase, they would have a short test, and at the end there would be a final test. Each test would be different from the other tests, and the tests were meant to check what they had noticed about the language that they listened to. They were instructed to decide, by pressing a Yes or a No button, if the words that they heard in the tests could be possible in the language that they heard. The experiment lasted around 5 minutes.

Test string types

All test items were 3-syllable strings designed as four different types: grammatical familiar, ungrammatical novel, grammatical novel, and ungrammatical familiar (as presented in section 4 above). Each of the three intermediate tests had four test strings (one of each type), and the final test had eight strings (two of each type). Thus, there were (4 +4 +4 +8 =) 20 test strings in total, and they were used in all three entropy conditions (complete test item set in Appendix).

Experiment 1: results

In order to test the effect of *input complexity* on generalization, the High Entropy, Medium Entropy and Low Entropy conditions were compared in a Generalized Linear Mixed Model, with Accuracy (correct acceptance/rejection) as dependent variable and Entropy condition, Test String Type x Entropy condition interaction, Test phase x Entropy condition interaction as fixed factors, and Subject and Trial as random factors. An alpha level of *.05* was used for all statistical tests. We started fitting the data from the intercept-only model and added the random and fixed factors one by one. We report here the best fitting model, both in terms of model's accuracy in predicting the observed data, and in terms of AICc (Akaike Information Criterion Corrected). There was a statistically significant Test String Type x Entropy condition interaction ($F(9, 679) = 6.363$, $p = .000$). There was no statistically significant main effect of Entropy condition ($F(2, 679) = 0.401$, $p = .67$). Results indicated a non-significant trend in the predicted direction for Test phase x Entropy condition interaction ($F(9, 679) = 1.243$, $p = .26$).

Figure 1 presents the mean rate of acceptance (percentage of acceptances per group) across conditions for Familiar-syllable XXY and New-syllable XXY. The mean acceptance rate of New-syllable XXY in High Entropy was 80% (Mean = .80, SD = .403), in Medium Entropy was 73% (Mean = .73, SD = .446), and in Low Entropy was 65% (Mean = .65, SD = .480). One-sample Wilcoxon Signed-Rank tests indicated a statistically significant above-chance mean acceptance for New-syllable XXY in High Entropy ($Z = 4.648$, $SE = 118.12$, $p = .000$; Cohen's effect size $d = 0.6$), in Medium Entropy ($Z = 3.615$, $SE = 118.12$, $p = .000$; Cohen's effect size $d = 0.47$), and in Low Entropy ($Z = 2.292$, $SE = 103.82$, $p = .022$, Cohen's effect size $d = 0.31$). In High Entropy there was a significant difference between acceptance of Familiar-syllable XXY and acceptance of New-syllable XXY (M =.167, SD =.376; $t(3) = 2.721$, SE =0.853, $p = .007$); in Medium Entropy there was also a significant difference between performance on these tests (M =.233, SD =.427; $t(3) = 3.454$, SE =0.838, $p = .001$); and in Low Entropy the difference between performance on these tests was also significant (M =.327, SD =.511; $t(3) = 3.566$, SE =1.158, $p = .000$). Further, Cohen's effect size value (d = 0.36) and the effect-size correlation (r = 0.18) for the difference between performance

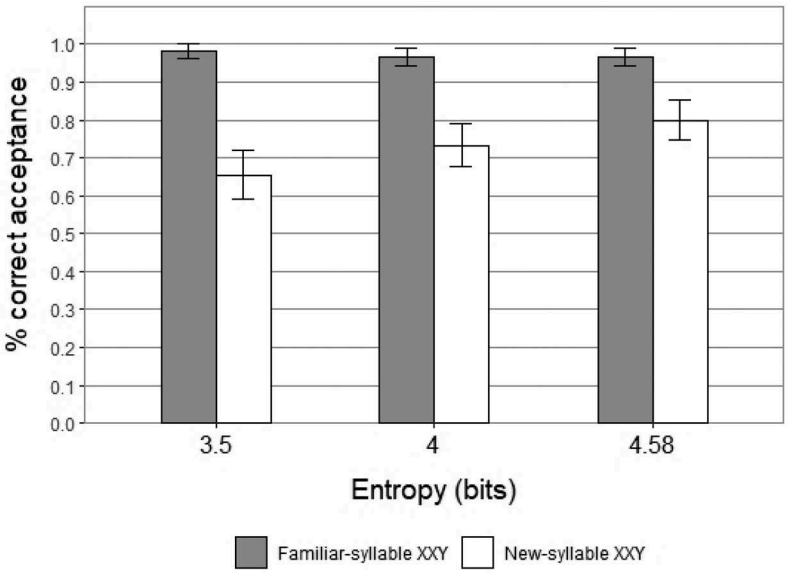

Figure 1. Percentage of correct acceptance for familiar-syllable XXY & new-syllable XXY. Error bars show standard error of the mean. Experiment 1.

on these tests in High Entropy vs. Low Entropy were higher than the same values for High Entropy vs. Medium Entropy (d = 0.15, r = 0.07), and also higher than the same values for Low Entropy vs. Medium Entropy (d = 0.21, r = 0.1).Figure 2 shows the mean rate of rejection for Familiar-syllable XYZ and New-syllable XYZ. The mean rejection of Familiar-syllable XYZ in High Entropy was 82% (Mean = .82, SD = .39), significantly different from the mean acceptance of Familiar-syllable XXY (*t (3) =2.529, SE = 0.851, p = .012*); 77% in Medium Entropy (Mean = .77, SD = .427), significantly different from the mean acceptance of Familiar-syllable XXY (*t(3) =3.147, SE = 0.837, p = .002*); and 91% in Low Entropy (Mean = .91, SD = .290), near-significantly different from the mean acceptance of Familiar-syllable XXY (*t(3) =1.683, SE = 1.185, p = .093*).

Figure 2 shows the mean rate of rejection for Familiar-syllable XYZ and New-syllable XYZ. The mean rejection of Familiar-syllable XYZ in High Entropy was 82% (Mean = .82, SD = .39), significantly different from the mean acceptance of Familiar-syllable XXY (*t(3) =2.529, SE = 0.851, p = .012*); 77% in Medium Entropy (Mean = .77, SD = .427), significantly different from the mean acceptance of Familiar-syllable XXY (*t(3) =3.147, SE = 0.837, p = .002*); and 91% in Low Entropy (Mean = .91, SD = .290), near-significantly different from the mean acceptance of Familiar-syllable XXY (*t(3) =1.683, SE = 1.185, p = .093*).

Discussion

The results of Experiment 1 show that the mean acceptance of *new* XXY strings increases as a function of increasing entropy. Moreover, there were differences between the rates of acceptance of new XXY vs. familiar XXY strings depending on the entropy group. This shows differences between groups in terms of how learners encode the XXY strings: if the participants do not make a clear distinction between a new XXY and a familiar XXY, we conclude that they formed a *category-based generalization* (XXY) which applies equally to both familiar and new XXY strings. Thus, a smaller difference between the means of acceptance of these test types shows a higher tendency to make *category-based generalizations*. The results showed that in the high entropy group this difference is smaller than in the medium entropy one, which is smaller than in the low entropy

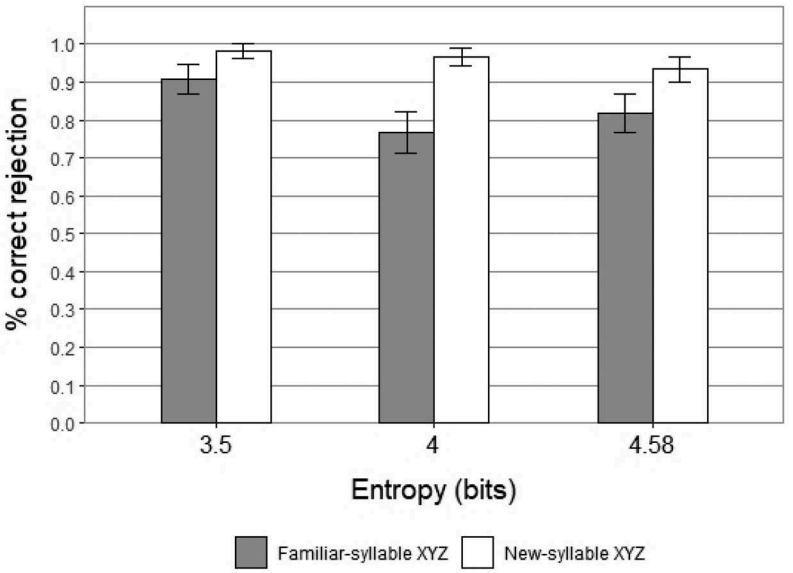

Figure 2. Percentage of correct rejection for familiar-syllable XYZ & new-syllable XYZ. Error bars show standard error of the mean. Experiment 1.

group. Hence these results indicate that learners exposed to higher *input complexity* had a higher tendency to make *category-based generalizations* and to generalize to novel strings displaying the underlying XXY pattern, which is in line with the predictions of our entropy model.

The rate of correct rejection for XYZ strings with familiar syllables is very high in the low entropy group, although the rate of acceptance for new XXY strings is rather low (Figure 3). As it agrees with our predictions, this result suggests that the *input complexity* did not exceed the *channel capacity* and it enabled learners to extract rules of specific sequencing of the memorized items (i.e. ITEM is dominant and signals a clear mismatch between grammatical and ungrammatical strings of specific items). In the high entropy group, there was also a firm rejection of XYZ strings with familiar syllables, but only as high as the acceptance of new XXY strings. This indicates that CATEG is strong enough to drive rejection of the XYZ strings. As predicted, the medium entropy group yielded the lowest performance of all groups. The interpretation is that increased *input complexity* prevents a strong memory trace of the entire strings, and thus ITEM cannot support a consistent and confident rejection of the XYZ strings. At the same time, CATEG is not strongly developed to consistently reject the incorrect XYZ pattern. To sum up, the results showed a roughly U-shaped performance on XYZ with familiar syllables, as a function of increased input entropy. Similar tendencies toward a U-shaped curve of learning were found in previous language acquisition studies, and they were argued to be due to the dynamics reflected by different mechanisms working simultaneously and interfering with each other (Rogers, Rakinson, & McClelland, 2004). Therefore, we interpret this U-shape pattern of results to show the two forms of encoding – *item-bound* and *category-based generalizations* – competing against each other with almost similar strength, thus creating the most uncertain situation for this task.[10]

The results showed that the decreasing trend of the rejection of familiar-syllable XYZ changes into an increasing trend roughly at the same entropy level where it meets the increasing trend of acceptance of new XXY. We hypothesize that the lowest point of the U-shaped trend of the rejection of familiar-syllable XYZ is the intersection point of the decreasing trend of XYZ and the increasing

[10]A similar U-shaped effect of stimulus complexity (entropy) on allocation of visual attention was found in infants – the "Goldilocks effect" (Kidd, Piantadosi, & Aslin, 2012).

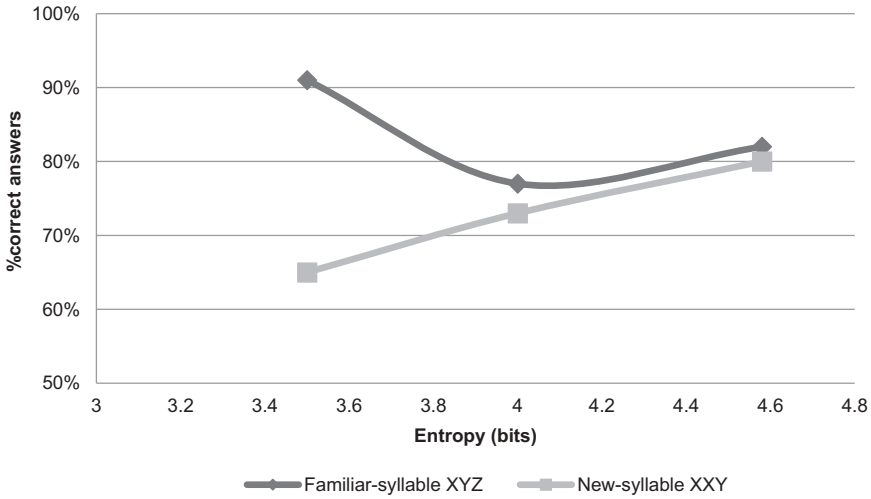

Figure 3. Percentage of correct acceptance of new-syllable XXY and correct rejection of familiar-syllable XYZ plotted against input entropy. Experiment 1.

trend of XXY. The calculated intersection point of the two trends – *y(New-syllable XXY) = y (Familiar-syllable XYZ)* – is *H = 4.2* (*y = 0.72)*, which allows the prediction that the rate of rejection of Familiar-syllable XYZ decreases until 72%, if the *input complexity* is H =4.2 bits. This value is predicted to be the point where the decreasing trend for Familiar-syllable XYZ reaches its minimum and changes into an increasing function, given that CATEG outperforms ITEM. This point is hypothesized to roughly mark the excess limit of the *channel capacity*.

A subsequent re-thinking of the XYZ strings with familiar syllables raised the question that these strings should have had an X_1X_2Y pattern (X_1 is different from X_2), to ensure that the reason for the rejection of these strings does not involve the inconsistency of using X-syllables in the last position of the strings, or Y-syllables in the first or second position of the string. Only two out of five Familiar-syllable XYZ strings did not have an X_1X_2Y pattern. However, this confound would have helped rejection of these strings more in the low entropy group, where it was easier to remember the specific familiar X-syllables and Y-syllables. An ANOVA with familiarization group (High Entropy, Medium Entropy and Low Entropy as between-subjects variable and test item (X_1X_2Y vs. non-X_1X_2Y) as within-subjects variable revealed no statistically significant difference between the rejection rate of X_1X_2Y strings and the rejection rate of the non-X_1X_2Y strings in any of the conditions (High Entropy: Mean[X_1X_2Y] = .81, Mean[non-X_1X_2Y] = .83, F(1,58) = .072, p = *.79*; Medium Entropy: Mean[X_1X_2Y] = .79, Mean[non-X_1X_2Y] = .73, F(1,58) = .293, p = *.59*; Low Entropy: Mean[X_1X_2Y] = .91, Mean[non-X_1X_2Y] = .91, F(1,53) = .000, p = *1.00*. Therefore, such a confound is highly unlikely to explain the results.

We designed intermediate tests to investigate the learning process as an interaction between input entropy and exposure time. On the one hand, we predicted that longer exposure to the familiarization items would strengthen the memory trace of the specific items, and thus it would make it easier to encode the specific syllables/strings. Thus, the tendency to make *category-based generalizations* will decrease as a function of increasing exposure time, as was shown in Reeder et al. (2013). On the other hand, a high input entropy would make remembering the specific items more difficult than a medium entropy and a low entropy. Thus, an interaction between input entropy and exposure time was predicted to show the following results: the acceptance of new XXY strings across the intermediate tests through the final test is expected to decrease in all entropy groups due to exposure time. But at a different rate, depending on the input entropy, as follows: the percentage of acceptance of new XXY strings should have a slowly decreasing trend in high entropy (because the more

complex input prevents forming memory trace of specific items and strings), a slightly steeper decreasing trend in medium entropy, and an even steeper decreasing trend in low entropy (because the more repetitive input allows remembering of specific items and strings). Although the results did not reach statistical significance, the trends match the predictions: in low entropy group the performance curve decreases slightly steeper than in the medium entropy, and steeper than in the high entropy one. Further research would need to be conducted with larger samples and longer exposure time to further investigate the generalization curve as an interaction between input entropy and exposure time.

Experiment 2

In Experiment 2, we further tested the effect of *input complexity* on generalization when learners are exposed to three other degrees of *entropy*. The purpose was to replicate the pattern of results obtained in Experiment 1, i.e. to find a gradually increasing tendency to make *category-based generalizations* as a function of increasing input entropy. We exposed adults to an XXY grammar similar to the one used in Experiment 1, but the three conditions had other degrees of entropy. For the Low Entropy (Experiment 2) condition we chose a lower entropy value than for Low Entropy (Experiment 1) (2.8 bits – 4 × 7 Xs/4 x 7 Ys) to test the prediction made by the simple linear regression equation that we fitted for the new *XXY* strings: at a lower entropy value (*H =2.8 bits*) the induction tendency will approach chance level (around 54%). The entropy value for the Medium Entropy (Experiment 2) condition (4.25 bits – 2 × 14 Xs/2 x 14 Ys) was chosen to test the specific prediction made by the simple linear regression equation that the mean performance on X1X2Y strings with familiar syllables will decrease as compared to the performance for Medium Entropy (Experiment 1) (for *H =4 bits* the performance was 77%): at *H =4.2 bits* the mean performance predicted is 72%. For the High Entropy (Experiment 2) condition we chose a higher entropy (4.8 bits – 1 × 28 Xs/1 x 28 Ys) than High Entropy (Experiment 1) in order to test if the tendency to abstract away from the specific input increases further or it stabilizes at a certain ceiling. The prediction is that at a certain degree of entropy the tendency to generalize will stabilize at a certain ceiling regardless of how much the entropy increases, due to the finite *channel capacity*, i.e. there will be no further increase in the tendency toward *category-based encoding*.

Method

Participants

Thirty-six Dutch speaking adults (30 females and 6 males, age range 18–34, mean 22) participated in the experiment. Only healthy participants that had no known language, reading or hearing impairment or attention deficit were included. They were paid 5 EUR for participation.

Familiarization stimuli

As in Experiment 1, participants were exposed to 3-syllable XXY strings. The same recorded syllables from Experiment 1 were used, but spliced together using Praat to form other strings than those used in Experiment 1, to obtain different degrees of *input complexity*. All three conditions (High Entropy, Medium Entropy, Low Entropy) had equal number of familiarization strings – 84 XXY strings in total (28 XXY strings in each familiarization phase) – which were presented in a randomized order per participant (complete stimulus set in Appendix). This was also a between-subjects design, and participants were assigned randomly to one of the three conditions.

Entropy values of familiarization conditions

The Shannon entropy formula and the entropy calculations were applied in the same manner as for Experiment 1 to obtain other three different values for *input complexity*, as follows:

1. **Low Entropy**: 7 X-syllables and 7 Y-syllables (with each syllable used 4 times in each familiarization phase). To generate the XXY strings for the Low Entropy condition, the 7 XX pairs were concatenated with the 7 Y-syllables to obtain 7 strings, which were repeated 4 times to obtain 28 strings which were used in all familiarization phases. The same procedure was applied to the other conditions. All three familiarization phases had the same entropy values: the average bigram entropy (H[bigram]) was 2.8, the average trigram entropy (H[trigram]) was 2.8, and the total average entropy (H[total]) was 2.8 (see Table 5 for complete entropy calculations).

2. **Medium Entropy**: 14 X-syllables and 14 Y-syllables (7 different X-syllables and 7 different Y-syllables were added to those used for Low Entropy with each syllable used 2 times. All three familiarization phases had the same entropy values: the average bigram entropy (H[bigram]) was 4.05, the average trigram entropy (H[trigram]) was 4.46, and the total average entropy (H[total]) was 4.25.

3. **High Entropy**: 28 X-syllables and 28 Y-syllables (14 X-syllables and 14 Y-syllables were added to those used for Medium Entropy with each syllable used one time. All three familiarization phases had the same entropy values: the average bigram entropy (H[bigram]) was 4.8), the average trigram entropy (H[trigram]) was 4.8, and the total average entropy (H[total]) was 4.8.

These values were different from the values in the entropy conditions used in Experiment 1 (repeated here for quick comparison $H[total]_{HiEN} = 4.58$, $H[total]_{MedEN} = 4$, $H[total]_{LowEN} = 3.5$).

Procedure

The procedure was the same as for Experiment 1.

Test string types and performance predictions

Participants in Experiment 2 were tested on the same types of test strings as for Experiment 1. Each test phase had the same number of test items as the phases for Experiment 1 (4 items per test), and the total number of test items was the same – 20 items in total (complete test item set in Appendix):

Familiar-syllable XXY – correct answer: yes – accept
New-syllable X_1X_2Y (three different new syllables) – **correct answer: no – reject**
New-syllable XXY – correct answer: yes – accept
Familiar-syllable X_1X_2Y (three different familiar syllables) – **correct answer: no – reject**

The predictions are similar to the those presented for Experiment 1 in section 4.

Experiment 2: results

In order to test the effect of *input complexity* on the process of generalizing, the High Entropy, Medium Entropy and Low Entropy conditions were compared in a Generalized Linear Mixed Model, with Accuracy (correct acceptance/rejection) as dependent variable and Entropy condition, Test String Type x Entropy condition interaction, Test phase x Entropy condition interaction as fixed factors, and Subject and Trial as

Table 5. Entropy values for Experiment 2.

Low Entropy	Medium Entropy	High Entropy
H[bX] =H[7] = 2.8	H[bX] =H[14] = 3.8	H[bX] =H[28] = 4.8
H[XX] = H[7] = 2.8	H[XX] = H[14] = 3.8	H[XX] = H[28] = 4.8
H[XY] = H[7] = 2.8	H[XY] = H[28] = 4.8	H[XY] = H[28] = 4.8
H[Ye] = H[7] = 2.8	H[Ye] = H[14] = 3.8	H[Ye] = H[28] = 4.8
H[bXX] = H[7] = 2.8	H[bXX] = H[14] = 3.8	H[bXX] = H[28] = 4.8
H[XXY] = H[XYe] = H[7] = 2.8	H[XXY] = H[XYe] = H[28] = 4.8	H[XXY] =H[XYe] = H[28] = 4.8
H[bigram] = 2.8	H[bigram] = 4.05	H[bigram] = 4.8
H[trigram] = 2.8	H[trigram] = 4.46	H[trigram] = 4.8
H[total] = $\frac{H[bigram]+H[trigram]}{2}$ = 2.8	H[total] = 4.25	H[total] = 4.8

random factors. An alpha level of *.05* was used for all statistical tests. We started fitting the data from the intercept-only model and added the random and fixed factors one by one. We report here the best fitting model, both in terms of model's accuracy in predicting the observed data, and in terms of AICc (Akaike Information Criterion Corrected). There was a statistically significant Test String Type x Entropy condition interaction ($F(9, 699) = 5.038$, $p = .000$). There was no statistically significant main effect of Entropy condition ($F(2, 699) = 0.260$, $p = .77$). Results indicated a non-statistically significant trend in the predicted direction for Test phase x Entropy Group interaction ($F(9, 699) = 1.163$, $p = .32$).

Figure 4 shows the mean acceptance rates across conditions for Familiar-syllable XXY and New-syllable XXY. The mean rate of acceptance for New-syllable XXY in High Entropy was 80% (Mean = .80, SD = .403), for Medium Entropy was 77% (Mean = .77, SD = .427), and for Low Entropy was 57% (Mean = .57, SD = .5). One-sample Wilcoxon Signed-Rank tests indicated a statistically significant above-chance mean acceptance for New-syllable XXY in High Entropy ($Z = 4.648$, $SE = 118.12$, $p = .000$; Cohen's $d = 0.6$) and in Medium Entropy ($Z = 4.131$, $SE = 118.12$, $p = .000$; $d = 0.53$), but in Low Entropy the mean acceptance was not significantly above chance ($Z = 1.033$, $SE = 118.12$, $p = .3$, $d = 0.13$). In High Entropy there was a significant difference between acceptance of Familiar-syllable XXY and acceptance of New-syllable XXY (M =.167, SD =.376; $t(3) = 2.161$, SE =0.643, $p = .031$); in Medium Entropy there was also a significant difference between performance on these tests (M =.233, SD =.427; $t(3) = 2.542$, SE =0.624, $p = .011$); and in Low Entropy the difference between performance on these tests was also significant (M =.327, SD =.511; $t(3) = 4.335$, SE =0.683, $p = .000$). Further, Cohen's d (d = 0.73) and the effect-size correlation (r = 0.34) for the difference between acceptance of Familiar-syllable XXY and acceptance of New-syllable XXY in High Entropy vs. Low Entropy were higher than the same values for High Entropy vs. Medium Entropy (d = 0.09, r = 0.04), and also higher than the same values for Low Entropy vs. Medium Entropy (d = 0.63, r = 0.3).

Figure 5 displays the mean rate of rejection for Familiar-syllable X_1X_2Y and New-syllable X_1X_2Y. The mean rejection rate for Familiar-syllable X_1X_2Y was 90% for High Entropy (Mean = .90, SD = .303), not significantly different from the mean acceptance of Familiar-syllable XXY ($t(3) = 0.647$, $SE = 0.704$, $p = .518$); 73% for Medium Entropy (Mean = .73, SD = .446), significantly different from the mean acceptance of Familiar-syllable XXY ($t(3) = 2.856$, $SE = 0.619$, $p = .004$); and 83% for Low Entropy (Mean = .83, SD = .376), significantly different from the mean acceptance of Familiar-syllable XXY ($t(3) = 2.028$, $SE = 0.711$, $p = .043$).

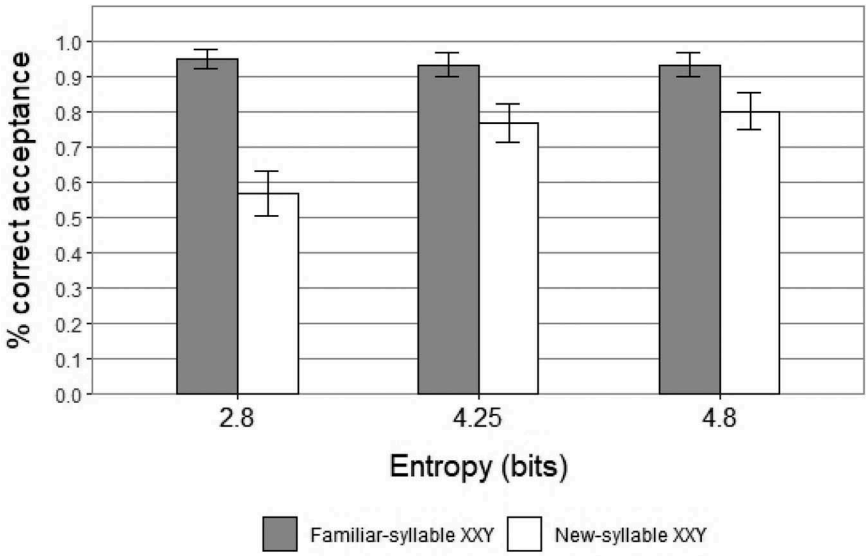

Figure 4. Percentage of correct acceptance for familiar-syllable XXY & new-syllable XXY. Error bars show standard error of the mean. Experiment 2.

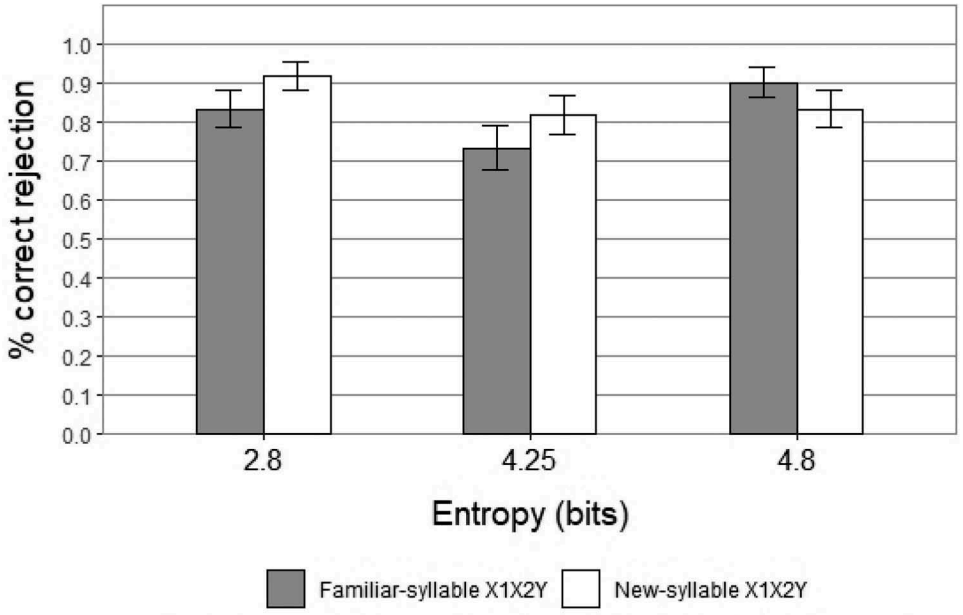

Figure 5. Percentage of correct rejection for familiar-syllable X1X2Y & new-syllable X1X2Y. Error bars show standard error of the mean. Experiment 2.

Comparing experiment 1 and experiment 2

To further test the effect of *input complexity* on the process of making generalizations, all the conditions from Experiment 1 and Experiment 2 were combined in an omnibus Generalized Linear Mixed Model, with Accuracy (correct acceptance/rejection) as dependent variable and Entropy condition, Test String Type x Entropy condition interaction, Test phase x Entropy condition interaction as fixed factors, and Subject and Trial as random factors. An alpha level of *.05* was used for all statistical tests. We started fitting the data from the intercept-only model and added the random and fixed factors one by one. We report here the best fitting model, both in terms of model's accuracy in predicting the observed data, and in terms of AICc (Akaike Information Criterion Corrected). There was a statistically significant Test String Type x Entropy condition interaction ($F(18, 1{,}378) = 5.782$, $p = .000$). There was no statistically significant main effect of Entropy condition ($F(5, 1{,}378) = 1.165$, $p = .32$), and no statistically significant Test phase X Entropy condition interaction ($F(18, 1{,}378) = 1.150$, $p = .29$).

Figure 6 shows the distribution of individual mean rates per type of test item in each group, namely Low Entropy, Medium Entropy, High Entropy, in Experiment 1 and Experiment 2.

A simple linear regression was calculated (Figure 7) to predict the rate of acceptance of New-syllable XXY based on the amount of input entropy. A significant regression equation was found ($F(1,4) = 243.54$, $p < .000$), with an R^2 of .98. Input entropy was a significant predictor for the rate of acceptance of New-syllable XXY.

Discussion

The results of Experiment 2 show that the mean acceptance of new XXY strings as grammatical increases as a function of increasing entropy. These results reveal a similar pattern to the results from Experiment 1: an increase in the tendency to abstract away from the memorized input as the *input complexity* increases. The different degrees of discrimination between XXY strings with novel

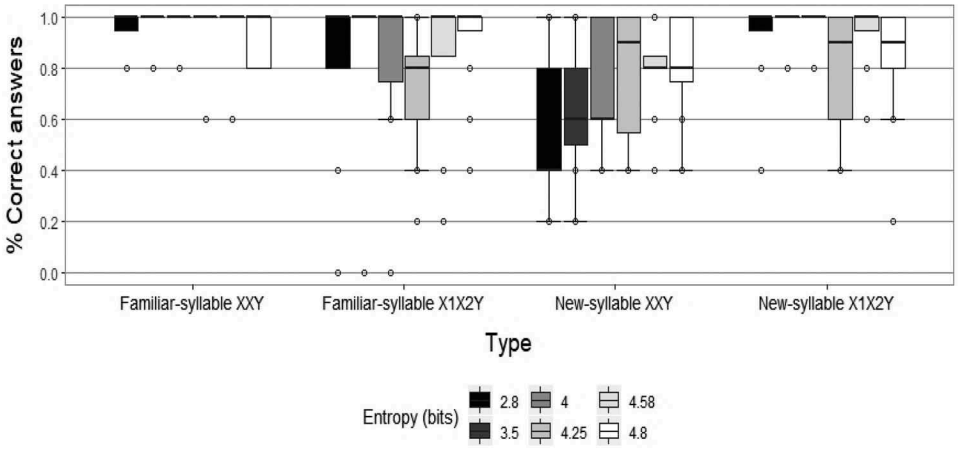

Figure 6. On the X-axis, the four types of test items: Familiar-syllable XXY; Familiar-syllable X1X2Y; New-syllable XXY; New syllable X1X2Y. On the Y-axis the mean rate of correct answers: correct acceptance for XXY strings (with familiar or new syllables) and correct rejection for X1X2Y strings (with familiar or new syllables). Experiments 1 & 2.

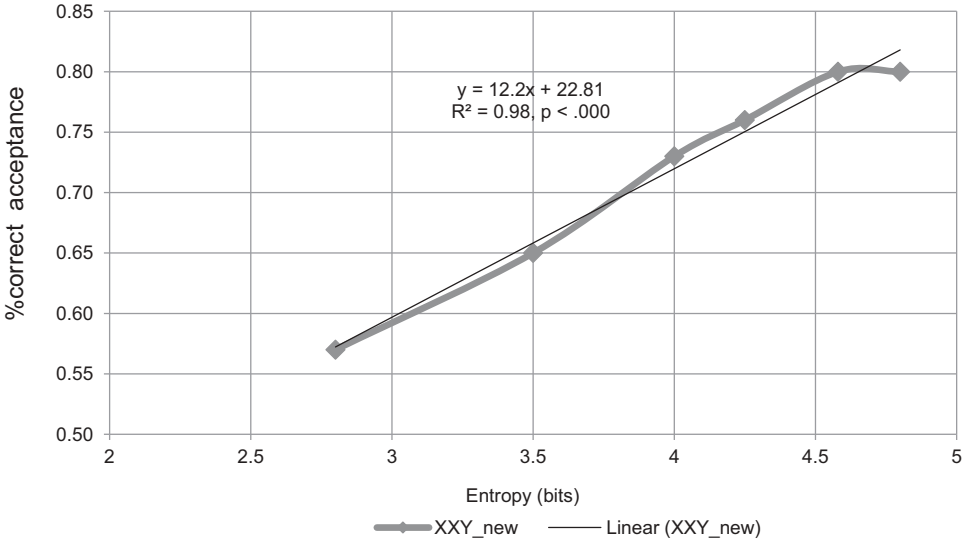

Figure 7. Percentage of acceptance for new-syllable XXY as a function of input entropy. Experiments 1 & 2.

syllables and XXY strings with familiar syllables show differences between groups in terms of their tendency to generalize to new items: in High Entropy this discrimination is lower than in Medium Entropy, which is lower than in Low Entropy. This difference suggests that learners in the High Entropy group had the highest tendency to fully generalize to novel XXY strings. Similar to Experiment 1, the roughly U-shaped performance in the case of ungrammatical Familiar-syllable X1X2Y strings may point to the competition between the two forms of encoding (the *item-bound* and *category-based generalization*).

When analyzed together, the results from Experiment 1 and Experiment 2 show that the rate of accepting XXY strings with new syllables as grammatical increases as the entropy increases. These results suggest an increasing tendency to make *category-based generalizations* as the *input complexity* increases, which is

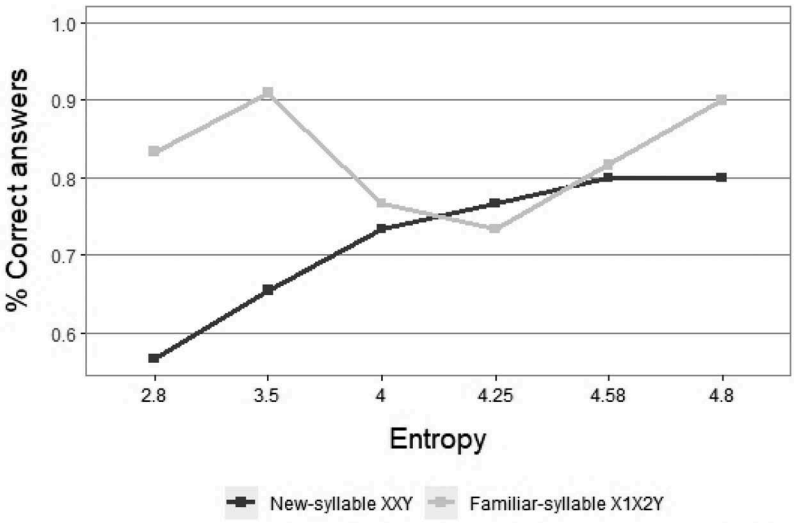

Figure 8. Percentage of correct acceptance of new-syllable XXY and correct rejection of familiar-syllable X1X2Y. Experiments 1 & 2.

consistent with the predictions made by our model. The same tendency is also shown by the decrease in the discrimination between XXY strings with novel syllables and XXY strings with familiar syllables, as the *input complexity* increases. As predicted, the mean acceptance of XXY strings with new syllables decreases to very close to chance level (57%) when the *input complexity* decreases to an entropy of 2.8 bits (Figure 8). When entropy increases from 4 bits (Medium Entropy – Experiment 1) to 4.2 bits (Medium Entropy – Experiment 2), the mean rejection rate for X1X2Y with familiar syllables decreases below 77% (the rejection rate at *H =4 bits*), to reach 73%, which is very close to the value predicted in section 8 (72%). The results show that when entropy increases from 4.58 bits (High Entropy – Experiment 1) to 4.8 bits (High Entropy – Experiment 2), the mean rate of acceptance for new XXY strings stabilizes at the value of 80% acceptance. This result suggests that around this amount of entropy (4.5 bits), the tendency to abstract away from specific items might stabilize at this ceiling regardless of how much the entropy increases. According to our entropy model, this ceiling effect is hypothesized to be due to the limitations in *channel capacity*.

The results of the experiments presented here can be also interpreted in terms of the degree of uncertainty of the cognitive system regarding the abstract structure of the input. The percentages of acceptance of novel XXY strings can be interpreted as the probability that a learner will abstract away from the specific items in the input and generalize to new XXY strings (for example, a probability of *0.8* at an input entropy of *4.8* bits, a probability of *0.57* at an input entropy of *2.8*, etc.). Under this interpretation, we used the information-theoretic measure of information load – I = – *log(p)* – to quantify the amount of uncertainty about input structure. A logarithmic curve was estimated (Figure 9) to predict uncertainty regarding the XXY structure of the input, based on the amount of input entropy. A significant logarithmic equation was found ($F(1,4) = 321.63$, $p < .000$), with an R^2 of .98. As shown in Figure 9, the uncertainty about structure is predicted to decrease logarithmically as the input entropy increases.

General discussion and conclusions

This study contributes to the ongoing debate on the learning mechanisms underlying rule induction. Some authors argued for two separate and qualitatively different mechanisms: *statistical learning* and *abstract rule learning* (Endress & Bonatti, 2007; Marcus et al., 1999), while others proposed that *statistical learning* underlies both types of generalizations (Aslin & Newport, 2012; 2014; Frost & Monaghan, 2016; Perruchet &

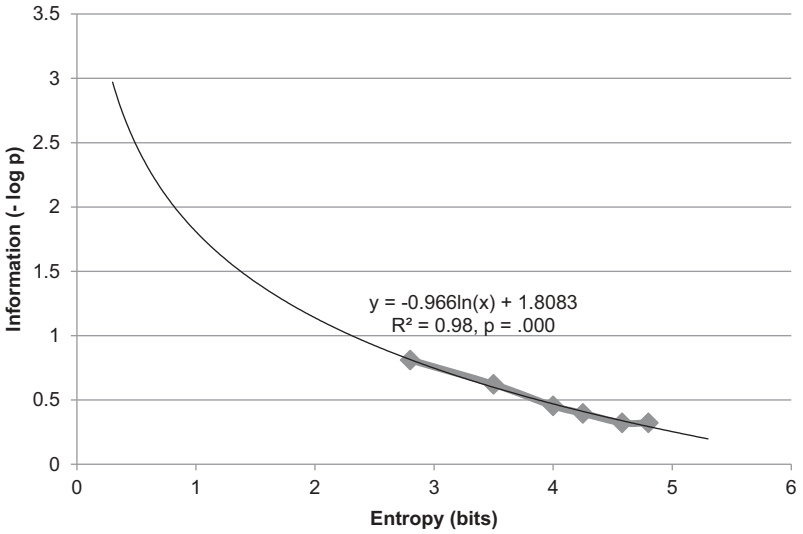

Figure 9. Uncertainty regarding the structure in the input.

Pacton, 2006). Recent computational models suggest that learners might combine statistical learning and rule-based learning (Adriaans & Kager, 2010; Frank & Tenenbaum, 2011). However, these studies do not explain how the two mechanisms relate to each other, and it has remained unclear if and how two qualitatively different forms of encoding (*item-bound* and *category-based generalizations)* can arise from a single mechanism. Our model and the results of our experiments support the view put forth by Aslin and Newport (2012; 2014). These authors suggested that it is the (in)consistency of the distribution of contextual cues that triggers a narrow generalization (*item-bound generalization*, in our terminology) or a broader generalization (*category-based generalization*). However, they did not provide a precise description of the pattern of such (in)consistencies, and their hypothesis cannot answer the following questions: 1) What is the specific pattern of (in)consistencies and how much (in)consistency is needed to move from *item-bound* to *category-based generalization*? 2) What triggers this transition? 3) Why infants (children) and adults need different degrees of (in)consistency? Some studies pointed to memory constraints, under the "Less-is-More" hypothesis, but without clear evidence or explanation (Hudson Kam & Chang, 2009; Hudson Kam & Newport, 2005, 2009; Newport, 1990, 2016). 4) Why does increased exposure to the same distribution of (in) consistent cues reduce the tendency to make *category-based generalizations*?

Our entropy model answers these questions and it accounts for both types of encoding by identifying two factors whose interplay is predicted to be the source of both types of generalizations: *input complexity (entropy)* and the *encoding power (channel capacity)* of the brain. Entropy captures and quantifies the specific pattern of (in)consistencies (i.e. input variability and surprise) that triggers rule induction. Thus, it allows for precise predictions on the generalizations that are made by learners exposed to any degree of *input complexity*. According to our model, learning starts out by memorizing specific items and by encoding these items and relations between them as *item-bound generalizations*. If the *input entropy* exceeds the *encoding power* of the brain, a higher-order form of encoding (*category-based generalization*) develops gradually.

Our model is in line with the general "Less-is-More" hypothesis, and it offers an extended and more refined formal approach to this hypothesis. Moreover, our model is in line with evidence from neurobiology (Frankland, Köhler, & Josselyn, 2013; Hardt, Nader, & Wang, 2013; Migues et al., 2016; Richards & Frankland, 2017) and neural networks research (Hawkins, 2004; Kumaran, Hassabis, & McClelland, 2016; MacKay, 2003) that converge on the findings and hypothesis that the memory system is designed to remember a certain degree of specificity (i.e. of entropy, in our terminology) in

order to prevent underfitting (missing specific parameters to correctly capture the underlying structure of the data), but also to prevent overfitting to past data/events (inadequately remembering and representing noise as underlying structure). According to these hypotheses and findings, rather than being faithful in-detail representations of the past data/events, memories are models optimized for future data integration, i.e. for better generalization and prediction of future data/events, in order to allow for more flexibility and better adaptability to noisy environments. As a refined information-theoretic extension of the "Less-is-More" hypothesis and in accord with these current developments in neurobiology and neural networks research, our entropy model offers a basis for conceptualization and quantification of the specific pattern of variability (input entropy) that the brain is naturally sensitive to, and which drives in a gradual fashion the rule induction mechanism in order to prevent overfitting to the input and to allow for representations of novel future input. From an information-theoretic point of view, our model proposes *channel capacity* (amount of entropy processed per unit of time) to reflect and quantify this design feature of the memory system proposed in neurobiology and neural network research that naturally and automatically places a lower and an upper bound on the degree of specificity (quantified in bits of information) represented in the neural pathways when encoding information, i.e. creating memory representations as actively predictive models of novel data/events. *Channel capacity* adds into the rule induction "formula" the essential dimension of time, i.e. a rate of encoding the entropy in the environment, as a natural physical system that is sensitive to a time-dependent and noisy (= highly entropic) inflow of information (Radulescu, Murali, Wijnen, & Avrutin (2019).

In two artificial grammar experiments we tested the model by investigating the effect of one factor of the model, namely *input entropy*, on rule induction. The findings strongly support the predictions of our entropy model, namely: *item-bound generalization* and *category-based generalization* are not independent outcomes of two qualitatively different mechanisms. Rather, they are outcomes of the same information encoding mechanism that *gradually* moves from a lower-level *item-bound* encoding to a higher-order abstract encoding (*category-based generalization)*, as triggered by the *input entropy*: the lower the *input entropy*, the higher the tendency toward *item-bound generalizations*, and, consequently, the lower the tendency to make a *category-based generalization*. The higher the *input entropy*, the higher the tendency to make a *category-based generalization*. These findings support our hypotheses, and bring evidence in favor of the validity of this entropy model for rule induction.

To further test the predictions of the entropy model proposed in this paper, the following outstanding questions should be investigated.

What is the effect of *input entropy* on infant rule induction? Further investigation is needed in order to probe whether the same pattern of results found in adults is replicated in infants, i.e. infants' tendency toward *category-based generalization* increases *gradually* as a function of increasing *input entropy*. Given that infants are hypothesized to have an overall lower *channel capacity*, they should be exposed to a lower range of entropy than adults. Previous research into infants' generalization mechanisms have already hinted at the significance of surprise (in our terminology, *entropy*) as a triggering factor for generalization (Gerken et al., 2015). However, the necessary amount and nature of input variability (or surprise) remains unclear: some studies show that at least three or four examples are needed for infants to generalize (Gerken, 2006, 2010; Gerken & Bollt, 2008; Peterson, 2011), but Gerken et al. (2015) claim that a single example suffices for generalization. Gerken et al. (2015) interpreted their results to support a Bayesian account of generalization, also suggested by Griffiths & Tenenbaum (2007): when an input is inconsistent with learners' prior model, hence surprising, learners seek a new hypothesis to accommodate the new (surprising) input. However, we think that these results raise concerns. Firstly, the authors used a very reduced exposure time (21 seconds) compared to previous studies – 2 minutes in Gerken (2006, 2010). Reduced exposure time is a crucial component in the mechanisms of rule induction, as noted in previous studies with adults (Reeder et al., 2013), and as explicitly predicted by the time-dependent *channel capacity* component of our entropy model. Secondly, the authors claim that generalization occurred from a single example, which is surprising compared to their prior model. Formally, learners' analysis

encompasses also their prior model, not just the one example they were exposed to in the lab. And we think (although the authors do not take this into account) that infants' analysis and learning extend also over the very long test phase (much longer than the familiarization phase itself), which includes 12 test trials with added variability (four different examples). Considering these concerns, the conclusion that infants generalize only from a single example is not decisive, and thus further research is needed to capture the nature and specific pattern of entropy (i.e. surprise) that drives infant rule induction.

In this paper, we proposed an original implementation of entropy as a quantitative measure of input complexity to artificial grammar learning with adults, but testing our model of a *gradual* transition from *item-bound* to *category-based generalization* with infants will require a different approach to implementation, in terms of calculations of entropy which should be different for infants, given that their cognitive system is still under development, so their *channel capacity* is hypothesized to be reduced. Infants might be more sensitive to local statistical properties of the input, rather than the entire set of items, and they might update their memory representations in an incremental fashion, as suggested already by evidence found in infant research (Gerken, 2010; Gerken & Quam, 2016). Thus, indeed due to a lower encoding power (*channel capacity*), underpinned by more plasticity of their developing memory system, infants' learning system may not be sensitive to average of bigrams/trigrams over the entire set of stimuli, since their encoding "window" might be more locally tuned (lower channel capacity). Moreover, since infants' sensitivity to similarities vs differences might develop gradually in the first year of life, given evidence that a primitive similarity detector is in place from birth (Gervain, Macagno, Cogoi, Pena, & Mehler, 2008) and a detector for differences might develop later around 6–7 month old, as suggested by our recent findings (Radulescu, Wijnen, Avrutin, & Gervain, (2019). As hypothesized by our model, sensitivity to entropy encompasses both a sensitivity to similar (or identical) features, and also a sensitivity to differences, thus these should be developmentally available for the sensitivity to entropy to be fully fledged.

The natural follow-up question would then be if differences in rule induction across developmental stages could be explained by variations in *channel capacity*, as hypothesized by our model. *Channel capacity*, our finite time-dependent entropy-processor, is hypothesized to increase with age, as cognitive capacities mature, and thus reduce the need to move to a higher-order category-based form of encoding. Infants are expected to have a higher tendency to make *category-based generalizations* compared to adults, when exposed to the same *input entropy*, due to their having a lower *channel capacity* than the adults. Indeed, such hypotheses have long been put forward (e.g. the "Less-is-More" hypothesis) in order to suggest an important role of perceptual and memory constraints on rule induction (Endress & Bonatti, 2007; Newport, 1990). Furthermore, these cognitive capacities mature in time, so there should be differences between developmental stages: it is an obvious truth that children outperform adults at language learning even though their non-linguistic cognitive capacities are yet to develop. Research also showed that adults are more likely to reproduce the statistical properties in their input, while children turn the statistical specificity into general rules (Hudson Kam & Newport, 2005, 2009). The same authors suggest that it is an interaction of age and input properties that leads to generalization. However, as these researchers also pointed out, it is not age *per se*, but the cognitive abilities that mature with age, and therefore memory was proposed as a factor. We also consider this interaction to be key to the mechanisms of generalization, as children are more likely than adults to "forget" the statistical specificity of the input and abstract away from it. But it is still not clear if it is both perceptual and memory constraints, and what memory component is at stake. Our model gives a more refined and formal approach to such hypotheses formulated in the psychology literature, and it makes the connection in information-theoretic terms between behavioral evidence found in psychological research and current developments and hypotheses formulated in neurobiology regarding the essential role of memory transience ("forgetting") in overfitting vs generalization design features of the memory system (Richards & Frankland, 2017) and converging views from neural networks research (Kumaran et al., 2016).

The results presented in this paper point to a ceiling effect of input entropy on rule induction, which is the result of the brain's finite encoding power, captured by the *channel capacity* factor in our model. We hypothesize that the encoding power varies according to individual differences in (unintentional) incidental memory and a general pattern-recognition capacity. We have already found evidence for a negative effect of incidental memorization and a positive effect of a visual pattern-recognition capacity on rule induction (Radulescu, Giannopoulou, Wijnen, & Avrutin, (2019).

Further research should be conducted to investigate the suitability and feasibility of entropy as a quantitative measure of input complexity and of learners' uncertainty (i.e. surprise) in rule induction, and also to assess the generalizability of this model to more complex non-repetition grammars. As suggested by previous studies (Endress, Dehaene-Lambertz, & Mehler, 2007; Endress, Nespor, & Mehler, 2009) a low-level perceptual identity-detector ("repetition detector"), which is in place from birth (Gervain, Berent, & Werker, 2012; Gervain et al., 2008), might aid learning of repetition-based grammars. Indeed, we assume that our entropy model is generalizable to all grammars and further investigations are needed to probe its implementation and feasibility. In a recent study on non-adjacent dependencies learning that extends and refines previous findings by Gómez (2002) we found that the mere set size of items was not the only factor to drive generalization, but it was the specific pattern of variability captured by input entropy, as predicted by our entropy model (Radulescu, Grama, Avrutin, & Wijnen, (2019).

As suggested before (Gerken, 2010), the human brain is not sensitive to the mere number of items or to their frequencies, but to a specific pattern of variability. We have shown in our experiments and in the reinterpretation of previous studies (section 3) that entropy captures this pattern. This result adds to a growing body of evidence showing that human language processing is sensitive to entropy (Baayen, Feldman, & Schreuder, 2006; Milin, Kuperman, Kostic, & Baayen, 2009). Moreover, entropy was shown to have an effect on lexical access in unimpaired adults as well as in elderly populations and individuals with non-fluent aphasia (Van Ewijk, 2013; Van Ewijk & Avrutin, 2016). Entropy also plays an important role in other cognitive mechanisms beyond language learning, for instance in decision-making (Tversky & Kahneman, 1992) and problem-solving (Stephen et al., 2009). Entropy was used to quantify the complexity levels within neural systems (Pereda, Quiroga, & Bhattacharya, 2005), in theories on the emergence of consciousness (Tononi, 2008), and in identifying features of brain organization that underlie the emergence of cognition and consciousness (Guevara Erra, Mateos, Wennberg, & Perez Velazquez, 2007). Recent research asks the question of how encoding input entropy at a cognitive level relates to brain responses to uncertainty at a neurobiological level (Hasson, 2017).

The phenomena investigated in this study mark a qualitative developmental step in the mechanisms underpinning language learning: moving away from an item-bound learning that memorizes and produces constructions encountered in the input or with items encountered in the input, toward category-based generalization that applies abstract rules productively. By showing that it is the interaction between *input entropy* and the finite *channel capacity* that drives the *gradual* transition to an abstract-level generalization, this research fills in an important gap in the puzzle about the induction problem for language acquisition.

Acknowledgments

This work was supported by the Netherlands Organization for Scientific Research (NWO), project number 322-75-009/1116. We would like to thank the editor and two anonymous reviewers for their valuable comments on the manuscript.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Netherlands Organization for Scientific Research (NWO) [322-75-009/1116].

References

Adriaans, F., & Kager, R. (2010). Adding generalization to statistical learning: The induction of phonotactics from continuous speech. *Journal of Memory and Language*, *62*, 311–331. doi:10.1016/j.jml.2009.11.007

Aslin, R. N, & Newport, E. L. (2012). Statistical learning: from acquiring specific items to forming general rules. *Current Directions in Psychological Science*, *21*, 170–176. doi:10.1177/0963721412436806

Aslin, R. N, & Newport, E. L. (2014). Distributional language learning: mechanisms and models of category formation. *Language Learning*, *64*, 86–105.

Aslin, R. N. & Newport, E. L. (2012). Statistical Learning: From Acquiring Specific Items to Forming General Rules. Current Directions in Psychological Science 21, 170–176.Aslin, R. N. & Newport, E. L. (2014). Distributional Language Learning: Mechanisms and Models of Category Formation. Language Learning 64, 86–105. Aslin, R. N., Saffran, J., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*, 321–324. doi:10.1111/1467-9280.00063

Baayen, R. H., Feldman, L., & Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language*, *53*, 496–512.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). CELEX2 LDC96L14. Web Download. Linguistic Data Consortium, Philadelphia, USA.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, *5*(9/10), 341–345.

Boersma, P., & Weenink, D. (2014). Praat: Doing phonetics by computer [Computer program]. Version 5. 4.02, Retrieved from http://www.praat.org/

Brooks, L. R., & Vokey, J. R. (1991). Abstract analogies and abstracted grammars: Comments on Reber (1989) and Mathews et al. (1989). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *120*, 316–323.

Chambers, K., Onishi, K., & Fisher, C. (2003). Infants learn phonotactic regularities from brief auditory experience. *Cognition*, *87*, B69–B77. doi:10.1016/s0010-0277(02)00233-0

Endress, A. D., & Bonatti, L. L. (2007). Rapid learning of syllable classes from a perceptually continuous speech stream. *Cognition*, *105*, 247–299. doi:10.1016/j.cognition.2006.09.010

Endress, A. D., Dehaene-Lambertz, G., & Mehler, J. (2007). Perceptual constraints and the learnability of simple grammars. *Cognition*, *105*, 577–614. doi:10.1016/j.cognition.2006.12.014

Endress, A. D., Nespor, M., & Mehler, J. (2009). Perceptual and memory constraints on language acquisition. *Trends in Cognitive Sciences*, *13*, 348–353. doi:10.1016/j.tics.2009.05.005

Frank, M. C., & Tenenbaum, J. B. (2011). The ideal observer models for rule learning in simple languages. *Cognition*, *120*, 360–371. doi:10.1016/j.cognition.2010.10.005

Frankland, P. W., Köhler, S., & Josselyn, S. A. (2013). Hippocampal neurogenesis and forgetting. *Trends in Neurosciences*, *36*, 497–503. doi:10.1016/j.tins.2013.05.002

Frost, R. L. A., & Monaghan, P. (2016). Simultaneous segmentation and generalisation of non-adjacent dependencies from continuous speech. *Cognition*, *147*, 70–74. doi:10.1016/j.cognition.2015.11.010

Gerken, L., Dawson, C., Chatila, R., & Tenenbaum, J. (2015). Surprise! Infants consider possible bases of generalization for a single input example. *Developmental Science*, *18*, 80–89. doi:10.1111/desc.2014.18.issue-1

Gerken, L. A. (2006). Decisions, decisions: Infant language learning when multiple generalizations are possible. *Cognition*, *98*, B67–B74. doi:10.1016/j.cognition.2005.03.003

Gerken, L. A. (2010). Infants use rational decision criteria for choosing among models of their input. *Cognition*, *115* (2), 362–366. doi:10.1016/j.cognition.2010.01.006

Gerken, L. A., & Bollt, A. (2008). Three exemplars allow at least some linguistic generalizations: Implications for generalization mechanisms and constraints. *Language Learning and Development*, *4*(3), 228–248. doi:10.1080/15475440802143117

Gerken, L. A., & Quam, C. M. (2016). Infant learning is influenced by local spurious 16 generalizations. *Developmental Science*. doi:10.1111/desc.12410

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*, *24*(3), 564–574. doi:10.1162/jocn_a_00157

Gervain, J., Macagno, F., Cogoi, S., Pena, M., & Mehler, J. (2008). The neonate brain detects speech structure. Proceedings of the National Academy of Sciences, United States of America, 105, 14222–14227.

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science*, *13*(5), 431–436. doi:10.1111/1467-9280.00476

Gómez, R. L., & Gerken, L. A. (2000). Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences*, *4*, 178–186. doi:10.1016/S1364-6613(00)01467-4

Griffiths, T.L., & Tenenbaum, J.B. (2007). From mere coincidences to meaningful discoveries. Cognition *103*(2), 180–226. Guevara Erra, R., Mateos, D. M., Wennberg, R., & Perez Velazquez, J. *L.* (2016). Statistical mechanics of consciousness: Maximization of information content of network is associated with conscious awareness. Physical Review, E94(5)Physical Review, E, 94(5–1), 52402..

Hardt, O., Nader, K., & Wang, Y.-T. (2013). GluA2-dependent AMPA receptor endocytosis and the decay of early and late long-term potentiation: Possible mechanisms for forgetting of short- and long-term memories. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*, 20130141. doi:10.1098/rstb.2013.0141

Hasson, U. (2017). The neurobiology of uncertainty: Implications for statistical learning. *Philosophical Transactions of the Royal Society B*, *372*, 20160048. doi:10.1098/rstb.2016.0048

Hawkins, D. M. (2004). The problem of overfitting. *Journal of Chemical Information and Computer Science*, *44*, 1–12. doi:10.1021/ci0342472

Hudson Kam, C., & Newport, E. L. (2005). Regularizing unpredictable variation. *Language Learning and Development*, *1*, 151–195. doi:10.1080/15475441.2005.9684215

Hudson Kam, C., & Newport, E. L. (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology*, *59*, 30–66. doi:10.1016/j.cogpsych.2009.01.001

Hudson Kam, C. L., & Chang, A. (2009). Investigating the cause of language regularization in adults: Memory constraints or learning effects? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 815–821. doi:10.1037/a0015097

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2012). The goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. *PloS One*, *7*(5), e36399. doi:10.1371/journal.pone.0036399

Knowlton, B. J., & Squire, L. R. (1994). The information acquired during artificial grammar learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 79–91. doi:10.1037/0278-7393.20.1.79

Knowlton, B. J., & Squire, L. R. (1996). Artificial grammar learning depends on implicit acquisition of both abstract and exemplar-specific information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 169–181. doi:10.1037/0278-7393.22.1.169

Kumaran, D., Hassabis, D., & McClelland, J. L. (2016). What learning systems do intelligent agents need? Complementary learning systems theory updated. *Trends in Cognitive Sciences*, *20*, 512–534. doi:10.1016/j.tics.2016.05.004

MacKay, D. J. C. (2003). *Information theory, inference and learning algorithms*. Cambridge: Cambridge University Press.

Marcus, G. F., Vijayan, S., Rao, S. B., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science*, *283*, 77–80. doi:10.1126/science.283.5398.77

Migues, P. V., Liu, L., Archbold, G. E. B., Einarsson, E. Ö., Wong, J., Bonasia, K., … Hardt, O. (2016). Blocking synaptic removal of GluA2-containing AMPA receptors prevents the natural forgetting of long-term memories. *Journal of Neuroscience*, *36*, 3481–3494. doi:10.1523/JNEUROSCI.3333-15.2016

Milin, P., Kuperman, V., Kostic, A., & Baayen, R. H. (2009). Paradigms bit by bit: An information theoretic approach to the processing of paradigmatic structure in inflection and derivation. In J. P. Blevins & J. And Blevins (Eds.), *Analogy in grammar: Form and acquisition* (pp. 214–252). Oxford, UK: Oxford University Press.

Newport, E. L. (1990). Maturational constraints on language learning. *Cognitive Science*, *14*, 11–28. doi: 10.1207/s15516709cog1401_2

Newport, E. L. (2016). Statistical language learning: Computational, maturational, and linguistic constraints. *Language and Cognition*, *8*, 447–461. doi:10.1017/langcog.2016.20

Pereda, E., Quiroga, R., & Bhattacharya, J. (2005). Nonlinear multivariate analysis of neurophysiological signals. *Progress in Neurobiology*, *77*, 1–37. doi:10.1016/j.pneurobio.2005.10.003

Perruchet, P., & Pacteau, C. (1990). Synthetic grammar learning: Implicit rule abstraction or explicit fragmentary knowledge? *Journal of Experimental Psychology: General*, *119*, 264–275. doi:10.1037/0096-3445.119.3.264

Perruchet, P., & Pacton, S. (2006). Implicit learning and statistical learning: One phenomenon, two approaches. *Trends in Cognitive Sciences*, *10*, 233–238. doi:10.1016/j.tics.2006.03.006

Peterson, M. A. (2011). Variable exemplars may operate by facilitating latent perceptual organization. *Infancy*, *16*(1), 52–60. doi:10.1111/infa.2010.16.issue-1

Pothos, E. M. (2010). An entropy model for artificial grammar learning. *Frontiers in Cognitive Science*, *1*, 1–13.

Radulescu, S., Giannopoulou, E., Wijnen, F., & Avrutin, S. (2019) Cognitive constraints on rule induction. An entropy model. *Unpublished Manuscript*.

Radulescu, S., Grama, I., Avrutin, S., & Wijnen, F. (2019) Entropy drives rule induction in non-adjacent dependency learning. *Unpublished Manuscript*.

Radulescu, S., Murali, M., Wijnen, F., & Avrutin, S. (2019) Effect of channel capacity on rule induction. An entropy model. *Unpublished Manuscript*.

Radulescu, S., Wijnen, F., Avrutin, S., & Gervain, J. (2019) Same processing costs for encoding sameness and difference in 6-7-month-olds: An fNIRS study. *Unpublished Manuscript*.

Reeder, P. A., Newport, E. L., & Aslin, R. N. (2009). The role of distributional information in linguistic category formation. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st annual conference of the cognitive science society* (pp. 2564–2569). Austin, TX: Cognitive Science Society.

Reeder, P. A., Newport, E. L., & Aslin, R. N. (2013). From shared contexts to syntactic categories: The role of distributional information in learning linguistic form-classes. *Cognitive Psychology*, *66*, 30–54. doi:10.1016/j.cogpsych.2012.09.001

Richards, B. A., & Frankland, P. W. (2017). The persistence and transience of memory. *Neuron*, *94*, 1071–1084. doi:10.1016/j.neuron.2017.04.037

Rogers, T., Rakinson, D., & McClelland, J. (2004). U-shaped curves in development: A PDP approach. *Journal of Cognition and Development*, *5*, 137–145. doi:10.1207/s15327647jcd0501_14

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928. doi:10.1126/science.274.5294.1926

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423. doi:10.1002/bltj.1948.27.issue-3

Stephen, D. G., Dixon, J. A., & Isenhower, R. W. (2009). Dynamics of representational change: Entropy, action, and cognition. *Journal of Experimental Psychology: Human Perception and Performance*, *35*, 1811–1832. doi:10.1037/a0014510

Tononi, G. (2008). Consciousness and integrated information: A provisional manifesto. *The Biological Bulletin*, *215*, 216–242. doi:10.2307/25470707

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*, 297–323. doi:10.1007/BF00122574

Van Ewijk, L. (2013). Word retrieval in acquired and developmental language disorders. PhD dissertation, Utrecht University

Van Ewijk, L., & Avrutin, S. (2016). Lexical access in non-fluent aphasia: A bit more on reduced processing. *Aphasiology*, *30*(12), 1275–1282. doi:10.1080/02687038.2015.1135867

Vokey, J. R., & Higham, P. A. (2005). Abstract analogies and positive transfer in artificial grammar learning. *Canadian Journal of Experimental Psychology*, *59*(1), 54–61. doi:10.1037/h0087461

Appendix

Familiarization items. Experiment 1.

High Entropy			Medium Entropy			Low Entropy		
Phase 1	Phase 2	Phase 3	Phase 1	Phase 2	Phase 3	Phase 1	Phase 2	Phase 3
xoxoʃi	xoxoʊu	xoxoʊu	xoxoʃi	xoxoka:	xoxoke:	xoxoʃi	xoxoʃi	xoxoke:
pypydy	pypyʃø:	pypysa:	pypyʃi	pypyka:	pypysa:	pypyʃi	pypysa:	pypysa:
tø:tø:sa:	Tø:tø:by	tø:tø:by	tø:tø:dy	tø:tø:my	tø:tø:sa:	tø:tø:ʃi	tø:tø:dy	tø:tø:ʃi
ve:ve:fø:	ve:ve:mo	ve:ve:da:	ve:ve:dy	ve:ve:my	ve:ve:my	ve:ve:ʃi	ve:ve:ʃi	ve:ve:ʃi
ʊoʊomo	ʊoʊofa:	ʊoʊoke:	ʊoʊosa:	ʊoʊoɣo	ʊoʊoɣo	ʊoʊody	ʊoʊody	ʊoʊosa:
loloke:	Loloko	lolody	lolosa:	loloɣo	loloɣo	lolody	loloʃi	loloʃi
xuxuʃu	xuxumø:	xuxuʃi	xuxufø:	xuxuʃi	xuxuʃi	xoxody	xoxosa:	xoxody
hø:hø:ʊø:	hø:hø:xi	hø:hø:ko	hø:hø:fø:	hø:hø:ʃi	hø:hø:ʃi	pypydy	pypydy	pypydy
jyjyfi	jyjyzy	jyjyʃø:	jyjymo	jyjydy	jyjydy	tø:tø:sa:	tø:tø:sa:	tø:tø:dy
ninika:	ninify	ninifø:	ninimo	ninidy	ninifø:	ve:ve:sa:	ve:ve:dy	ve:ve:dy
roromy	rorobo	roromo	roroke:	rorosa:	rorofø:	ʊoʊosa:	ʊoʊoʃi	ʊoʊoʃi
vyvyɣo	vyvyhy	vyvyfa:	vyvyke:	vyvysa:	vyvymo	lolosa:	lolosa:	lolosa:
ha:ha:ʊu	ha:ha:ʃi	ha:ha:mø:	xoxoʃu	xoxofø:	xoxoʊø:	xoxofø:	xoxofø:	xoxosa:
hihiʃø:	hihidy	hihizy	pypyʃu	pypyfø:	pypyfi	pypyfø:	pypyke:	pypyke:
jijiby	jijisa:	jijixi	tø:tø:ʊø:	tø:tø:mo	tø:tø:fi	tø:tø:fø:	tø:tø:mo	tø:tø:fø:
jujuda:	jujufø:	jujufy	ve:ve:ʊø:	ve:ve:mo	ve:ve:mo	ve:ve:fø:	ve:ve:fø:	ve:ve:fø:
jø:jø:fa:	jø:jø:da:	jø:jø:bo	ʊoʊofi	ʊoʊoke:	ʊoʊoke:	ʊoʊomo	ʊoʊomo	ʊoʊoke:
liliko	lilike:	lilihy	lolofi	loloke:	lolody	lolomo	lolofø:	lolofø:
lylymø:	lylyʃu	lylyʃu	xuxuka:	xuxuʃu	xuxuʃu	xoxomo	xoxoke:	xoxomo
nonoxi	nonoʊø:	nonoʊø:	hø:hø:ka:	hø:hø:ʃu	hø:hø:ʃu	pypymo	pypymo	pypymo
nunuzy	nunufi	nunufi	jyjymy	jyjyʊø:	jyjyʊø:	tø:tø:ke:	tø:tø:ke:	tø:tø:mo
ryryfy	ryryka:	ryryka:	ninimy	niniʊø:	ninika:	ve:ve:ke:	ve:ve:mo	ve:ve:mo
vivibo	vivimy	vivimy	roroɣo	rorofi	roroka:	ʊoʊoke:	ʊoʊofø:	ʊoʊofø:
vø:vø:hy	vø:vø:ɣo	vø:vø:ɣo	vyvyɣo	vyvyfi	vyvymy	loloke:	loloke:	loloke:

Test items. Experiment 1.

Test 1		Test 2		Test 3		Final Test		
Familiar-syllable XXY	xoxoʃi	Familiar-syllable XXY	ve:ve:mo	Familiar-syllable XXY	ʊoʊoke:	Familiar-syllable XXY	pypysa:	lolody
New-syllable XYZ	dova:sø:	New-syllable XYZ	rø:luxe:	New-syllable XYZ	mita:zu	New-syllable XYZ	fuse:bi	kø:sodo
New-syllable XXY	pø:pø:de:	New-syllable XXY	totosy	New-syllable XXY	vovofo	New-syllable XXY	ʊa:ʊa:zø:	xø:xø:ki
Familiar-syllable XYZ	tø:dysa:	Familiar-syllable XYZ	pyʊofø:	Familiar-syllable XYZ	loxomo	Familiar-syllable XYZ	ve:dyʊo	tø:ve:ke:

Familiarization items. Experiment 2.

High Entropy Phase 1/2/3	Medium Entropy Phase 1/2/3	Low Entropy Phase 1/2/3
ke:ke:my	ke:ke:my	ke:ke:my
jujuɣo	jujuɣo	jujuɣo
da:da:li	da:da:li	da:da:li
pypyve:	pypyve:	pypyve:
tø:tø:rø:	tø:tø:rø:	tø:tø:rø:
hihisa:	hihisa:	hihisa:
fofoʃu	fofoʃu	fofoʃu
nonoʃø:	nonoʃø:	ke:ke:my
nunuvø:	nunuvø:	jujuɣo
kykyʊa:	kykyʊa:	da:da:li
jø:jø:vi	jø:jø:vi	pypyve:
totomø:	totomø:	tø:tø:rø:
ha:ha:vy	ha:ha:vy	hihisa:
fyfyʃi	fyfyʃi	fofoʃu
dodoɣø:	da:da:mø:	ke:ke:my
bybyro	pypyvy	jujuɣo
bibimo	tø:tø:ʃi	da:da:li
kikiɣu	hihimy	pypyve:
fifizy	fofoɣo	tø:tø:rø:
fufuʊø:	nonoli	hihisa:
hø:hø:ʊo	nunuve:	fofoʃu
ka:ka:zø:	kykyrø:	ke:ke:my
kø:kø:lu	jø:jø:sa:	jujuɣo
boboɣe:	totoʃu	da:da:li
de:de:va:	ha:ha:ʃø:	pypyve:
hyhysø:	fyfyvø:	tø:tø:rø:
fa:fa:ly	ke:ke:ʊa:	hihisa:
jyjyxi	jujuvi	fofoʃu

Test items. Experiment 2.

Test 1		Test 2		Test 3		Final Test		
Familiar-syllable XXY	da:da:li	Familiar-syllable XXY	hihisa:	Familiar-syllable XXY	ke:ke:my	Familiar-syllable XXY	tø:tø:rø:	jujuɣo
New-syllable X1X2Y	poxa:ru	New-syllable X1X2Y	runyni	New-syllable X1X2Y	xa:misy	New-syllable X1X2Y	syniny	mininy
New-syllable XXY	dydyta:	New-syllable XXY	zuzuvo	New-syllable XXY	sosory	New-syllable XXY	jijifø:	ʊuʊuse:
Familiar-syllable X1X2Y	juda:sa:	Familiar-syllable X1X2Y	pytø:my	Familiar-syllable X1X2Y	ke:fove:	Familiar-syllable X1X2Y	hida:rø:	tø:pyɣo