Patterns bit by bit. An Entropy Model for Rule Induction

Silvia Rădulescu, Frank Wijnen, Sergey Avrutin

UiL-OTS, Utrecht University

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 triggers 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 triggered by the discrepancy between input complexity (entropy) and the finite encoding power of the human brain (channel capacity)*. In two artificial grammar experiments with adults we probed the effect of input entropy on rule induction. Results showed that as the input entropy increases, the tendency to infer abstract rules increases gradually.

1. Introduction

1.1 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. However, 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 as such 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 features¹ 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 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

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

thoroughly understand language acquisition is how learners converge on these higher-order *category-based generalizations*.

1.2 Statistical learning vs. algebraic rules

A traditional 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 (Saffran, Aslin & Newport, 1996; Aslin, Saffran & Newport, 1998), 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 (Marcus, Vijayan, Rao, & Vishton, 1999; Endress & Bonatti, 2007) 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 manipulates 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 consider 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?

1.3 Rule induction in infants

Gerken (2006) took a step towards 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, Gerken’s account is incomplete. Gerken argues that only the second group had two equally plausible generalizations at hand, but we consider 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 very much 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 (“*leledi*”) 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 (Gomez, 2002; Gerken & Boltt, 2008) 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?

1.4 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, Aslin & Newport (2009; 2013). In a series of eight artificial language experiments (Exp. 1-4, 5A-5D), adults were trained on 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,

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

participants were tested on the withheld (novel) grammatical strings, plus 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 et al. (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 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.

2. An Entropy Model for Linguistic Generalization

2.1. 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 predicts that *rule induction is an encoding mechanism triggered by the interaction between input complexity (entropy) and the limited 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 trigger the inductive leap from *item-bound* to *category-based generalization* are *input complexity (entropy)* and *the limited encoding power of the human brain (channel capacity)*.

This model thus specifies a quantitative measure for the likelihood of taking this inductive leap by capturing the specific pattern of variable input interacting with cognitive constraints.

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)³. Entropy is used in our model to describe a property of the input – *complexity*, 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 $p=0.25$, $H = 2$). 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.

³ For a random variable X , with n values $\{x_1, x_2 \dots x_n\}$, 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.

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.

*Channel capacity*⁴, describes the amount of entropy that can be processed per unit of time (Shannon, 1948). *Channel capacity* is employed here to model the limitations of the information encoding system, i.e. the limited rate of information encoding. Intuitively, the capacity to encode specific items and relations between them is limited. Thus, depending on the degree of input complexity and the available encoding power (i.e. channel capacity), different forms of information encoding are necessary to encode the complexity of a given input.

2.2. 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 limited 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. Low *input complexity*, which does not exceed *channel capacity*, facilitates encoding of specific items and enables *item-bound generalizations*, which are based on memorizing specific physical items and their perceptual features. This would be the case for rules like “*end in di*”, or rules specifying what specific items would follow each other (e.g. *ba* or *ge* follows *wo*).

⁴ *Channel 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 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.

- b. If the *input complexity (entropy)* exceeds the available *channel capacity* of the encoding system, i.e. the inflow of information exceeds the encoding power, our encoding system cannot encode each and every item and specific features. If the *channel capacity* is limited at a certain amount of entropy, then to encode an input that has a higher entropy than the available channel capacity, the new form of encoding needs to reduce the number of specific features that individual items are coded for, to group them in abstract categories, and to encode relations between these categories. This abstract form of encoding allows for higher *input entropy* to be encoded using the same available *channel capacity*, thus yielding higher-level *category-based generalizations*. 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.
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 limited encoding power of the human mind. We hypothesize that it is modulated by (unintentional) incidental memory capacity, attention and a general pattern-recognition capacity.

2.3. Application of the Model to AGL

Given that entropy is defined as a property of a variable⁵, 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 forms 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*,

⁵ A variable X is a set of x values, where x ranges from $\{0, x_1, x_2 \dots x_n\}$.

wi, *ji*, *de*, etc. Each category of bigrams and trigrams forms 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 training 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 (Perruchet and Pacteau, 1990; Knowlton and Squire, 1994). Studies also showed that performance is predicted by the frequency of these chunks (Knowlton and Squire, 1994). There is also evidence for transfer of the knowledge to novel chunks, based on abstract analogy to the specific training items (Brooks and Vokey, 1991; Vokey and 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 strings⁶. 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 training 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. 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

⁶ The author provides a method for calculating entropy of every test string based on the training items. There are several inconsistencies in his method and his model has a few implementation issues. However, we will not discuss these issues here, because they do not fall under the scope of this paper.

display the training stimulus sets for the two conditions tested by Gerken (2006), plus additional entropy calculations as per the entropy formula presented in this paper. The entropy values include bigram and trigram entropy, as well as the average bigram entropy ($H[\text{bigram}]$), the average trigram entropy ($H[\text{trigram}]$)⁷. 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*”.

Diagonal condition
[A A B] le le di wi wi je ji ji li de de we
Entropy values $H[bA] = -[(p(le)*\log_2 p(le)) + (p(wi)*\log_2 p(wi)) + (p(ji)*\log_2 p(ji)) + (p(de)*\log_2 p(de))]= 2$ $H[Be] = -[(p(di)*\log_2 p(di)) + (p(je)*\log_2 p(je)) + (p(li)*\log_2 p(li)) + (p(we)*\log_2 p(we))]= 2$ $H[AA] = -[(p(lele)*\log_2 p(lele)) + (p(wiwi)*\log_2 p(wiwi)) + (p(jiji)*\log_2 p(jiji)) + (p(dede)*\log_2 p(dede))]= 2$ $H[AB] = -[(p(ledi)*\log_2 p(ledi)) + (p(wije)*\log_2 p(wije)) + (p(jili)*\log_2 p(jili)) + (p(dewe)*\log_2 p(dewe))]= 2$ $H[AAB] = -[(p(leledi)*\log_2 p(leledi)) + (p(wiwije)*\log_2 p(wiwije)) + (p(jijili)*\log_2 p(jijili)) + (p(dedewe)*\log_2 p(dedewe))]= 2$ $H[\text{bigram}] = 2 \quad H[\text{trigram}] = 2$
Table 1. Entropy values of the input in the Diagonal Condition in Gerken (2006)

Column condition
[A A B] le le di wi wi di ji ji di de de di
Entropy values

⁷ 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.

$H[bA] = - [(p(le)*\log_2 p(le)) + (p(wi)*\log_2 p(wi)) + (p(ji)*\log_2 p(ji)) + (p(de)*\log_2 p(de))]= 2$ $H[Be] = - [p(di)*\log_2 p(di)]= 0$ $H[AA] = - [(p(lee)*\log_2 p(lee)) + (p(wiwi)*\log_2 p(wiwi)) + (p(jiji)*\log_2 p(jiji)) + (p(dede)*\log_2 p(dede))]= 2$ $H[AB] = - [(p(ledi)*\log_2 p(ledi)) + (p(widi)*\log_2 p(widi)) + (p(jidi)*\log_2 p(jidi)) + (p(dedi)*\log_2 p(dedi))]= 2$ $H[AAB] = - [(p(leledi)*\log_2 p(leledi)) + (p(wiwiidi)*\log_2 p(wiwiidi)) + (p(jijidi)*\log_2 p(jijidi)) + (p(dededi)*\log_2 p(dededi))]= 2$
$H[\text{bigram}] = 1.5 \quad H[\text{trigram}] = 2$
Table 2. Entropy values of the input in the Column Condition in Gerken (2006)

An entropy-based reinterpretation of the results by Reeder, Aslin & Newport (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 too vague 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.

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

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 modulates *channel capacity*⁸). 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 towards *category-based generalizations*.

⁸ Recall *channel capacity* quantifies the amount of entropy that can be processed per unit of time.

4. 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*⁹. The following predictions were tested:

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

To test these hypotheses, we designed several versions of the same artificial grammar (3-syllable XXY structure¹⁰) in order to expose participants to different input entropy in three groups: high, medium and low entropy. An ensuing test phase presented participants with a grammaticality judgement task, where they were asked if they accepted the test strings as possible in the training language. The test included four test string types that were designed to test each type of rule induction, as presented below.

⁹ We ignored variations in channel capacity here.

¹⁰ An XXY pattern describes strings consisting of two identical syllables (XX) followed by another different syllable (Y): e.g. *goo_goo_sjie*, *puu_puu_saa*, *joe_joe_feu*

Type_1: trained_syll_XXY (XXY structure with trained X-syllables and trained Y-syllables) – **correct answer: [ACCEPT]**¹¹ – this is a positive test case¹², which is intended to check learning of the trained 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.

Type_2: untrained_syll_XYZ (XYZ structure with untrained syllables) – **correct answer: [REJECT]** – this is the complementary negative test case, which is intended to check learning of the trained strings and string pattern. It is designed to back up and complement results for Type_1 as follows: if the forms of encoding – either ITEM¹³ or CATEG – trigger acceptance of trained XXY strings, then they should trigger rejection of the negative test cases (untrained 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 XXY trained cannot be valid.

Type_3: untrained_syll_XXY (XXY structure with untrained syllables) – **correct answer: [ACCEPT]** – this is a positive test case, which 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

¹¹ [ACCEPT] means the Yes button is to be pressed, meaning the participant considers the test item as a possible string in the training language; [REJECT] means the No button is to be pressed, coding for the participant's rejection of the test item as a possible string in the training language.

¹² A positive test case is designed to check the behavior of a system in the intended working case, i.e. if a system has been trained on XXY strings, the positive test will check the system's behavior on XXY strings. A negative test case is designed to check the behavior of a system in a different (or opposite) situation than the intended one, i.e. if a system is trained on XXY strings, the negative test will check the system's behavior on XYZ strings, or XYY strings, etc. For the validity of the results and their interpretation, it is crucial to have a perfect balance between positive and (complementary) negative test cases in order to prevent biases in participants' tendencies to accept any test items, and also in order to strengthen the evidence provided by positive test cases.

¹³ For ease of presentation, *item-bound generalization* is denoted ITEM, and *category-based generalization* – CATEG).

compared against the mean rates of acceptance for the trained 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 **Type_1: trained_syll_XXY** and **Type_3: untrained_syll_XXY**, i.e. the rates of acceptance of these test types will be significantly different. A strong discrimination between these test types (**Type_1: trained_syll_XXY** significantly more accepted than **Type_3: untrained_syll_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 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.

Type_4: trained_syll_XYZ (XYZ structure with trained syllables¹⁴) – **correct answer: [REJECT]** – this is the negative test case complementary to Type 3: if the new XXY strings are accepted in a different proportion due to differences in types of encoding developed, then the negative test case (XYZ with trained syllables) 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 trained syllables drives acceptance of these ungrammatical strings with trained syllables. Hence differences in performance are expected across conditions, depending on the extent to which ITEM and CATEG are developed: 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 flash out clear mismatches

¹⁴ A subset of the syllables used in training were concatenated to create XYZ test strings with trained syllables. Any of the X-syllables and Y-syllables were randomly assigned to the X, Y or Z slot of the XYZ pattern.

between trained and new strings 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 trained 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.

5. Experiment 1

5.1. Method

5.1.1. 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.

5.1.2. Training 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. *goo* [xo:], *sjie* [ʃ i:]). 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 database¹⁵, 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 Praat¹⁶.

The experiment consisted of three exposure phases with intermediate test phases, followed by a final test phase, as per the timeline below:

In the exposure phases, a total of 72 XXY strings were presented, 24 per separate 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 (HiEn_Exp1), Medium Entropy (MedEn_Exp1) and Low Entropy (LowEn_Exp1).

5.2. Entropy values of training 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 the entropy formula¹⁷, three different values for *input complexity* were obtained, as follows:

¹⁵ Baayen, R, R Piepenbrock, and L Gulikers. CELEX2 LDC96L14. Web Download. Philadelphia: Linguistic Data Consortium, 1995.

¹⁶ www.praat.org

¹⁷ The same entropy formula that was presented in previous sections was used here: $H(X) = -\sum_{i=1}^n p(x_i) \log p(x_i)$

1. **LowEn_Exp1:** 4 * 6 X-syllables / 4 * 6 Y-syllables (each syllable was used 4 times in each training phase). To generate the XXY strings for LowEn_Exp1 condition, all 6 XX pairs were concatenated with all 6 Y-syllables, but different subsets (consisting of 24 XX_Y combinations) were used for each training phase. The same procedure was applied to the other conditions. All three training phases had the same entropy values: the average bigram entropy ($H[\text{bigram}]$) was 3.08, the average trigram entropy ($H[\text{trigram}]$) was 3.91, and the total average entropy ($H[\text{total}]$) was 3.5 (see Table 4 for complete entropy calculations).

2. **MedEn_Exp1:** 2 * 12 X-syllables / 2 * 12 Y-syllables; 6 different X-syllables and 6 different Y-syllables were added to the syllables used for LowEn_Exp1; each syllable was used 2 times. All three training phases had the same entropy values: the average bigram entropy ($H[\text{bigram}]$) was 3.83, the average trigram entropy ($H[\text{trigram}]$) was 4.25, and the total average entropy ($H[\text{total}]$) was 4.

3. **HiEn_Exp1:** 1 * 24 X-syllables / 1 * 24 Y-syllables; 12 X-syllables and 12 Y-syllables were added to those used for MedEn_Exp1; each syllable was used one time. All three training phases had the same entropy values: the average bigram entropy ($H[\text{bigram}]$) was 4.58, the average trigram entropy ($H[\text{trigram}]$) was 4.58, and the total average entropy ($H[\text{total}]$) was 4.58.

HiEn_Exp1	MedEn_Exp1	LowEn_Exp1
$H[bX] = H[24] = -\Sigma[0.042 * \log 0.042] = 4.58$ $H[XX] = H[24] = 4.58$ $H[XY] = H[24] = 4.58$ $H[Ye] = H[24] = 4.58$ $H[bXX] = H[XXY] = H[XYe] = H[24] = 4.58$ $H[\text{bigram}] = 4.58$ $H[\text{trigram}] = 4.58$ $H[\text{total}] = \frac{H[\text{bigram}] + H[\text{trigram}]}{2} = 4.58$	$H[bX] = H[12] = -\Sigma[0.083 * \log 0.083] = 3.58$ $H[XX] = H[12] = 3.58$ $H[XY] = H[24] = 4.58$ $H[Ye] = H[12] = 3.58$ $H[bXX] = H[12] = 3.58$ $H[XXY] = H[XYe] = H[24] = 4.58$ $H[\text{bigram}] = 3.83$ $H[\text{trigram}] = 4.25$ $H[\text{total}] = 4$	$H[bX] = H[6] = -\Sigma[0.167 * \log 0.167] = 2.58$ $H[XX] = H[6] = 2.58$ $H[XY] = H[24] = 4.58$ $H[Ye] = H[6] = 2.58$ $H[bXX] = H[6] = 2.58$ $H[XXY] = H[XYe] = H[24] = 4.58$ $H[\text{bigram}] = 3.08$ $H[\text{trigram}] = 3.91$ $H[\text{total}] = 3.5$
Table 4. Entropy values for Experiment 1		

5.3. 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 instructed 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 training phases. After each training 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.

5.4. 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).

6. Experiment 1: Results

In order to test the effect of *input complexity* on generalization, the HiEn_Exp1, MedEn_Exp1 and LowEn_Exp1 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, Test¹⁸ 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 x Entropy condition interaction ($F(9, 679) = 1.243, p = .26$).

Fig. 1 presents the mean rate of acceptance (percentage of acceptances per group) across conditions for Type_1: trained_syll_XXY and Type_3: untrained_syll_XXY. The mean acceptance rate of Type_3: untrained_syll_XXY in HiEn_Exp1 was 80% (Mean = .80, SD = .403), in MedEn_Exp1 was 73% (Mean = .73, SD = .446), and in LowEn_Exp1 was 65% (Mean = .65, SD = .480). One-sample Wilcoxon Signed-Rank tests indicated a statistically significant above-chance mean acceptance for Type_3 in HiEn_Exp1 ($Z = 4.648, SE = 118.12, p = .000$; Cohen's effect size $d = 0.6$), in MedEn_Exp1 ($Z = 3.615, SE = 118.12, p = .000$; Cohen's effect size $d = 0.47$), and in LowEn_Exp1 ($Z = 2.292, SE = 103.82, p = .022$, Cohen's effect size $d = 0.31$). In HiEn_Exp1 there was a significant difference between acceptance of Type_1 and acceptance of Type_3 ($M=.167, SD=.376; t(3) = 2.721, SE=0.853, p = .007$); in MedEn_Exp1 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 LowEn_Exp1 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 on these tests in HiEn_Exp1 vs. LowEn_Exp1 were higher than

¹⁸ The intermediate tests and the final test in all entropy groups.

the same values for HiEn_Exp1 vs. MedEn_Exp1 ($d = 0.15$, $r = 0.07$), and also higher than the same values for LowEn_Exp1 vs. MedEn_Exp1 ($d = 0.21$, $r = 0.1$).

Fig. 2 shows the mean rate of rejection for Type_2: untrained_syll_XYZ & Type_4: trained_syll_XYZ. The mean rejection of Type_4: trained_syll_XYZ in HiEn_Exp1 was 82% (Mean = .82, SD = .39), significantly different from the mean acceptance of Type_1 ($t(3) = 2.529$, $SE = 0.851$, $p = .012$); 77% in MedEn_Exp1 (Mean = .77, SD = .427), significantly different from the mean acceptance of Type_1 ($t(3) = 3.147$, $SE = 0.837$, $p = .002$); and 91% in LowEn_Exp1 (Mean = .91, SD = .290), near-significantly different from the mean acceptance of Type_1 ($t(3) = 1.683$, $SE = 1.185$, $p = .093$).

7. 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. trained 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 trained XXY, we conclude that they formed a *category-based generalization* (XXY) which applies equally to both trained 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 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 trained syllables is very high in the low entropy group, although the rate of acceptance for new XXY strings is rather low (Fig. 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 trained 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 trained syllables, as a function of increased input entropy. Similar tendencies towards 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.

The results showed that the decreasing trend of the rejection of trained 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 trained XYZ is the intersection point of the decreasing trend of XYZ and the increasing trend of XXY. The calculated intersection point of the two trends – $y(XXY_new) = y(XYZ_trained)$ – is $H = 4.2$ ($y = 0.72$), which allows the prediction that the rate of rejection of XYZ_{trained} 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 Trained_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 trained 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 XYZ_{trained} 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 trained X-syllables and Y-syllables. An ANOVA with training group (HiEn_Exp1, MedEn_Exp1 and LowEn_Exp1) 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 (HiEn_Exp1: Mean[X_1X_2Y] = .81, Mean[non- X_1X_2Y] = .83, $F(1,58) = .072$, $p = .79$; MedEn_Exp1: Mean[X_1X_2Y] = .79, Mean[non- X_1X_2Y] = .73, $F(1,58) = .293$, $p = .59$; LowEn_Exp1: 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 training 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 it 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 (presumably because of the low number of data points per each test and because of a very short total exposure time – about 4 minutes), 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.

8. 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 LowEn_Exp2 condition we chose a lower entropy value than for LowEn_Exp1 (2.8 bits - 4 x 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 MedEn_Exp2 condition (4.25 bits - 2 x 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 trained syllables will decrease as compared to the performance for MedEn_Exp1 (for $H=4$ bits the performance was 77%): at $H=4.2$ bits the mean performance predicted is 72%. For the HiEn_Exp2 condition we chose a higher entropy (4.8 bits - 1 x 28 Xs / 1 x 28 Ys) than HiEn_Exp1 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, i.e. there will be no further increase in the tendency to make category-based generalizations.

8.1. Method

8.1.1. 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.

8.1.2. Training 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 (HiEn_Exp2, MedEn_Exp2, LowEn_Exp2) had equal number of training strings – 84 XXY strings in total (28 XXY strings in each training 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.

8.2. Entropy values of training conditions

The entropy formula¹⁹ was applied in the same manner as for Experiment 1 to obtain other three different values for *input complexity*, as follows:

1. **LowEn_Exp2:** 4 * 7 X-syllables / 4 * 7 Y-syllables (each syllable was used 4 times in each training phase). To generate the XXY strings for the LowEn_Exp2 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 training phases. The same procedure was applied to the other conditions. All three training phases had the same entropy values: the average bigram entropy ($H[\text{bigram}]$) was 2.8, the average trigram entropy ($H[\text{trigram}]$) was 2.8, and the total average entropy ($H[\text{total}]$) was 2.8 (see Table 5 for complete entropy calculations).

2. **MedEn_Exp2:** 2 * 14 X-syllables / 2 * 14 Y-syllables; 7 different X-syllables and 7 different Y-syllables were added to those used for LowEn_Exp2; each syllable was used 2 times. All three training phases had the same entropy values: the average bigram entropy ($H[\text{bigram}]$) was 4.05, the average trigram entropy ($H[\text{trigram}]$) was 4.46, and the total average entropy ($H[\text{total}]$) was 4.25.

3. **HiEn_Exp2:** 1 * 28 X-syllables / 1 * 28 Y-syllables; 14 X-syllables and 14 Y-syllables were added to those used for MedEn_Exp2; each syllable was used one time. All three training phases had the same entropy values: the average bigram entropy ($H[\text{bigram}]$) was 4.8), the average trigram entropy ($H[\text{trigram}]$) was 4.8, and the total average entropy ($H[\text{total}]$) was 4.8.

These values were different from the values in the entropy conditions used in Experiment

1 ($H[\text{total}]_{\text{HiEn}} = 4.58$, $H[\text{total}]_{\text{MedEn}} = 4$, $H[\text{total}]_{\text{LowEn}} = 3.5$).

HiEn_Exp2	MedEn_Exp2	LowEn_Exp2
$H[\text{bX}] = H[28] = 4.8$	$H[\text{bX}] = H[14] = 3.8$	$H[\text{bX}] = H[7] = 2.8$
$H[\text{XX}] = H[28] = 4.8$	$H[\text{XX}] = H[14] = 3.8$	$H[\text{XX}] = H[7] = 2.8$
$H[\text{XY}] = H[28] = 4.8$	$H[\text{XY}] = H[28] = 4.8$	$H[\text{XY}] = H[7] = 2.8$
$H[\text{Ye}] = H[28] = 4.8$	$H[\text{Ye}] = H[14] = 3.8$	$H[\text{Ye}] = H[7] = 2.8$
$H[\text{bXX}] = H[28] = 4.8$	$H[\text{bXX}] = H[14] = 3.8$	$H[\text{bXX}] = H[7] = 2.8$
$H[\text{XXY}] = H[\text{XYe}] = H[28] = 4.8$	$H[\text{XXY}] = H[\text{XYe}] = H[28] = 4.8$	$H[\text{XXY}] = H[\text{XYe}] = H[7] = 2.8$
$H[\text{bigram}] = 4.8$	$H[\text{bigram}] = 4.05$	$H[\text{bigram}] = 2.8$

¹⁹ The same entropy formula that was presented in previous sections was used here: $H(X) = -\sum_{i=1}^n p(x_i) \log p(x_i)$

H[trigram] = 4.8 H[total] = 4.8	H[trigram] = 4.46 H[total] = 4.25	H[trigram] = 2.8 H[total] = 2.8
Table 5. Entropy values for Experiment 2		

8.3. Procedure The procedure was the same as for Experiment 1.

8.4. 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):

Type_1: trained_syll_XXY – correct answer: [ACCEPT]

Type_2: untrained_syll_X₁X₂Y (three different untrained syllables) – correct answer: [REJECT]

Type_3: untrained_syll_XXY – correct answer: [ACCEPT]

Type_4: trained_syll_X₁X₂Y (three different trained syllables) – correct answer: [REJECT]

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

9. Experiment 2: Results

In order to test the effect of *input complexity* on the process of generalizing, the HiEn_Exp2, MedEn_Exp2 and LowEn_Exp2 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 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 ($F(9, 699) = 5.038, p = .000$). 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 x Entropy Group interaction ($F(9, 699) = 1.163, p = .32$).

Fig. 4 shows the mean acceptance rates across conditions for Type_1: trained_syll_XXY and Type_3: untrained_syll_XXY. The mean rate of acceptance for Type_3: untrained_syll_XXY in HiEn_Exp2 was 80% (Mean = .80, SD = .403), for MedEn_Exp2 was 77% (Mean = .77, SD = .427), and for LowEn_Exp2 was 57% (Mean = .57, SD = .5). One-sample Wilcoxon Signed-Rank tests indicated a statistically significant above-chance mean acceptance for Type_3 in HiEn_Exp2 ($Z = 4.648$, $SE = 118.12$, $p = .000$; Cohen's effect size $d = 0.6$) and in MedEn_Exp2 ($Z = 4.131$, $SE = 118.12$, $p = .000$; Cohen's effect size $d = 0.53$), but in LowEn_Exp2 the mean acceptance was not significantly above chance ($Z = 1.033$, $SE = 118.12$, $p = .3$, Cohen's effect size $d = 0.13$). In HiEn_Exp2 there was a significant difference between acceptance of Type_1 and acceptance of Type_3 ($M = .167$, $SD = .376$; $t(3) = 2.161$, $SE = 0.643$, $p = .031$); in MedEn_Exp2 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 LowEn_Exp2 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 effect size value ($d = 0.73$) and the effect-size correlation ($r = 0.34$) for the difference between acceptance of Type_1 and acceptance of Type_3 in HiEn_Exp2 vs. LowEn_Exp2 were higher than the same values for HiEn_Exp2 vs. MedEn_Exp2 ($d = 0.09$, $r = 0.04$), and also higher than the same values for LowEn_Exp2 vs. MedEn_Exp2 ($d = 0.63$, $r = 0.3$).

Fig. 5 displays the mean rate of rejection for Type_2: untrained_syll_ X_1X_2Y & Type_4: trained_syll_ X_1X_2Y . The mean rejection rate for Type_4: trained_syll_ X_1X_2Y was 90% for HiEn_Exp2 (Mean = .90, SD = .303), not significantly different from the mean acceptance of Type_1 ($t(3) = 0.647$, $SE = 0.704$, $p = .518$); 73% for MedEn_Exp2 (Mean = .73, SD = .446), significantly different from the mean acceptance of Type_1 ($t(3) = 2.856$, $SE = 0.619$, $p = .004$); and 83% for LowEn_Exp2 (Mean = .83, SD = .376), significantly different from the mean acceptance of Type_1 ($t(3) = 2.028$, $SE = 0.711$, $p = .043$).

10. Comparing Experiment 1 and Experiment 2

To further test the effect of *input complexity* on the process of making generalizations, all the conditions from both 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 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 X Entropy condition interaction ($F(18, 1,378) = 1.150, p = .29$).

A simple linear regression was calculated (Fig. 6) to predict the rate of acceptance of new 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 XXY.

11. 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 syllables and XXY strings with trained syllables show differences between groups in terms of their tendency to generalize to new items: in HiEn_Exp2 this discrimination is lower than in MedEn_Exp2, which is lower than in LowEn_Exp2. This difference suggests that learners in the HiEn_Exp2 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 XYZ strings with trained syllables 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 mean acceptance of XXY strings with untrained 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 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 trained syllables, as the *input complexity* increases. As predicted, the mean acceptance of XXY strings with untrained syllables decreases to very close to chance level (57%) when the *input complexity* decreases to an entropy of 2.8 bits (Fig. 7). When entropy increases from 4 bits (MedEn_Exp1) to 4.2 bits (MedEn_Exp2), the mean rejection rate for X1X2Y with trained syllables decreases below 77% (the rejection rate at $H=4\text{bits}$), 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 (HiEn_Exp1) to 4.8 bits (HiEn_Exp2), 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 (Fig. 8) 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 Fig. 8, the information load (uncertainty) about the input structure (rules) is predicted to decrease logarithmically as the input entropy increases.

12. 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* (Marcus et al., 1999; Endress & Bonatti, 2007), while others proposed that *statistical learning* underlies both types of generalizations (Aslin & Newport, 2012; 2014; Perruchet & Pacton, 2006; Frost & Monaghan, 2016). 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 and other studies 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, but without clear evidence or explanation (Newport, 1999; Hudson Kam & Newport, 2005; 2009; Hudson Kam & Chang, 2009; Newport, 2016)); 4) why increased exposure to the same distribution of (in)consistent cues reduces the tendency to make *category-based generalizations*.

Our entropy model answers these questions and it accounts for both types of encoding by identifying the same 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 at 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.

In two artificial grammar experiments we investigated the effect of *input complexity* 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 towards *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, we will carry out future research to answer the following outstanding questions.

What is the effect of *input entropy* on infant rule induction? To answer this question, we will probe the effect of *input complexity* on rule induction in infants, to investigate whether the same pattern of results found in adults is replicated in infants, i.e. infants' tendency towards *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 will be exposed to a lower range of entropy than adults. Results of these experiments will provide further insights into child rule induction. 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; Peterson, 2011; Gerken & Boltt, 2008), but Gerken et al. (2015) claim that a single example suffices for generalization. Gerken et al. (2015) interpreted these results to support a Bayesian account of generalization, also suggested by Griffiths & Tennenbaum (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 consider 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 marginally investigated and noted in previous studies with adults (Reeder et al., 2013), and as explicitly predicted by the *channel capacity* component of our entropy model. Secondly, the authors claim that generalization occurred from a single example, but this interpretation runs contrary to another assumption in their argumentation, namely that the example is surprising compared to their prior model. Thus, 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 extends 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.

Can differences in rule induction across developmental stages be explained by variations in *channel capacity*? *Channel capacity* 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. This change in *channel capacity* is predicted to account for the differences between infants and adults. To test this hypothesis, we will keep *input entropy* constant and expose two populations with different *channel capacity* - infants and adults - to the same amount of entropy. 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. Previous research hypothesized the role of perceptual and memory constraints on rule induction (Newport, 1990; Endress & Bonatti, 2007). 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 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.

What is *channel capacity* in psychological terms? To answer this question, we will investigate the cognitive capacities that directly modulate *channel capacity*. 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 limited encoding power. We hypothesize that the encoding power varies according to individual differences in (unintentional) incidental memory and a general pattern-recognition capacity. We have already conducted experiments that show a negative effect of incidental memorization and a positive effect of a visual pattern-recognition capacity on rule induction (*article in prep.*).

Further research will 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. 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, Kostiç, & Baayen, 2009). In several reaction time experiments Kostiç and colleagues found that processing speed of inflected nouns and verbs in Serbian depended on the information load of the individual nouns or verbs, and on the entropy of their entire inflection class. Entropy also plays an important role beyond language, for instance in decision-making (Tversky & Kahneman, 1992) and problem-solving (Stephen, Dixon & Isenhowe, 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, 2016). 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).

Given that our model makes predictions about a domain-general rule induction mechanism, we will also investigate the effects of *input entropy* and *channel capacity* on rule induction with *visual input*. The results would be useful to investigate the domain-general application of our entropy model.

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, towards category-based generalization that applies abstract rules productively. By showing that it is the interaction between *input complexity* and the limitations of *channel capacity* that triggers the inductive leap to an abstract-level generalization, this research fills in an important gap in the puzzle about the induction problem for language acquisition.

Appendix

Training items. Experiment 1.

High Entropy			Medium Entropy			Low Entropy		
Exposure 1	Exposure 2	Exposure 3	Exposure 1	Exposure 2	Exposure 3	Exposure 1	Exposure 2	Exposure 3
googoosjie	googoowoe	googoowoe	googoosjie	googookaa	googookeee	googoosjie	googoosjie	googookeee
puupuuduu	puupuusjeu	puupuusaa	puupuusjie	puupuukaa	puupuusaa	puupuusjie	puupuusaa	puupuusaa
teuteusaa	Teuteubuu	teuteubuu	teuteuduu	teuteumuu	teuteusaa	teuteusjie	teuteuduu	teuteusjie
veeveefeu	veeveemoo	veeveedaa	veeveeduu	veeveemuu	veeveemuu	veevesjie	veevesjie	veevesjie
woowoomoo	woowoofaa	woowookeee	woowoosaa	woowoochoo	woowoochoo	woowooduu	woowooduu	woowoosaa
loolookeee	Loolookoo	loolooduu	looloosaa	looloochoo	looloochoo	loolooduu	looloosjie	looloosjie
goegoesjoe	goegoemeu	goegoesjie	goegoefeu	goegoesjie	goegoesjie	googooduu	googoosaa	googooduu
heuheuweu	heuheuchie	heuhekoo	heuheufeu	heuheusjie	heuheusjie	puupuuduu	puupuuduu	puupuuduu
juujuufie	juujuuzuu	juujuusjeu	juujuumoo	juujuuduu	juujuuduu	teuteusaa	teuteusaa	teuteuduu
nieniekaa	nieniefuu	nieniefeu	nieniemoo	nienieduu	nieniefeu	veevesaa	veeveeduu	veeveeduu
rooroomuu	roorooboo	rooroomoo	roorookeee	rooroosaa	rooroofeu	woowoosaa	woowoosjie	woowoosjie
vuuvuuchoo	vuuvuuhuu	vuuvuufaa	vuuvuukee	vuuvuusaa	vuuvuumoo	looloosaa	looloosaa	looloosaa
haahaawoe	haahaasjie	haahaameu	googoosjoe	googoofeu	googooweu	googoofeu	googoofeu	googoosaa
hiebiesjeu	hiehieduu	hiehiezuu	puupuusjoe	puupuufeu	puupuufie	puupuufeu	puupuukee	puupuukee
jiejiebuu	jiejiesaa	jiejiechie	teuteuweu	teuteumoo	teuteufie	teuteufeu	teuteumoo	teuteufeu
joejoedaa	joejoefeu	joejoefuu	veeveeweu	veeveemoo	veeveemoo	veeveefeu	veeveefeu	veeveefeu
jeujeufaa	jeujeudaa	jeujeuboo	woowoofie	woowookeee	woowookeee	woowoomoo	woowoomoo	woowookeee
lielikoo	lieliekee	lieliehuu	looloojie	loolookeee	loolooduu	looloomoo	looloofoe	looloofoe
luuluumeu	luuluusjoe	luuluusjoe	goegoekaa	goegoesjoe	goegoesjoe	googoomoo	googookeee	googoomoo
noonoochie	noonooweu	noonooweu	heuhekkaa	heuheusjoe	heuheusjoe	puupuumoo	puupuumoo	puupuumoo
noenoezoo	noenoejie	noenoejie	juujuumuu	juujuuweu	juujuuweu	teuteukee	teuteukee	teuteumoo
ruuruufuu	ruuruukaa	ruuruukaa	nieniemuu	nienieweu	nieniekaa	veeveekkee	veeveemoo	veeveemoo
vievieboo	vieviemuu	vieviemuu	roorochoo	rooroofie	roorookaa	woowookeee	woowoofeu	woowoofeu
veuveuhuu	veuveuchoo	veuveuchoo	vuuvuuchoo	vuuvuufie	vuuvuumuu	loolookeee	loolookeee	loolookeee

Test items. Experiment 1.

Test 1		Test 2		Test 3		Final Test	
type_1	googoosjie	type_1	veeveemoo	type_1	woowookeee	type_1	puupuusaa
type_2	doovaaseu	type_2	reuloegree	type_2	mietaazoe	type_2	foeseebie
type_3	peupeudee	type_3	tootoosuu	type_3	voovoofoo	type_3	waawaazeu
type_4	teuduusaa	type_4	puuwoofeu	type_4	loogoomoo	type_4	veeduuwoo

Training items. Experiment 2.

High Entropy Exposure 1/2/3	Medium Entropy Exposure 1/2/3	Low Entropy Exposure 1/2/3
keekeemuu	keekeemuu	keekeemuu
joejoegoo	joejoegoo	joejoegoo
daadaalie	daadaalie	daadaalie
puupuuee	puupuuee	puupuuee
teuteureu	teuteureu	teuteureu
hiebiesaa	hiebiesaa	hiebiesaa
fofoosjoe	fofoosjoe	fofoosjoe
noonosjeu	noonosjeu	keekeemuu
noenoeveu	noenoeveu	joejoegoo
kuukuuwaa	kuukuuwaa	daadaalie
jeujeuvie	jeujeuvie	puupuuee
tootoomeu	tootoomeu	teuteureu
haahaavuu	haahaavuu	hiebiesaa
fuufusjie	fuufusjie	fofoosjoe
doodoogeu	daadaameu	keekeemuu
buubuuroo	puupuuvuu	joejoegoo
biebiemoo	teuteusjie	daadaalie
kiekiegoe	hiehiemuu	puupuuee
fiefiezuu	fofoogoo	teuteureu
fofoeweu	noonoolie	hiebiesaa
heuheuwoo	noenoevee	fofoosjoe
kaakaazeu	kuukuureu	keekeemuu
keukeuloe	jeujeusaa	joejoegoo

booboogee deedeevaa huuhuuseu faaiaaluu juujuuchie	tootoosjoe haahaasjeu fuufuuveu keekeewaa joejoevie	daadaalie puupuuvee teuteureu hiebiesaa foofosjoe
--	---	---

Test items. Experiment 2.

Test 1		Test 2		Test 3		Final Test		
type_1	daadaalie	type_1	hie_hie_saa	type_1	kee_kee_muu	type_1	teuteureu	joejoegoo
type_2	poogaaroe	type_2	roe_nuu_nie	type_2	gaa_mie_suu	type_2	suunienuu	mienienuu
type_3	duu_duu_taa	type_3	zoe_zoe_voo	type_3	soo_soo_ruu	type_3	jiejiefeu	woewoese
type_4	joe_daa_saa	type_4	puu_teu_muu	type_4	kee_foo_vee	type_4	hiedaareu	teupuugoo

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.
- Aslin, R. N., Saffran, J., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, 9, 321–324.
- Aslin, R.N., and Newport, E. (2012). Statistical learning: From acquiring specific items to forming general rules. *Current Directions in Psychological Science*, 21, 170–176.
- Aslin, R.N., and Newport, E. (2014). Distributional Language Learning: Mechanisms and Models of Category Formation. *Language Learning*, 64, 86–105.
- 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.
- Brooks, L. R., & Vokey, J. R. (1991). Abstract analogies and abstracted grammars: comments on Reber (1989) and Mathews et al. (1989). *J. Exp. Psychol. Learn. Mem. Cogn.* 120, 316–323.
- Chambers, K., Onishi, K., & Fisher, C. (2003). Infants learn phonotactic regularities from brief auditory experience. *Cognition*, 87, B69–B77.
- Endress, A. D., & Bonatti, L. L. (2007). Rapid learning of syllable classes from a perceptually continuous speech stream. *Cognition*, 105, 247–299.
- Frank, M. C., & Tenenbaum, J. B. (2011) The ideal observer models for rule learning in simple languages. *Cognition*, 120, 360 – 371.
- Frost, R. L. A., & Monaghan, P. (2016). Simultaneous segmentation and generalisation of non-adjacent dependencies from continuous speech. *Cognition*, 147, 70–74.
- Gerken, L. A. (2006). Decisions, decisions: Infant language learning when multiple generalizations are possible. *Cognition*, 98, B67–B74.
- Gerken, LA, & 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.
- Gerken LA. (2010). Infants use rational decision criteria for choosing among models of their input. *Cognition*, 115(2), 362–366.
- 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.
- Gómez, R.L. & Gerken, L.A. (2000). Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences* 4, 178–186.

- Gómez, R.L. (2002). Variability and detection of invariant structure. *Psychological Science*, 13(5), 431 – 436.
- 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).
- Hasson U. (2017). The neurobiology of uncertainty: implications for statistical learning. *Phil. Trans. R. Soc. B* 372: 20160048. <http://dx.doi.org/10.1098/rstb.2016.0048>
- Hudson Kam, C. , & Newport , E. L . (2005). Regularizing unpredictable variation. *Language Learning and Development* , 1, 151 – 195 .
- Hudson Kam, C. , & Newport , E. L . (2009). Getting it right by getting it wrong: when learners change languages. *Cognitive Psychology*, 59, 30 – 66
- 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.
- Knowlton, B. J. & Squire, L. R. (1994). The information acquired during artificial grammar learning. *J. Exp. Psychol. Learn. Mem. Cogn.* 20, 79–91.
- Knowlton, B. J. & Squire, L. R. (1996). Artificial grammar learning depends on implicit acquisition of both abstract and exemplar-specific information. *J. Exp. Psychol. Learn. Mem. Cogn.* 22, 169–181.
- Marcus, G. F., Vijayan, S., Rao, S. B., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science*, 283, 77–80.
- Marslen-Wilson, W. D., & Tyler, L. K. (1981). Central processes in speech understanding. *Philosophical Transactions of the Royal Society of London*, B295(1077), 317–322.
- 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 Blevins, J.P. And Blevins, J. (Eds), *Analogy in grammar: Form and acquisition*, Oxford University Press, Oxford, 2009, 214–252.
- 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? *J. Exp. Psychol. General*, 119, 264–275.
- Perruchet, P. & Pacton, S. (2006). Implicit learning and statistical learning: One phenomenon, two approaches. *Trends in Cognitive Sciences*, 10, 233–238
- Peterson M.A. (2011). Variable exemplars may operate by facilitating latent perceptual organization. *Infancy*, 16(1), 52–60.
- Pothos, E. M. (2010). An entropy model for Artificial Grammar Learning. *Frontiers in Cognitive Science*, 1, 1–13.
- Reber, A. S. & Allen R. (1978). Analogic and abstraction strategies in synthetic grammar learning, a functional interpretation. *Cognition* 6, 189–221.
- 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.
- Rogers, T., Rakinson, D., & McClelland, J. (2004). U-shaped curves in development: A PDP approach. *Journal of Cognition and Development*, 5, 137-145.
- Saffran, J.R., Aslin, R.N., & Newport, E.L. (1996). Statistical Learning by 8-Month-Old Infants. *Science* 274 (5294), 1926-1928.
- Shannon, C. E. (1948). *A mathematical theory of communication*. Bell System Technical Journal, 27, 379-423.
- Stephen, D. G., Dixon, J. A., & Isenhowe, 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. *Biol. Bull.* 215, 216-242.
- 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
- Vokey, J. R. & Higham, Ph. A. (2005). Abstract analogies and positive transfer in artificial grammar learning. *Canadian Journal of Experimental Psychology*, 59, (1), 54-61.