Patterns bit by bit. An Entropy Model for Linguistic Generalizations

Abstract

When confronted with the challenge of learning their native language, children manage impressively fast to infer generalized rules from a limited set of examples, and apply those rules to strings of words never heard before.

This paper addresses the puzzle of what triggers and what limits the inductive leap from memorizing specific linguistic items to extracting general rules. An innovative entropy model for linguistic generalization is proposed, which is designed to bridge the gap between previous findings on the factors that modulate the process of making linguistic generalizations and to unify them under one consistent account based on an information-theoretic approach to rule induction. The prediction made by this model is that generalization is a cognitive mechanism that results from the interaction of input complexity (entropy) and the processing limitations of the human brain, expressed as limited channel capacity.

In a pilot experiment with adults, a miniature artificial grammar was designed to probe the effect of input complexity on the process of generalization. The number and frequency of linguistic items was manipulated to obtain different degrees of input complexity. Entropy was used as a measure of input complexity, given that entropy varies as a function of the number of items in the input and their probability of occurrence. Results showed that the more complex the linguistic environment, the higher the tendency to create generalized rules in response to the input complexity.

Silvia Rădulescu / 3761991

Utrecht University | Master Thesis |Supervisors: Frank Wijnen, Sergey Avrutin

1. The induction problem for human language

Children are faced with the difficult task of acquiring the rules of the language they are being exposed to. Inferring generalized rules from a limited set of examples, and applying those rules to strings of words never heard before, describes the induction problem for language acquisition (also known as the logical problem for language acquisition.) The fact that children can go about the induction problem and acquire the rules of their language in an amazingly short period of time is a fundamental research problem for linguistics. Despite extensive efforts, we still know very little about how infants so brilliantly manage to do so.

Linguistic generalization entails inferring general rules (principles) from a limited amount of specific examples, and applying those rules to novel strings of linguistic items (sounds, words, etc.). For example, English learners extract a general rule of forming past tense "add *–ed* to the verb stem"; Italian children infer from a limited number of examples that feminine nouns end in *a,* and take the definite article *la*. Children's generalization abilities extend also to forming abstract linguistic categories (such as 'noun', 'determiner', etc.) and to inferring abstract relations between them. For example, learners who hear sequences like *Dad drives slowly* will abstract away from the idiosyncratic properties of the specific linguistic items and generalize to abstract linguistic categories: 'noun' ('dad'), 'verb' ('drive'), 'adverb' ('slowly'), and they also infer a general rule that a Noun-Verb-Adverb sequence is well-formed with other linguistic items belonging to these three linguistic categories, e.g. $Eagles_{[noun]}$ $fly_{[verb]}$ $fast_{[adverb]}$. This is a very powerful phenomenon in language, because it enables a potentially infinite number of strings. Although both artificial grammar learning studies and natural language studies carried out so far on generalization have been informative, there are still a lot of unanswered questions about how infants acquire abstract relations (like those described above, and others, like phonological generalizations, inflectional agreement, binding, etc.) without any explicit instruction.

The linguistic input infants are exposed to is rich in patterns of sounds and groups of sounds. It has been argued that infants can exploit the richness of the sensory world and detect patterns in the auditory/visual input, like phonotactic information (Chambers, Onishi, & Fisher, 2003), and word boundaries (Saffran, Aslin & Newport, 1996), by using a learning mechanism dubbed statistical learning. This mechanism relies on computing transitional probabilities between items (e.g. the probability that a certain sound occurs after another sound). However, statistical learning cannot account for the task of

abstracting rules beyond specific items, because it is confined to patterns of linguistic items that have been encountered in the input, and it is blind to novel items.

Marcus, Vijayan, Bandi Rao, and Vishton (1999) proposed that an algebra-like system is in place in humans, which enables them to extract algebraic rules and apply them to new input, after having been briefly exposed to small amounts of input. These algebraic rules describe relationships between categories (variables) such as "the first item is the same as the third item" (e.g. *li_na_li*). This conclusion was reached after carrying out an artificial grammar study in which after a couple of minutes of exposure to AAB, ABB and ABA strings of syllables (such as *le_le_di, ga_ti_ti, li_na_li*) 7-month-olds generalized the underlying structure to new strings, such as *ko_ko_ba, wo_fe_fe, de_ko_de*. An algebra-like system addresses the case of generalizing to novel input, but it does not explain how humans tune into such algebraic rules, and what the factors (if any) in the linguistic input are that facilitate or impede this process.

Gómez and Gerken (2000) drew a conceptual distinction between the types of abstractions that language learners make: pattern-based abstractions and category-based abstractions. A pattern-based abstraction is a relation between perceptual characteristics of elements, such as a relation based on physical identity, e.g. *ba-ba* (*ba* is followed by *ba*), while a category-based abstraction is an operation over abstract variables (X is followed by X, where X is a variable), for example, a Noun-Verb-Noun generalization 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 different values).

This paper addresses the following question: what triggers and what limits the inductive leap from memorizing specific linguistic items to extracting general rules? More specifically, what triggers the above-mentioned mechanisms of rule extraction: perceptually-bound rules vs. category-based abstractions? Are these mechanisms two different outcomes of the same learning system governed by the same principle? What ignites generalizations that go beyond perceptually-bound abstractions to rules that apply to linguistic categories?

2. An Entropy Model for Linguistic Generalization

This paper proposes a new approach to rule extraction and generalization from an information-theoretic perspective, namely an entropy model. Entropy, as an information-theoretic concept, quantifies the amount of uncertainty, i.e. the average unpredictability in the input (Shannon, 1948). For a random variable X, with n outcomes $\{x_1, x_2 \ldots x_n\}$, Shannon's entropy, denoted by $H(X)$, is defined as:

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

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. Probabilities are numbers between 0 and 1, where 0 shows improbability that a value occurs, while a probability of 1 indicates that something is 100% certain. The entropy of a set of items increases with the number of discrete items, and also if the probabilities of the items are more similar. The entropy of a set is maximal when the probabilities of all the elements in it are the same. Given that entropy 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), entropy is used in this model as a measure of *input complexity.*

Another information-theoretic concept that is employed by this model is *channel capacity,* which describes the amount of entropy that can be processed per unit of time (Shannon, 1948). If the amount of entropy that is fed into a communication channel (e.g. of humans) exceeds its capacity, the amount of entropy has to be reduced in order to get the data through the channel error-free. One way of doing this is to re-organize and regularize the information, in such a way that its entropy is reduced, and it can be processed using the available channel capacity.

The entropy model proposed here puts together these two concepts and makes the prediction that *generalization is a cognitive mechanism that results from the interaction of input complexity (entropy) and the processing limitations of the human brain, i.e. a limited channel capacity*.

How can this model apply to language learning? Learners need to memorize many linguistic items (e.g. English – *a, an, the*), and each item has several perceptually tractable features: the particular sounds and their combinations. However, learners need to acquire also abstract linguistic categories (e.g. the category of articles). Such categories may have more or less items, and each item may have a higher or a lower frequency in language use. For example, suppose a set has only a couple of items – *a, the* – and one occurs in learners' environment more often than the other one (*a* – 75%, *the* – 25%). The complexity of this set will be lower than the complexity of a set of eighteen items with roughly similar probabilities of occurrence (e.g. the Italian determiner system). Learners exposed to the first set will memorize the two items and will expect the high-probability item to appear most often. Learners of the eighteen-item set will have a hard time memorizing every item, and will be uncertain which one will occur in a particular situation. When faced with such situations where learners need to make sense of the richness and

[1] *Log* should be read as *log* to the base 2 here and throughout the paper.

complexity of the linguistic environment, in order to acquire language, learners must adopt some learning strategy to cope with the *input complexity*, and with the *uncertainty* in their representation of the linguistic input. They would try to memorize as many linguistic items as possible tracking different perceptual features, which are automatically grasped (in this case through the auditory system). But memory capacity and processing capacities are limited (they mature in time). Due to overall limited capacity (*channel capacity*) learners fail to learn all the complexity in the sensory world immediately by rote, and they will unconsciously look for ways to reduce the burden put on their memory and processing systems by the input complexity. They will try to find similarities and associations to group the perceptually different items in a linguistic category. It follows that less complexity in the input will allow memorization of specific items, while a higher input complexity that overloads the available channel capacity will drive the tendency to regularize the input (i.e. reduce the number of dimensions/features that individual items can be coded for) and make generalizations based on categories.

Therefore a prediction made by the entropy model is that a low entropy of the set of linguistic items enables learners to extract perceptually-bound rules: this would be the case for rules like "*ends in -ed",* or rules about relations specifying what particular items would follow each other (e.g. *ba* or *ge* follows *wo*). A high entropy of the set of linguistic items enables learners to tune into category-based abstractions, i.e. rules made over abstract categories: such as AAB or AXB patterns, which allow for other new items to be included in the category. Therefore, according to this entropy model, perceptually-bound learning and categorization/rule-induction are outcomes of the same learning mechanism, which is governed by the same principle of creating structure (rules) in response to the degree of complexity (entropy) in the environment.

Another prediction is that an increase of *channel capacity* (i.e., processing and memory capacity) that occurs in the course of human development limits the tendency to regularize the input. The generalization that will be made is the one that reduces entropy (i.e. reduces the number of dimensions/features that individual items can be coded for) to a level that allows getting the data through the available channel error-free. Channel capacity is lower in infants than in adults and it increases at later developmental stages: Marslen-Wilson and Tyler (1981) showed that processing speed increases with age. As a consequence, if infants and adults are exposed to the same amount of entropy in the linguistic input, adults will have a reduced tendency to make a category-based generalization, due to the fact that adults have a higher channel capacity than infants.

What is the relevance of the entropy construct for language acquisition and for the human cognitive system in general? The human brain has been shown to be sensitive to entropy in a series of recent linguistic studies (Baayen, Feldman & Schreuder, 2006; Milin, Kuperman, Kostiç, & Baayen, 2009). For example, in a series of reaction time experiments on processing speed of inflected nouns and verbs in Serbian, Kostiç and colleagues found that participants' performance was dependent on the information load of the individual nouns or verbs and on the entropy of the entire inflection class they belonged to. Moreover, entropy was shown to play an important role in phenomena like decision-making (Tversky & Kahneman, 1992) and problem-solving (Stephen, Dixon & Isenhower, 2009). More specifically, Stephen et al. (2009) showed in a problem-solving task (a gear-system problem) that an increase in the entropy of participants' responses to the task above an overwhelming threshold led to a change in solution strategy. The concept of entropy has been used in quantifying the entropy levels within neural systems (Pereda, Quiroga, & Bhattacharya, 2005), and also in theories on the emergence of consciousness (Tononi, 2008).

3. A previous entropy model and its limitations

Pothos (2010) proposes an information-theoretic model to describe performance in acquiring knowledge about a finite state grammar. More specifically, he employs Shannon's entropy as a measure of quantifying the ease of predicting if a string of symbols is compatible with a trained language, i.e. if a string would possibly be part of the trained language. Pothos (2010) presents an entropy model for artificial grammar learning (AGL) which is based on what he defines as *Information Premise*: the cognitive system prefers choices that allow for as great a reduction in uncertainty as possible. Each string has a certain level of complexity (entropy), depending on the extent to which it can be predicted from the trained language. Suppose a string of symbols is presented to participants without any previous exposure to the language, they would presumably deem any arrangement of symbols into strings to be possible strings in the language (this describes a high uncertainty environment). Conversely, if participants are trained on a language and they take into account the strings of the trained language, then some arrangements of symbols into strings would be more likely than others to be considered grammatical (i.e. part of the trained language), hence a lower level of uncertainty when predicting if a string is part of the trained language. As per his suggested *Information Premise,* the lower the entropy, the easier it will be to recognize the string as being part of the trained language, and consequently performance will be higher.

Pothos suggests the following way to compute the entropy of the stimuli: each string of symbols (item) is divided into parts, namely all the chunks formed by two sequential symbols (bigrams) and the chunks formed by three sequential symbols (trigrams), including the anchor positions (the beginning and ending of the strings - which will be denoted by letters "b" and "e", respectively). For example, a string like *MSV* would be split into four bigrams (*bM, MS, SV, Ve*) and three trigrams (*bMS, MSV, SVe*).

Suppose a participant is exposed to the following strings in training: *MMM*, *MMM*, *SSS*, and asked to evaluate the following string (test item): *MSV*[2]. By applying Shannon's entropy formula, the author computes how familiar the first bigram *bM* is, as follows:

H(bM) = -2/3*log2/3 - 1/3*log1/3 = 0.92[3],

because the continuation after the anchor *b* can be either an *M* with a probability of ⅔ or an *S* with a probability of ⅓. If the number of *bM* had been equal to the number of *bS*, the uncertainty would be higher *H(bX) =* 1. But if there had been 99 *bM* and one *bS*, the uncertainty would be greatly reduced: *H(bM) =* 0.08. The same calculations are done for trigrams in a similar way to the calculations for the bigrams.

The next steps would be to compute *H(MS), H(SV) and H(Ve)* in the same way. But how could entropy be calculated, if a bigram has not been seen in training? The author suggests that the corresponding entropy in such a case should be computed by assuming all possible continuations. Namely, he proposes *H(MS)* = (-1/4log1/4)*4 = 2, meaning that any symbol can have four possible and equiprobable continuations - *M, S, V, e*. However, as the author states himself, there is no evidence in artificial grammar learning that such probabilities are relevant in performance in this way. The author gives no argument for his assumption that the entropy of untrained bigrams should be computed by considering that all symbols present in the test item would be equally probable.

Moreover, as shown in previous sections, entropy is by definition the average amount of information contained by a set of values that can be taken by a variable. In this particular case considered by Pothos (*MMM, MMM, SSS* in training, and *MSV* as test item), what he calculated and denoted as *H(bM)* is actually the entropy of the set of values that can be taken by the first bigram in all the strings {*bM, bM, bS*} from the training phase (not the entropy of the specific bigram *bM*). However, he refers to this entropy as the entropy of the specific bigram *bM*, which is a contradiction in terms, as mentioned before, because

[2] The author presumably denotes particular symbols generated by a grammar with *M, S, V.*
[3] The author uses *S(X)* to denote Shannon's entropy, but for consistency with the entropy formula presented above I will use *H(X)* throughout the paper to denote Shannon's entropy, as denoted in Shannon (1948).

bM is just a value taken from the set of first bigrams of all the training items. When he calculates *H(MS)* he does not make the calculations based on the training items, as he did for *H(bM)*, which is an inconsistency in his method. But what he actually calculates and denotes as *H(MS)* is the entropy of the set of symbols that make up the individual test string *MSVe* (in which all four distinct symbols appear with equal probability – *p=1/4*), without considering any training items. This is an irrelevant calculation for his argument, given that he proposes entropy to quantify predictability of test items from training items. Therefore I suggest that his assumptions and his model should be revised and a more precise and better-suited model should be devised to account for untrained items.

According to the entropy model proposed in this paper, this case would be analyzed as follows: given the following training items – *MMM*, *MMM*, *SSS* - the set of the first bigrams in the strings contains the following bigrams {*bM, bM, bS*}, the set of the second bigrams contains *{MM, MM, SS}*, and so on. The entropy of the set of the first bigrams is calculated as follows:

H{bM, bM, bS} = $-\sum_{i=1}^{n} p(x)logp(x)$ = -[p(bM)*log p(bM) + p(bS)*log p(bS)] = -[2/3*log2/3 + 1/3*log1/3] = 0.92. Similarly, the entropy of the second set of bigrams is calculated: H{MM, MM, SS} = $-\sum_{i=1}^{n} p(x)logp(x)$ = -[p(MM)*log p(MM) + p(SS)*log p(SS)] = -[2/3*log2/3 + 1/3*log1/3] = 0.92. These are low entropy values which allow for easy memorization of the two possible values of the first bigram, namely *bM* and *bS*, which predicts a perceptually-bound rule to be extracted of the form "first bigram is *bM* or *bS*", with *bM* having higher probability than *bS*. Similarly, a perceptually-bound rule is extracted for the second bigram "second bigram is *MM* or *SS*", with *MM* having higher probability than *SS*. When the test item *MSV* is given, the first bigram *bM* will be matched with certainty (or very low uncertainty – H{bM, bM, bS} = 0.92) as conforming to the rule, but the second bigram *MS* (which has zero probability of occurrence, considering the training items) will be rejected with the same certainty (or very low uncertainty – H{MM, MM, SS} = 0.92), because it does not match the rule "second bigram is *MM* or *SS*".

Pothos (2010) argues that an important aspect of an entropy model for AGL as opposed to a conditional probability approach (or transitional probability as described in the previous sections) would be the fact that an entropy model intrinsically takes into consideration not just the absolute frequency of the symbol or bigram, but most importantly the number and frequency of the competing ones. To put it simply, consider the following two training sets of symbols: {*XY, XY, XZ*} and {*XY, XY, XY, XY, XQ, XZ*}. A conditional probability approach would predict the same probability for *Y* to follow to follow *X,* in both sets, as the conditional probability of Y given X is P(Y|X) = ⅔ in both sets. On the other hand, an entropy model would predict different levels of predictability for *XY*, because the bigram entropy in the first set is

0.92, while in the second one 1.25. To further clarify what Pothos means, if trained on the larger set of symbols, *XY* would be endorsed with higher uncertainty, although it has the same probability, because there is also the question of how many alternative choices with their particular odds are taken into consideration in parallel when predicting a particular symbol with the highest probability.

Pothos (2010) analyzed two particular datasets from two studies: Pothos, E. M., Chater, N., & Ziori, E. (2006) and Pothos and Bailey (2000). Pothos et al. (2006) employed the grammar that Reber and Allen (1978) investigated, while Pothos and Bailey (2000) used the grammar from Knowlton and Squire (1996). The author examined correlations between four entropy measures (summed bigram entropy, average bigram entropy, summed trigram entropy, average trigram entropy) and the other predictors of AGL performance as considered by the authors of those studies. The results were partially as expected, meaning that higher entropy was correlated with lower endorsement rates. The average trigram entropy was the best predictor for the rate of grammaticality endorsements, followed by average bigram entropy, then summed trigram entropy and the last was the summed bigram entropy. However in the transfer conditions (training and test sequences are composed of different sets of symbols, but have the same underlying grammar) the correlations were not significant, although they were in the same direction (i.e. higher entropy was correlated with lower endorsement rates). The author argues that this finding shows the fact that entropy measures do not work well when they are applied to new symbols.

Wrapped up, the entropy-based proposal made by Pothos (2010) about learning a finite state grammar goes like this: the main claim is that if for a test item it is easy to guess all continuations from individual symbols and bigrams, then the overall entropy of the item will be low and thus the item will be more likely to be endorsed as grammatical. Easy to guess continuations are those where there are few options to choose from, in other words high probability bigrams and trigrams. It follows from this that strings with highly probable bigrams/trigrams and with a low number of bigrams/trigrams are going to be endorsed as grammatical more often than those with a lower probability.

Pothos's model only tackles the perceptually-bound abstractions, in that finite-state grammars contain only a finite number of items and the regularity in this type of grammars is defined by relations that specify what particular tokens would follow each other. Pothos's conclusions are in line with one of the predictions made by the entropy model proposed in this paper: a low entropy of the set of items (given by high probability bi-/trigrams and a low number of items) enables perceptually-bound rules to be extracted (rules about which particular items follow each other).

4. A Unified Account for Previous Studies. A Proof of Concept.

4.1. A Reinterpretation of Previous Findings

The model proposed in this paper is designed to bridge the gap between previous findings of relevant artificial grammar studies and to unify them under one consistent account by giving a reinterpretation in terms of brain's sensitivity to the entropy in the linguistic input. How does this entropy model provide a unified account for previous findings and how can such a model be applied to language learning studies?

In order for an entropy model to be applied, evidence is needed that knowledge is acquired about categories of items that can be construed as variables: there has been extensive evidence that grammaticality judgments in artificial grammar learning are shaped by knowledge acquired about chunks of trained items - bigrams and trigrams (Perruchet and Pacteau, 1990; Knowlton and Squire, 1994). Studies also showed that performance is predicted by the frequency of these chunks in the training phase (Knowlton and Squire, 1994). There is also evidence for a form of abstract learning, which allows for transfer of the knowledge to novel items, based on abstract analogy to the specific training items (Brooks and Vokey, 1991; Vokey and Higham, 2005). These are the prerequisites for applying an entropy model of artificial grammar learning.

In miniature artificial grammar studies using patterns of the form AAB, $A_i\_X_i\_B_i$, each position of the patterns can be considered to form a variable (a category of items), whose possible values are the specific items: for example, variable A in a study focusing on learning an AAB pattern (*le_le_di*) could be filled by *le, wi, ji, de,* etc. Each category of bigrams and trigrams can be also considered to form 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 grammar studies, the strings generated by the specific grammar can be construed as groups of bigrams and trigrams, and entropy can be calculated in a similar way. Therefore calculations can be made to quantify the amount of entropy contained in the training input in this kind of studies.

Previous studies focusing on either one or the other type of generalization found different kinds of mechanisms to account for the specific abstraction they were investigating. The entropy model proposed in this paper gives a conceptual analysis that accounts for both types of linguistic generalizations that have been observed and conceptualized in previous studies – perceptually-bound abstractions and

category-based abstractions – by identifying the same factors whose interplay is predicted to be the source of both types of generalizations: *input complexity* and *channel capacity*. This entropy model gives a quantitative measure for the likelihood of extracting a category-based generalization or a perceptually-bound rule, and thus it allows for precise predictions regarding the generalization process for any range of *input complexity*.

Gerken (2006) modified the design used by Marcus et al. (1999) and reconsidered their argument. In the training phase Marcus et al. (1999) presented 7-month-old infants with 16 AAB strings, of which four strings ended in *je* (*leleje, wiwije, jijije, dedeje*), another four strings ended in *li (leleli, wiwili, jijili, dedeli)*, other four strings ended in *di (leledi, wiwidi, jijidi, dededi)*, and other four strings ended in *we (lelewe, wiwiwe, jijiwe, dedewe)*. Gerken (2006) asked whether 9-month-old infants presented with two different subsets of these 16 strings would make the same generalization. More specifically when both rules are possible, the AAB rule and "ends in *di*" rule, what rule would be favored by the learners? In order to answer this question, Gerken (2006) presented one group of infants with four AAB strings ending in different syllables (Table 1 – named the "diagonal condition" by the author) and another group with four AAB strings ending only in *di* (Table 2 – named the "column condition" by the author). Infants in the second group had two equally plausible rules at hand: the more general AAB rule and the more restrictive "ends in *di*" rule.

Diagonal condition
[A A B] le le di wi wi je ji ji li de de we
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[bigram] = 2 H[trigram] = 2

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

The results showed that the second group failed to generalize to a novel string that displayed the AAB pattern (e.g. *ko ko ba*), but did not end in *di,* while the first group made this generalization. The explanation was that the only reliable rule that applied to the set of AAB strings ending in the same syllable *di*, was the more restrictive "ends in *di*" rule, because each and every string ended in *di*, but there was no evidence that any string could end in any other syllable, to support a generalization to novel strings. But to what extent is a certain rule reliable or likely to apply? What is the threshold for a rule to become reliable or likely enough to apply? And even more deeply, why do learners infer a general rule at all?

A reinterpretation according to the entropy model can be given to Gerken's findings, in order to help answer the unanswered questions as well. Tables 1 and 2 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 values, as well as the average bigram entropy (H[bigram]) and the average trigram entropy (H[trigram])[4].

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

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

The experiment condition that had an input characterized by a higher entropy (Table 1) yielded generalization to the more general AAB pattern, while the one with low entropy (Table 2) resulted in a more perceptually-bound rule "ends in *di*". As their follow-up experiment showed, the participants assigned to the "column condition" (Table 2) did generalize the AA category, and kept a specific

[4] 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.

perceptually-bound rule only for the last position of the strings, and therefore extracted an "*ends in di"* rule.

As can be seen in Tables 1 and 2, the two sets of stimuli have the same average trigram entropy, but they differ in terms of average bigram entropy. More specifically the difference in average bigram entropy comes from the difference in the amount of entropy of the final bigram H[Be]. These differences could explain the findings that learners made a generalization to variables A and B only when the amount of entropy was maximum, and they extracted a perceptually-bound rule when the amount of entropy was minimum (the entropy of the final bigram H[Be] in the "column condition" is 0).

Reeder, Aslin & Newport (2009) carried out a study to find the factors that trigger and modulate generalization in a series of experiments investigating the role of distributional cues on linguistic category formation. The authors examined adults' likelihood to generalize a category when exposed to input that had different distributional properties, such as richness of contexts, overlap in contexts, systematic gaps, and also their tendency to generalize when exposure time was increased.

More specifically, the participants were trained on an artificial grammar with strings of non-sense words having the underlying structure: (Q)_A_X_B_(R)[5]. Categories A, X, and B had three words each, forming 3x3x3=27 strings altogether. In a series of experiments designed to probe whether the participants can generalize X as a category of words, rather than just memorize the exact strings, the experimenters presented the participants with different subsets of strings from this grammar and in the test phase they were presented with the withheld grammatical strings, plus ungrammatical strings (A_X_A or B_X_B strings) and they were asked to evaluate on a scale from 1 to 5 if the strings sounded like they came from the grammar or not. Table 3 shows the exact stimuli presented in all experiments, together with the average evaluation rates given by the participants to the test strings.

In Experiment 1 the training subset densely sampled the language, in that all words within each category appeared in highly overlapping contexts - all *As* concatenated with all *Xs* and with all *Bs* (Figure 1.A.). Results showed no significant differences between ratings for grammatical familiar items and grammatical novel strings, but they were both significantly higher than the ratings for ungrammatical strings.

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

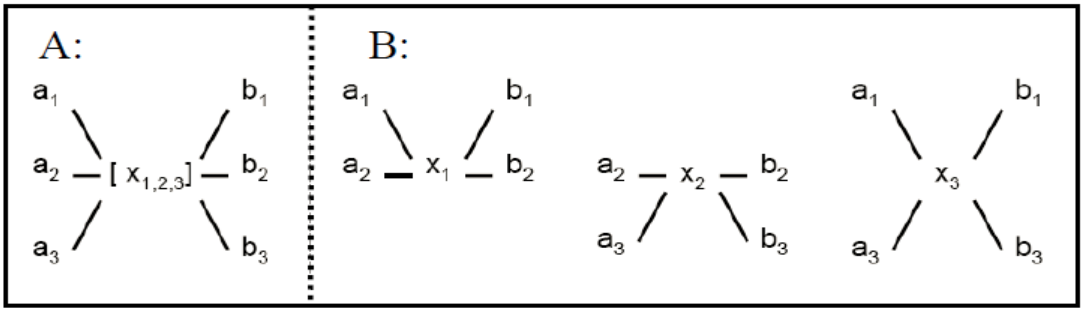


Figure 1. Side-by-side comparison of full overlap (A) and partial overlap (B)

(taken from Reeder et al., 2009)

Authors concluded that learners fully generalize the target category to novel strings, when there is richness of contexts for the target category *X*. In Experiment 2, training displayed a reduced number of exemplars (sparseness), and thus reduced exposure, but still all *As* concatenated with all *Xs* and with all *Bs,* there were insignificant differences in evaluation rates as compared to the performance for Experiment 1, and the authors reported that learners' ability to fully generalize the target category was unchanged as compared to the first experiment, when there was sparseness of exemplars.

	Experiment_1	Experiment_2	Experiment_3	Experiment_4
Stimuli	$A_1X_1B_1$ $A_1X_2B_2$ $A_1X_3B_1$ $A_1X_1B_3$ $A_1X_2B_3$ $A_1X_3B_2$ $A_2X_1B_2$ $A_2X_2B_1$ $A_2X_3B_1$ $A_2X_1B_3$ $A_2X_2B_2$ $A_2X_3B_3$ $A_3X_1B_1$ $A_3X_2B_1$ $A_3X_3B_2$ $A_3X_1B_2$ $A_3X_2B_3$ $A_3X_3B_3$	$A_1X_1B_3$ $A_1X_2B_2$ $A_1X_3B_1$ $A_2X_1B_2$ $A_2X_2B_1$ $A_2X_3B_3$ $A_3X_1B_1$ $A_3X_2B_3$ $A_3X_3B_2$	$A_1X_1B_3$ $A_2X_2B_1$ $A_1X_3B_1$ $A_2X_1B_2$ $A_3X_2B_1$ $A_1X_3B_2$ $A_2X_1B_3$ $A_3X_2B_3$ $A_3X_3B_2$	$A_1X_1B_3$ $A_2X_2B_1$ $A_1X_3B_1$ $A_2X_1B_2$ $A_3X_2B_1$ $A_1X_3B_2$ $A_2X_1B_3$ $A_3X_2B_3$ $A_3X_3B_2$
Rated (1-5)	Gram_Familiar: 3.78 Gram_New: 3.69 Ungrammatical: 2.58	Gram_Familiar: 3.54 Gram_New: 3.47 Ungrammatical: 2.73	Gram_Familiar: 3.79 Gram_New: 3.48 Ungrammatical: 2.85	Gram_Familiar: 4.05 Gram_New: 3.64 Ungram: 2.83
Entropy values	H[AX] = 3.169 H[bA] = H[Be] = 1.584 H[XB] = 3.169 H[AXB] = 4.169	H[AX] = 3.169 H[bA] = H[Be] = 1.584 H[XB] = 3.169 H[AXB] = 3.169	H[AX] = 2.503 H[bA] = H[Be] = 1.584 H[XB] = 2.503 H[AXB] = 2.584	H[AX] = 2.503 H[bA] = H[Be] = 1.584 H[XB] = 2.503 H[AXB] = 2.584

	H[bigram] = 2.376 H[trigram] = 3.502	H[bigram] = 2.376 H[trigram] = 3.169	H[bigram] = 2.043 H[trigram] = 2.530	H[bigram] = 2.043 H[trigram] = 2.530

Table 3. Stimuli, evaluation rates and entropy values of the input data in the experiments by Reeder, Aslin & Newport (2009)

In the third experiment, the subset of training strings was characterized by a drastically reduced overlap of contexts for the X words, as presented in Figure 1.B. As can be seen in Table 3, results showed that grammatical familiar strings were rated significantly higher than grammatical novel strings. The authors concluded that an input characterized by partial overlap, despite full coverage of the *X* words, led to a slight decrease in generalization, even though participants still continued to generalize, as shown by the significant difference between their ratings of grammatical novel strings and ungrammatical strings. In the fourth experiment, they increased the exposure to the same input used in Experiment 3. As presented in Table 3, results showed highly significant differences between all ratings, which was reported to show that, when exposure time is increased, participants' likelihood to generalize significantly decreases, because they have more certainty that the gaps in distributional contexts are meaningful. The authors hypothesize that an even longer exposure to input displaying partial overlap would lead to an eventual complete lack of generalization.

In conclusion, Reeder et al. (2009) found richness of contexts to drive generalization of the target category of words, while reduced number of exemplars did not affect generalization. Conversely incomplete overlap of contexts and increased exposure reduced the likelihood of generalization. While this study sheds light on some factors that trigger or impede generalization, there are still unanswered questions: why do these factors affect generalization in linguistic category formation and why do they modulate generalization in a different manner? Are they completely independent factors? Moreover, these factors presented as such do not allow for precise predictions to be formulated regarding the likelihood to make generalizations, given that the authors provide no measure to quantify different degrees of these factors in the input.

An entropy-based reinterpretation can be given to the findings of the experiments conducted by Reeder, Aslin & Newport (2009), which would help answer the above questions. A question raised about these findings is whether all the factors identified by the authors are independent factors that modulate generalization, and I suggest that it is the amount of entropy contained by each set of stimuli that accounts for the results of all these experiments. Table 3 includes also entropy calculations as per the entropy

formula presented in this paper. As can be seen in Table 3, the two data sets used in the first two experiments present similarity in terms of entropy values, which might explain the absence of significant difference in learners' performance, despite the fact that in Experiment 2 exposure is half as long and only half the number of exemplars were presented. The entropy values for the set of stimuli used in Experiment 3 were significantly reduced as compared to the first two experiments, which might explain learners' lower likelihood to generalize the categories. An increased exposure to the same stimulus set in the fourth experiment, might have led to an increased memory trace of the perceptual characteristics of the stimuli, therefore a reduced need of making a category-based abstraction. In conclusion, one unifying account for all the findings of these experiments can be found in the amounts of entropy displayed by the input: an increase in entropy drives the tendency to generalize to categories of items, while a decrease in entropy results in a decrease of the likelihood of making a category-based generalization, and a higher tendency to extract a perceptually-bound rule.

Reeder, Aslin & Newport (2009) note that at some point along the sparseness and overlap continuum, there must be a threshold for shifting from word-by-word learning to category generalization. However using their account no precise prediction can be made as to where this threshold would lie, given that they do not present any measure to quantify sparseness and degree of overlap, and therefore no prediction can be made from their hypotheses as to how this threshold can be found.

The reinterpretations of these previous findings call for an explanation of the apparently opposing predictions derived from an entropy model when applied to finite state grammars (Pothos, 2010) and the latter miniature artificial grammars investigating category-based generalizations (Gerken, 2006; Reeder, Aslin & Newport, 2009). Actually they are not contradictory predictions, but they both follow from the entropy model proposed in this paper: in order to extract perceptually-bound rules (rules based on perceptual characteristics of the specific items) a low-entropy environment (low *input complexity*) is needed, because those specific tokens that enter into relations have to be kept in working memory. Conversely, a high-entropy set of items (high *input complexity* that overloads *channel capacity*) increases the likelihood of making category-based abstractions. These predictions were clearly borne out in the studies using miniature artificial grammars to investigate category-based generalizations (Gerken, 2006; Reeder, Aslin & Newport, 2009), as shown in this section. As for the case of finite state grammars, the grammaticality endorsements are based on perceptually-based rules, i.e. on rules specifying what particular items would follow each other.

4.2. A Pilot Experiment. Purpose and hypotheses.

The previous sub-section presented a post-hoc analysis of previous studies in line with the proposed entropy model. The following sections present a pilot AGL experiment that specifically tested some predictions made by the entropy model presented in this paper. To the best of my knowledge, this is the first AGL experiment that investigates the role of the interplay between *input complexity* and *channel capacity* in the process of making linguistic generalizations by specifically testing entropy-based predictions.

Given that the primary prediction of the model is that generalization is a cognitive mechanism that results from the interaction between *input complexity* and *channel capacity*, these are the two factors that would need to be varied when testing this model. Thus, a first line of experiments would have to keep *channel capacity* constant and vary *input complexity,* from a low to medium to high entropy. One way to keep *channel capacity* constant is to test subjects in the same maturational stage, i.e. having roughly the same age or the same working memory and processing capacity. Channel capacity is assumed to be lower in infants than in adults, and it has been shown to increase with age (Marslen-Wilson & Tyler, 1981). A second line of experiments would have to keep *input complexity* constant and vary *channel capacity.* One way of easily doing that is to expose two groups in different maturational stages, i.e. infants and adults, to the same degree of *input complexity* and compare the results.

This pilot experiment was designed along the first line of research, namely *channel capacity* was kept constant, by testing adults of roughly the same age, and *input complexity* was varied to obtain three different degrees of entropy. An artificial language was created in order to precisely manipulate *input complexity.* The following hypotheses were probed:

i. the lower the *input complexity* (entropy), the higher the tendency towards perceptually-bound learning, and, consequently, the lower the tendency to make a category-based generalization;
ii. the higher the *input complexity* (entropy), the higher the tendency to make a category-based generalization;
iii. perceptually-bound learning and categorization/rule-induction are outcomes of the same learning mechanism, which is governed by the same principle of creating structure (rules) in response to the degree of complexity (entropy) in the environment.

According to the hypotheses, two tendencies are predicted to be developed by the participants throughout the experiment: perceptually-bound generalization and category-based generalization. Both

tendencies are hypothesized to develop gradually, depending on the amount of *input complexity* that can be processed given the available *channel capacity*. The higher the *input complexity* that needs to be dealt with having a given *channel capacity*, the higher the tendency to abstract away from a perceptually-bound rule to tune into a category-based abstraction.

5. Method

5.1. Participants

Thirty-five Dutch speaking adults (26 females and 9 males, age range 19-26, mean 22) participated in the experiment. An additional participant was tested, but excluded from the analysis because he reported being familiar with research on similar grammar learning in AGL setups. Only healthy non-dyslexic subjects that had no known hearing impairment or attention deficit were accepted in the experiment. They were paid 5 EUR for participation.

5.2. Training stimuli

Participants in the experiment were exposed to 3-syllable strings that implement a miniature artificial grammar, which closely resembles the structural pattern used by Gerken (2006), i.e. the strings followed an underlying XXY[6] 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, sjie*). A subset of the syllables was used in the two *X* slots of the pattern – to be called X-syllables – and another subset of syllables was 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.

All syllables and strings of syllables were generated using a Perl script, in order to precisely control for certain characteristics of the syllable structure and string structure, in order to ensure that no unintended patterns of syllable or string structure would be accidentally confounded in the miniature grammar. The script also checked if the strings were found in the CELEX database[7], to exclude actual

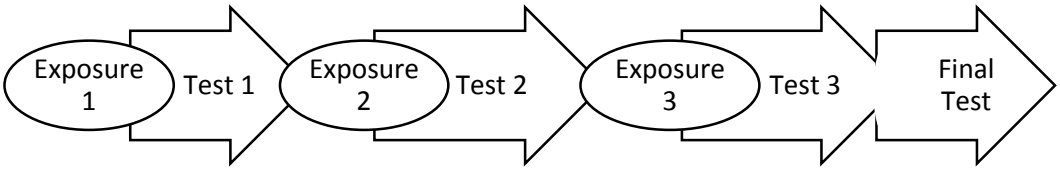
[6] An XXY pattern describes strings consisting of two identical syllables (X) followed by another different syllable (Y): e.g. *goo_goo_sjie, puu_puu_saa, joe_joe_feu*

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

existing words, if any, to ensure that they are non-sense syllables/strings in Dutch, in order to avoid any semantic cue that would strengthen the memory trace of any of the strings.

All the syllables were recorded in isolation by a female Dutch native speaker in a sound-proof booth, using a TASCAM DA-40 DAT-recorder. Syllables were recorded one by one, as they were presented to her on a screen, and she was instructed to use the same intonation for each syllable. The recorded syllables were spliced together to form the strings of the language using Praat[8].

The experiment had three conditions: High Entropy condition (HiEN), Medium Entropy condition (MedEN) and Low Entropy condition (LowEN). Each condition consisted of three exposure phases with intermediate test phases, followed by a final test phase, as per the timeline below:



All conditions had equal number of training strings - 72 XXY strings in total. All exposure phases had equal number of training strings in all conditions -24 XXY strings - which were presented in a randomized order per participant (complete stimulus set in Appendix 1). Intermediate tests were included to have a closer insight on the curve of learning, in order to gauge the learning process as a function of exposure (expressed in number of training strings). The intermediate tests and the test strings were presented in a randomized per participant fashion, in order to prevent any idiosyncratic properties of any of the test strings to be confounded in the results regarding the effect of exposure on learning. This was a between-subjects design, and participants were assigned randomly to one of the three conditions.

5.3. Entropy values of training conditions

In order to obtain the desired variation in *input complexity* 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 formula[9], three different values for *input complexity* were obtained - high, medium and low entropy. For the LowEN condition 6 X-syllables and 6 Y-syllables were used. For the MedEN, 6 extra X-syllables and 6 extra Y-syllables were added to those used

[8] www.praat.org

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

for the LowEN condition, in order to have 12 X-syllables and 12 Y-syllables. For the HiEN condition, 12 more X-syllables and 12 more Y-syllables were added to those used for MedEN, in order to have 24 X-syllables and 24 Y-syllables (see figures below).

To generate the XXY strings for the LowEN condition, all 6 XX pairs were concatenated with all 6 Y-syllables, but only a subset of 24 combinations (24 strings) was used for the first training phase, another subset of 24 combinations was used for the second training phase, and another subset for the third one. This ensured an equal number of XX_Y combinations in each training phase. The same procedure was applied to the other conditions.

Syllable frequency was controlled for in each training phase. Given that each training phase had 24 strings, in the LowEN condition, which had 6 x 6 syllables, each syllable was used 4 times (*p(syllable) = 0.167*) in each training phase. In the MedEN condition, which had 12 x 12 syllables, each syllable was used 2 times (*p(syllable) = 0.083*) in each training phase. In the HiEN condition, which had 24 x 24 syllables, each syllable was used only once (*p(syllable) = 0.042*) in each training phase (complete stimulus set in Appendix 1). The entropy values were kept constant in all the training phases of each condition. The complete entropy calculations per training phase in each condition are presented below:

<u>High entropy condition</u>

H[bX]=H[24]= -Σ[0.042*log0.042] = 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[bigram] = 4.58

H[trigram] = 4.58

High Entropy							
72 strings							
24 x 24 syllables							
Syllable X				Syllable Y			
goo	goe	haa	luu	sjie	sjoe	woe	meu
puu	heu	hie	noo	duu	weu	sjeu	chie
teu	juu	jie	noe	saa	fie	buu	zuu
vee	nie	joe	ruu	feu	kaa	daa	fuu
woo	roo	jeu	vie	moo	muu	faa	boo
loo	vuu	lie	veu	kee	choo	koo	huu

Medium entropy condition

H[bX]=H[12]= -Σ0.083*log0.083 = 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[bigram] = 3.83

H[trigram] = 4.246

Medium Entropy			
72 strings			
12 x 12 syllables			
Syllable X		Syllable Y	
2 * goo	2 * goe	2 * sjie	2 * sjoe
2 * puu	2 * heu	2 * duu	2 * weu
2 * teu	2 * juu	2 * saa	2 * fie
2 * vee	2 * nie	2 * feu	2 * kaa
2 * woo	2 * roo	2 * moo	2 * muu
2 * loo	2 * vuu	2 * kee	2 * choo

Low entropy condition

H[bX]=H[6]= -Σ0.167*log0.167 = 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[bigram] = 3.08

H[trigram] = 3.91

Low Entropy	
72 strings	
6 x 6 syllables	
Syllable X	Syllable Y
4 * goo	4 * sjie
4 * puu	4 * duu
4 * teu	4 * saa
4 * vee	4 * feu
4 * woo	4 * moo
4 * loo	4 * kee

As shown in these entropy calculations, the average bigram entropy in the HiEN condition (H[bigram] = 4.58) is higher than the average bigram entropy in the MedEN condition (H[bigram] = 3.83), which is higher than the average bigram entropy in the LowEN condition (H[bigram] = 3.08). Similarly, the average trigram entropy in the HiEN condition (H[trigram] = 4.58) is higher than the average trigram entropy in the MedEN condition (H[trigram] = 4.246), which is higher than the average trigram entropy in the LowEN condition (H[trigram] = 3.91).

5.4. Procedure

Participants were tested in a sound-proof booth and received instructions 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 explained that the experiment had three phases, and during each phase a number of 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 hear in the tests could be possible in the language that they heard. The experiment lasted around 5 minutes.

After the experiment, participants were briefly interviewed to get their feedback on any strategy they might have used in responding to the test items, their confidence about their responses, and their awareness of the moment when they started using their strategy, if any.

5.5. Test string types and performance predictions

The test strings were all 3-syllable strings designed as four different types: grammatical familiar (XXY strings with trained syllables), ungrammatical novel (XYZ[10] strings with untrained syllables), grammatical novel (XXY structure with untrained syllables), and ungrammatical with familiar syllables (XYZ strings with trained syllables). 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). Therefore, there were 20 test strings in total, and they were all included in all three conditions (complete test item set in Appendix 1). The test strings were also generated using the same Perl script that was used for the training stimuli. A subset of the syllables used in training were concatenated to create XYZ test strings with familiar syllables. Any of the X-syllables and Y-syllables were randomly assigned to the X, Y or Z slot of the XYZ pattern. X-syllables and Y-syllables that were not used in any of the exposure phases were spliced together to create XXY test strings with untrained syllables, and also to generate XYZ strings with unfamiliar syllables.

Each test string type was designed to test each mechanism of rule extraction in a specific way. According to the hypotheses, two tendencies are predicted to be developed by the participants gradually, depending on the amount of *input complexity* that can be processed given the available *channel capacity*. It is hypothesized that the answers that would be given to each type of test strings depend on the state

[10] XYZ pattern describes strings consisting of three different syllables (e.g. *doo_vaa_seu, teu_duu_saa*).

of development reached by each tendency, depending on whether the *channel capacity* is overloaded or not by the *entropy* in the input. For ease of presentation, the tendency towards perceptually-bound learning will be denoted as PERCEPT, and the tendency towards category-based generalization/rule-induction will be denoted as INDUCT. The four test string types and the predicted performance per test string type are presented below.

Type_1: trained_syll_XXY (XXY structure with trained X-syllables and trained Y-syllables) **– correct answer: [1]**[11] – this is a positive test case[12], which is intended to check learning of the trained strings. There is a very high probability that the correct answer would be given to Type_1 test strings in all three conditions, either due to INDUCT being strongly or at least satisfactorily developed in HiEN and MedEN, respectively, or due to PERCEPT being strongly developed in LowEN. As a result, all groups are expected to perform significantly above chance level, with no between-group difference.

Type_2: untrained_syll_XYZ (XYZ structure with untrained syllables) **– correct answer: [0]** – this is the complementary negative test case[13], which is intended to check learning of the trained strings and string pattern. It is designed to back up and complement results obtained for Type_1: if the tendencies hypothesized to yield the results described in the prediction for Type_1 indeed work as intended, then in the negative test case they should work consistently. There is a very high probability that the correct answer would be given to Type_2 test strings in all three conditions, either due to INDUCT being strongly or at least satisfactorily developed in HiEN and MedEN, respectively, or due to PERCEPT being strongly developed in LowEN. As a result, all groups are expected to perform significantly above chance level, with no between-group difference.

Type_3: untrained_syll_XXY (XXY structure with untrained syllables) **– correct answer: [1]** – this is a positive test case, which is intended to be the TARGET test string type to check generalization of rule to novel strings (categorization/rule-induction). All groups are expected to perform significantly above chance level, and also between-group differences in performance are expected: the highest number of correct answers are expected in HiEN condition, due to a strongly developed INDUCT, followed by MedEN

[11] [1] – means the Yes button is to be pressed, meaning the participant considered the test item as a possible string in the language; [0] – means the No button is to be pressed, coding for the participant's rejection of the test item as a possible string in the 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.

[13] 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.

due to a satisfactorily developed INDUCT, with LowEN having the lowest performance, because PERCEPT is strongly developed, while INDUCT is weak.

Type_4: trained_syll_XYZ (XYZ structure with trained syllables) **– correct answer: [0]** – this is the complementary negative test case, which is designed to back up and complement results obtained for Type_3: if the tendencies hypothesized to work as described in the prediction for Type_3 indeed work as intended, then in the negative test case they should work consistently. As such results at this test item should capture the two mechanisms working against each other, because the memory trace of trained syllables tends to drive acceptance of strings to be part of the language. Hence differences in performance are expected across conditions, depending on the extent to which PERCEPT is developed: the LowEN condition is expected to have the highest performance because PERCEPT is strongly developed to remember the syllables sequences; HiEN is predicted to follow with a similar performance, because INDUCT is efficient enough to help push the correct [0] response. MedEN is expected to display the lowest performance of all three conditions, because the memory trace of the individual trained syllables works against a [0] response, and because PERCEPT is too weak to have created a strong memory trace of the correct sequence of syllables, while INDUCT is not strongly developed to firmly prevent the wrong answer: in this case the two tendencies work against each other with almost similar strength.

Regarding the learning process as a function of exposure (the performance curves across tests), a difference between conditions is expected in the performance level for Type_3 test for rule-induction, in the sense that performance should have a slowly decreasing trend in HiEN, a slightly steeper decreasing trend in MedEN, and an even steeper decreasing trend in LowEN.

6. Results

In order to test the effect of *input complexity* on the process of making generalizations, when *channel capacity* was held constant (by testing adults of roughly the same age), the High Entropy condition, the Medium Entropy condition and the Low Entropy condition were compared in a Generalized Linear Mixed Model, with Entropy Group, Test String Type x Entropy Group interaction, Test x Entropy Group interaction as fixed factors, and Subject as random factor. An alpha level of *.05* was used for all statistical tests. There was a statistically significant Test String Type x Entropy Group interaction ($F(9, 679) = 6.428$, $p = .000$). There was no statistically significant main effect of Entropy Group ($F(2, 679) = 0.408$, $p = .66$), and no statistically significant Test X Entropy Group interaction ($F(9, 679) = 1.161$, $p = .31$). The performance levels

(in percentages of correct answers) of each Entropy Group for each Test String Type are presented below. Figure 2 presents the mean performance across conditions for Type_3: untrained_syll_XXY (i.e. correctly accepted XXY strings with untrained syllables): the mean performance for High Entropy was 80% (Mean = .80, SD = .403, SE = .052), for Medium Entropy was 73% (Mean = .73, SD = .446, SE = .058), and for Low Entropy was 65% (Mean = .65, SD = .480, SE = .065). The analysis indicated a statistically significant above-chance performance for Type_3 in all three conditions (*t(59) = 5.76, p = .000; t(59) = 4.05, p = .000; t(54) = 2.38, p = .02,* respectively).

The difference in performance between Type_3: untrained_syll_XXY and Type_1: trained_syll_XXY is

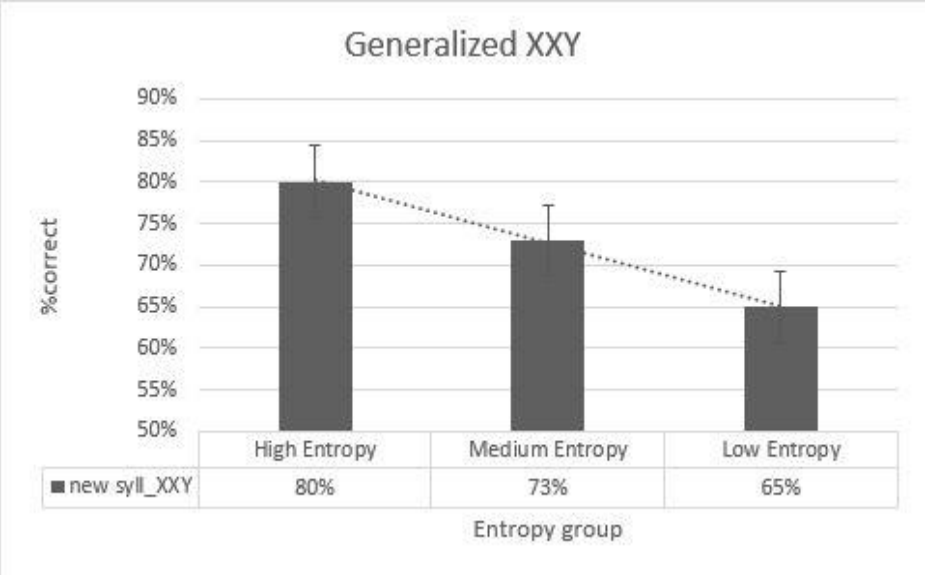


Figure 2. Percentage of correct answers for Type_3: untrained_syll_XXY

shown in Figure 3. An ANOVA revealed that the mean difference between performance in untrained XXY strings vs. trained XXY strings is higher in the Low Entropy condition (Mean = .33, SD = .511, SE = .069; F(1, 108) = 24.220, *p = .000*) than in the Medium Entropy condition (Mean = .23, SD = .427, SE = .055; F(1, 118) = 15.292, *p = .000*), and higher than in the High Entropy condition (Mean = .17, SD = .376, SE = .049; F(1, 118) = 9.012, *p = .003*). The mean difference between untrained XXY vs. trained XXY strings in the High Entropy condition was significantly lower than the corresponding mean difference in Low Entropy condition, F(1, 113) = 3.752*, p = .05.* Results indicated a non-significant trending in the same predicted direction for the mean difference between untrained XXY vs. trained XXY strings in the Low Entropy vs.

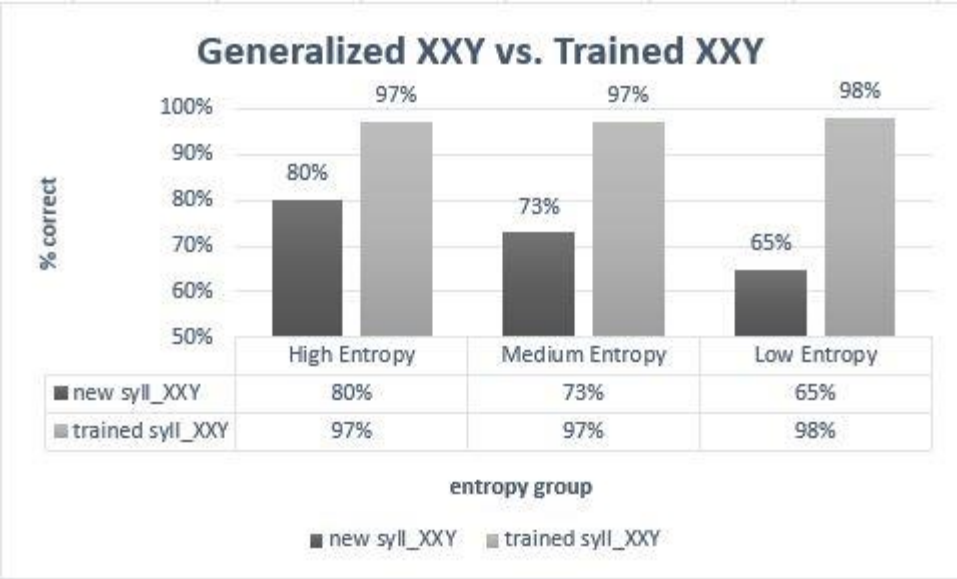


Figure 3. Percentage of correct answers for Type_3: untrained_syll_XXY vs. Type_1:trained_syll_XXY

Medium Entropy condition F(1, 113) = 1.153*, p = .28;* and for the mean difference between untrained XXY vs. trained XXY strings in the High Entropy vs. Medium Entropy condition F(1, 118) = 0.807*, p = .37.*

The analysis showed a statistically significant difference in performance for Type_4: trained_syll_XYZ (i.e. correctly rejected XYZ strings with trained syllables) as compared to the performance

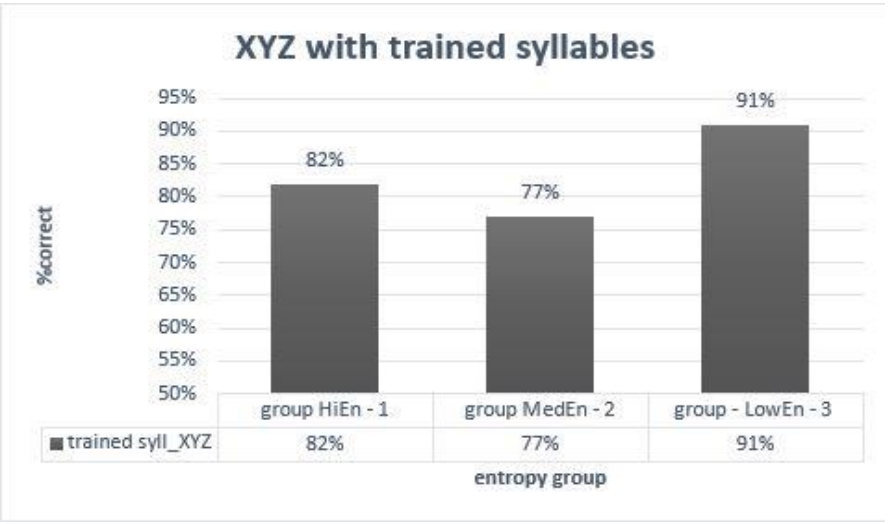


Figure 4. Percentage of correct answers for Type_4: trained_syll_XYZ

for Type_1: trained syll_XXY: for High Entropy condition (*p = .009*), for Medium Entropy condition (*p = .002*), and marginally significant for Low Entropy condition (*p = .099*). Figure 4 displays the mean performance across conditions: 82% for High Entropy (Mean = .82, SD = .39, SE = .050), significantly above chance (*t(59) = 6.28, p = .000*), 77% for Medium Entropy (Mean = .77, SD = .427, SE = .055), significantly above chance (*t(59) = 4.84, p = .000*), and 91% for Low Entropy (Mean = .91, SD = .290, SE = .039), significantly above chance (*t(54) = 10.45, p = .000*). An ANOVA showed a statistically significant better performance for Low Entropy condition compared to Medium Entropy condition (F(1,118) = 4.514, *p = 0.03*), and a non-significant trending in the predicted direction indicating a better performance in High Entropy condition vs. Medium Entropy condition (F(1,118) = 0.457, *p = 0.5*).

Figure 5 shows the percentage of correct answers for Type_2: untrained_syll_XYZ, for which there was no statistically significant difference as compared to performance for Type_1: trained_syll_XXY across conditions.

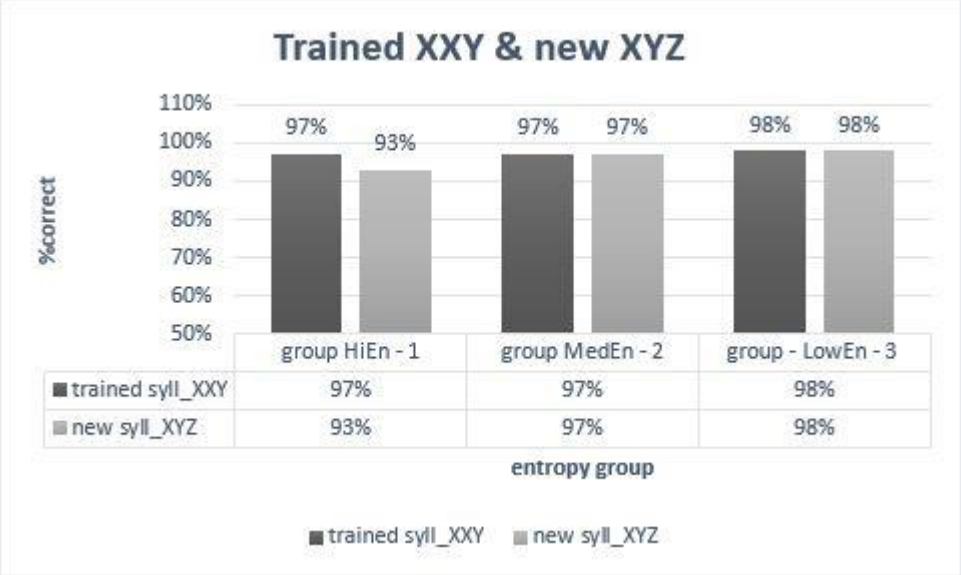


Figure 5. Percentage of correct answers for Type_1: trained_syll_XXY and Type_2:untrained_syll_XYZ

Results indicated a non-statistically significant trending in the predicted direction for Test x Entropy Group interaction. Figure 6 presents the percentage of correct answers for Type_3: untrained_syll_XXY (i.e. correctly accepted XXY strings with untrained syllables) shown by test.

Moreover, there was statistically significant above-chance performance across all test strings for all experimental conditions: the mean performance for High Entropy was 88% (Mean = .88, SD = .327, SE = .021; *t(239) = 17.985, p = .00*), for Medium Entropy was 86% (Mean = .86, SD = .349, SE = .023; *t(239) = 15.886, p = .00*), and for Low Entropy was 88% (Mean = .88, SD = .324, SE = .022; *t(219) = 17.503, p = .00*).

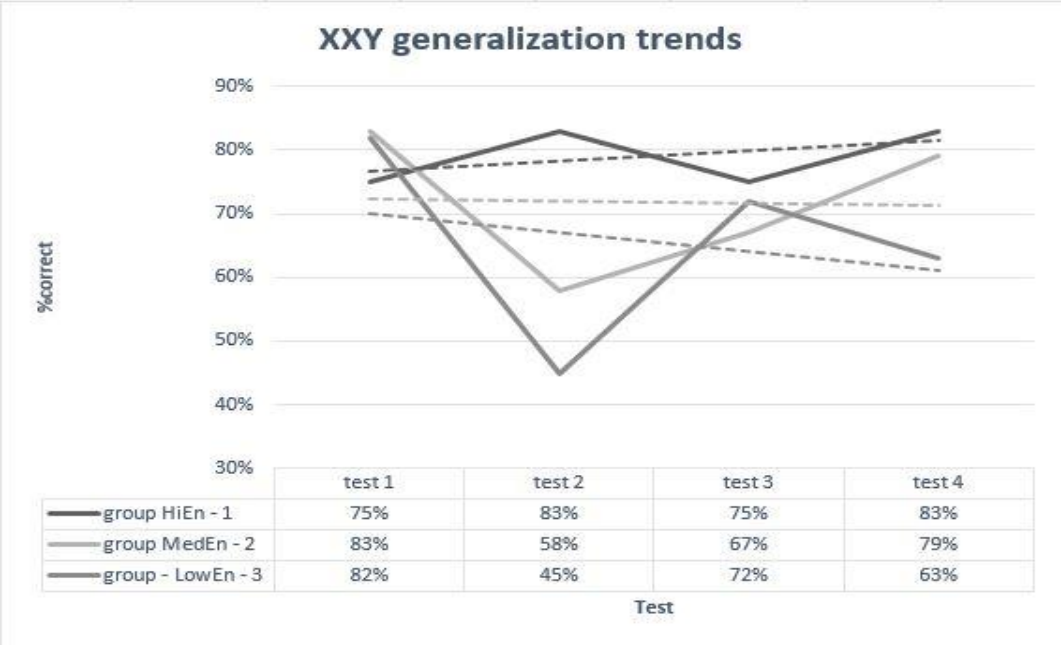


Figure 6. Percentage of correct answers for Type_3: untrained_syll_XXY detailed by test. The trend (dotted) lines show the direction of the generalization tendency as a function of exposure.

7. Discussion and Conclusions

This pilot study investigated in an artificial grammar experiment the effect of *input complexity* on the process of making generalizations, when *channel capacity* was kept constant, by testing adults of roughly the same age. According to the entropy model proposed in this paper, learning starts out with memorizing specific items and by developing perceptually-bound learning, which support dealing with the *input complexity* (entropy) as long as the *channel capacity* is not overloaded. When the *channel capacity* starts becoming overloaded and the efficiency of perceptually-bound learning in coping with the *input complexity* starts decreasing, because the complexity is too high compared to the available *channel capacity*, the tendency towards category-based generalization starts developing gradually.

The decrease across conditions in the mean performance for grammatical XXY strings with untrained syllables reveals a decrease in the tendency to abstract away from the memorized input as the

input complexity decreases. Moreover the different degrees of discrimination between XXY strings with novel syllables vs. XXY strings with trained syllables show differences between groups in terms of their tendency to generalize to new items: a lower difference shows a higher tendency to generalize. In the Low Entropy condition this difference is higher than in the Medium Entropy condition, which is higher than the difference in the High Entropy condition. Hence learners in the High Entropy condition had the highest tendency to fully generalize to novel strings displaying the underlying XXY pattern.

The roughly U-shaped performance revealed in the case of ungrammatical XYZ strings with trained syllables may point to the fact that the two tendencies (the perceptually-bound learning and rule induction) were working against each other, roughly as predicted, and pushing in different directions depending on their gained strength. Similar tendencies towards a U-shaped curve of learning have been found in previous language acquisition studies and they have been argued, among other explanations, to be due to the dynamics reflected by different mechanisms working simultaneously and interfering with each other (Rogers, Rakinson, & McClelland, 2004). Therefore, the results for XYZ strings with trained syllables may be considered to back up and complement the evidence brought by the results of untrained XXY strings, pointing to the interplay between the two tendencies towards generalizing.

Figure 7 depicts in a graph the hypothesized development and the interaction of the two tendencies. More specifically, the line labeled Trained_XYZ represents the function describing the measured values obtained from the Trained_XYZ results for *H = 3.5 bits, H = 4 bits, H = 4.58 bits*[14]. The line denoted INDUCT represents the function of rule-induction depending on the *input complexity*, increasing as per the measured values taken from the Untrained_XXY results for the same tested levels of entropy (the INDUCT values for *H = 3.5 bits, H = 4 bits, H = 4.58 bits* are the measured values for Untrained_XXY, and the other values in the table are predicted values according to the increasing trend of the measured values). PERCEPT represents the function of perceptually-bound learning, which is hypothesized to return constant high values of performance as long as the values of entropy are reduced, and decrease very abruptly (perhaps exponentially) as the values of entropy go up overloading *channel capacity*, and finally remain constant at a certain very low level without much change after a certain amount of entropy is reached, showing no change even if the entropy continues to grow. *F(H)* represents the calculated mathematical relation between the two functions – INDUCT and PERCEPT – based on the results obtained for Trained_XYZ strings and Untrained_XXY, and this function is hypothesized to predict the stage in

[14] These H values stand for the average entropy per training session: $\frac{\mathrm{H[bigram]+H[trigram]}}{2}$.

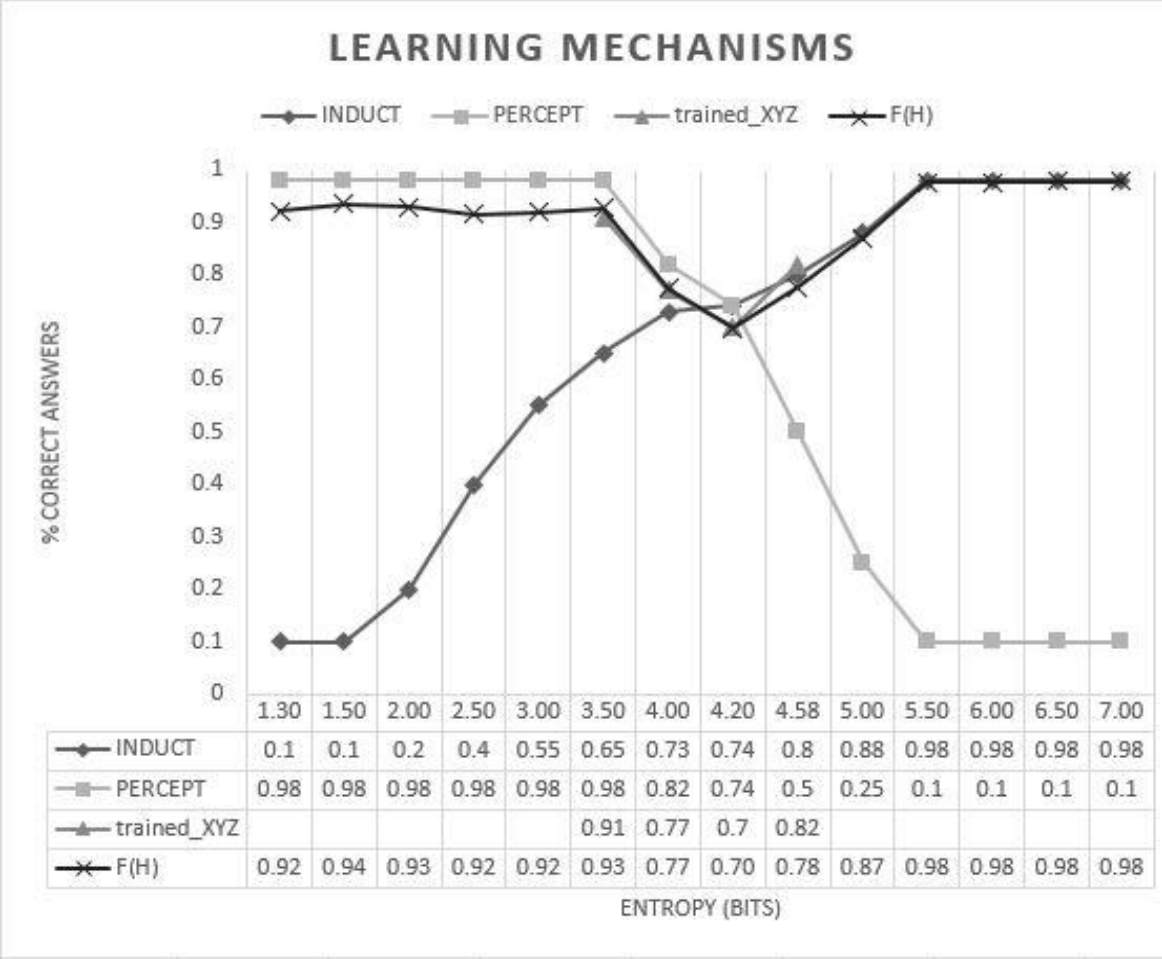
development of each tendency (measured in level of performance) and their interaction at any level of *input complexity*. As can be seen in the graph, the calculated values for *F(H)* are very close or identical to the actual measures (compare table values for *F(H)* to values for Trained_XYZ when *H = 3.5 bits, H = 4 bits, H = 4.58 bits,* in Figure 7).

Definition. Given *α,* so that *PERCEPT(α) = INDUCT(α)*:

$F(H) = PERCEPT(H) - 1/H^2 * INDUCT(H); H \leq \alpha$

$F(H) = INDUCT(H) - 1/H^2 * PERCEPT(H); H \geq \alpha$

According to the calculations, the predicted value for *α* is 4.2 bits, and this value is hypothesized to represent the point where the decreasing trend for Trained_XYZ reaches its minimum and changes into an increasing function, given that INDUCT outperforms PERCEPT. This point is predicted to roughly mark the overload limit of the *channel capacity*. Using this function, more predictions can be made about the



	1.30	1.50	2.00	2.50	3.00	3.50	4.00	4.20	4.58	5.00	5.50	6.00	6.50	7.00
INDUCT	0.1	0.1	0.2	0.4	0.55	0.65	0.73	0.74	0.8	0.88	0.98	0.98	0.98	0.98
PERCEPT	0.98	0.98	0.98	0.98	0.98	0.98	0.82	0.74	0.5	0.25	0.1	0.1	0.1	0.1
trained_XYZ						0.91	0.77	0.7	0.82					
F(H)	0.92	0.94	0.93	0.92	0.92	0.93	0.77	0.70	0.78	0.87	0.98	0.98	0.98	0.98

Figure 7. Learning mechanisms

level of performance and the development of each of the tendencies – perceptually-bound learning and rule induction - at specific levels of *input complexity* at the given *channel capacity.*

The mathematical relation expressed by this function predicts that the two tendencies work against each other with their respective amount of gained strength - each of them contributes a certain weight to the performance level, depending on their gained developmental stage: in the first stage, as long as the *input complexity* is low, PERCEPT has full weight in performance, while INDUCT has a lower weight ($1/H^2$), and at a certain amount of entropy - α, where the two tendencies lead to equal output (the function lines intersect on the graph), things change drastically and the weights are inversed: INDUCT has full weight in performance, and PERCEPT adds a reduced weight ($1/H^2$).

Regarding the learning process as a function of exposure (the performance curves across tests), as can be seen in Figure 6 depicting the percentage of correct answers for Type_3: untrained_syll_XXY detailed by test, although the results did not reach statistical significance (presumably because of the low number of data points per each test), the trends seem to match the predicted results: the performance curve decreases slightly steeper in LowEN, than in MedEn, and steeper than in HiEN. Further research would need to be carried out with larger samples in order to further investigate the generalization process as a function of exposure.

A post-hoc analysis of Type_4 test strings raised the question of a potential confound in the results for this type of strings: this type of test strings should have followed an X_1X_2Y pattern (where X_1 is different from X_2), in order to ensure that the reason for the rejection of these test 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 Type_4 test strings were found not to observe the X_1X_2Y pattern. However, if such confound interfered with the results, it is assumed that this fact would have helped rejection of these test strings. And if so, it is expected to have promoted rejection more in the LowEN condition, where it was easier for the participants to remember the distinction between trained X-syllables and Y-syllables. An ANOVA revealed no statistically significant difference between the X_1X_2Y strings and the non-X_1X_2Y strings in any of the conditions:

HiEN: Mean[X_1X_2Y] = .81, Mean[non-X_1X_2Y] = .83, $F(1,58) = .072$, $p = .79$.

MedEN: Mean[X_1X_2Y] = .79, Mean[non-X_1X_2Y] = .73, $F(1,58) = .293$, $p = .59$.

LowEN: Mean[X_1X_2Y] = .91, Mean[non-X_1X_2Y] = .91, $F(1,53) = .000$, $p = 1.00$.

Therefore, such confound is highly unlikely to have interfered with the results.

In conclusion, the entropy model proposed in this paper gives a conceptual analysis that accounts for both types of linguistic generalizations that have been found and conceptualized in previous studies – perceptually-bound generalizations and category-based abstractions – by identifying the same factors whose interplay is predicted to be the source of both types of generalizations: *input complexity* and *channel capacity*. This entropy model gives a quantitative measure for the likelihood of extracting a category-based generalization or a perceptually-bound rule, and thus it allows for precise predictions regarding the linguistic generalizations for any degree of *input complexity*. The results obtained in the experiment presented in this paper support the formulated hypotheses and bring evidence in favor of the validity of this entropy model for linguistic generalization.

In line with the predictions made by the proposed entropy model, the results of this experiment showed that in conditions of constant *channel capacity* (i.e. adults at roughly the same developmental stage) a low *input complexity* of the set of linguistic items drives the tendency towards perceptually-bound learning, and as a consequence reduces the tendency to generalize to novel input. Conversely, adults were shown to be more likely to tune into category-based generalizations when the linguistic input displayed a higher complexity. Therefore, in line with another prediction made by this entropy model, perceptually-bound learning and categorization/rule-induction were shown to be outcomes of the same learning mechanism, which is governed by the same principle of creating rules in response to the degree of *input complexity*.

In order to further test the entropy model proposed in this paper, further research should be carried out along the following lines:

a. keep *channel capacity* constant and vary *input complexity,* in order to replicate findings of this pilot project with a larger number of adults and to further investigate the curve of the learning under scrutiny, testing specific predictions made by the mathematical relation identified to describe the interplay between the two tendencies;
b. keep *channel capacity* constant at a lower level and vary *input complexity*, in order to replicate the findings of this experiment with infants, who have a lower *channel capacity*;
c. keep *input complexity* constant and vary *channel capacity,* by exposing both infants and adults to the same amount of entropy. Adults are expected to have a reduced tendency to make a category-based generalization as compared to infants.

d. test the effects of *input complexity* and *channel capacity* on the mechanisms of making generalizations in *visual input*. The results would be useful to investigate the question of domain-generality of generalization mechanisms.

The phenomena investigated in this study mark a qualitative developmental leap in the mechanisms underpinning language learning: moving away from a perceptually-bound mechanism that memorizes and produces constructions encountered in the input or with items encountered in the input, towards a category-based mechanism that applies abstract rules productively. By showing that generalization mechanisms are modulated by the interaction between *input complexity* and the limited *channel capacity* of the human brain, this research goes into the very heart of a core topic of linguistics and fills in an important gap in the puzzle about the induction problem for language acquisition.

Appendix 1

Training items

High Entropy		
Training 1	Training 2	Training 3
googoosjie	googoowoe	googoowoe
puupuuduu	puupuusjeu	puupuusaa
teuteusaa	Teuteubuu	teuteubuu
veeveefeu	veeveemoo	veeveedaa
woowoomoo	woowoofaa	woowookee
loolookee	Loolookoo	loolooduu
goegoesjoe	goegoemeu	goegoesjie
heuheuweu	heuheuchie	heuheukoo
juujuufie	juujuuzuu	juujuusjeu
nieniekaa	nieniefuu	nieniefeu
rooroomuu	roorooboo	rooroomoo
vuuvuuchoo	vuuvuuhuu	vuuvuufaa
haahaawoe	haahaasjie	haahaameu
hiehiesjeu	hiehieduu	hiehiezuu
jiejiebuu	jiejiesaa	jiejiechie
joejoedaa	joejoefeu	joejoefuu
jeujeufaa	jeujeudaa	jeujeuboo
lieliekoo	lieliekee	lieliehuu
luuluumeu	luuluusjoe	luuluusjoe
noonoochie	noonooweu	noonooweu
noenoezuu	noenoefie	noenoefie
ruuruufuu	ruuruukaa	ruuruukaa
vievieboo	vieviemuu	vieviemuu
veuveuhuu	veuveuchoo	veuveuchoo

Medium Entropy		
Training 1	Training 2	Training 3
googoosjie	googookaa	googookee
puupuusjie	puupuukaa	puupuusaa
teuteuduu	teuteumuu	teuteusaa
veeveeduu	veeveemuu	veeveemuu
woowoosaa	woowoochoo	woowoochoo
looloosaa	looloochoo	looloochoo
goegoefeu	goegoesjie	goegoesjie
heuheufeu	heuheusjie	heuheusjie
juujuumoo	juujuuduu	juujuuduu
nieniemoo	nienieduu	nieniefeu
roorookee	rooroosaa	rooroofeu

vuuvuukee	vuuvuusaa	vuuvuumoo
googoosjoe	googoofeu	googooweu
puupuusjoe	puupuufeu	puupuufie
teuteuweu	teuteumoo	teuteufie
veeveeweu	veeveemoo	veeveemoo
woowoofie	woowookee	woowookee
looloofie	loolookee	loolooduu
goegoekaa	goegoesjoe	goegoesjoe
heuheukaa	heuheusjoe	heuheusjoe
juujuumuu	juujuuweu	juujuuweu
nieniemuu	nienieweu	nieniekaa
rooroochoo	rooroofie	roorookaa
vuuvuuchoo	vuuvuufie	vuuvuumuu

Low Entropy		
Training 1	Training 2	Training 3
googoosjie	googoosjie	googookee
puupuusjie	puupuusaa	puupuusaa
teuteusjie	teuteuduu	teuteusjie
veeveesjie	veeveesjie	veeveesjie
woowooduu	woowooduu	woowoosaa
loolooduu	looloosjie	looloosjie
googooduu	googoosaa	googooduu
puupuuduu	puupuuduu	puupuuduu
teuteusaa	teuteusaa	teuteuduu
veeveesaa	veeveeduu	veeveeduu
woowoosaa	woowoosjie	woowoosjie
looloosaa	looloosaa	looloosaa
googoofeu	googoofeu	googoosaa
puupuufeu	puupuukee	puupuukee
teuteufeu	teuteumoo	teuteufeu
veeveefeu	veeveefeu	veeveefeu
woowoomoo	woowoomoo	woowookee
looloomoo	looloofeu	looloofeu
googoomoo	googookee	googoomoo
puupuumoo	puupuumoo	puupuumoo
teuteukee	teuteukee	teuteumoo
veeveekee	veeveemoo	veeveemoo
woowookee	woowoofeu	woowoofeu
loolookee	loolookee	loolookee

Test items

	Test 1
type_1	googoosjie
type_2	doovaaseu
type_3	peupeudee
type_4	teuduusaa
	Test 2
type_1	veeveemoo
type_2	reuloegee
type_3	tootoosuu
type_4	puuwoofeu
	Test 3
type_1	woowookee
type_2	mietaazoe
type_3	voovoofoo
type_4	loogoomoo
	Test 4
type_1	puupuusaa
type_2	foeseebie
type_3	waawaazeu
type_4	veeduuwoo
type_1	loolooduu
type_2	keusoodoo
type_3	geugeukie
type_4	teuveekee

References

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.

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

Gómez, R.L. & Gerken, L.A. (2000). Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences 4*, 178–186

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.

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.

Pothos, E. M. & Bailey, T. M. (2000). The importance of similarity in Artificial Grammar Learning. *J. Exp. Psychol. Learn. Mem. Cogn. 26*, 847–862.

Pothos, E. M. (2010). An entropy model for Artificial Grammar Learning. *Frontiers in Cognitive Science*, *1*, 1-13.

Pothos, E. M., Chater, N., & Ziori, E. (2006). Does stimulus appearance affect learning? *Am. J. Psychol. 119*, 277–301.

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.

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

Tononi, G. (2008). Consciousness and Integrated Information: a Provisional Manifesto. *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.